ClinicalNLP 2023

The 5th Workshop on Clinical Natural Language Processing (ClinicalNLP)

Proceedings of the Workshop

July 14, 2023

The ClinicalNLP organizers gratefully acknowledge the support from the following sponsors.

Lavita

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-959429-88-3

Preface

This volume contains papers from the 5th Workshop on Clinical Natural Language Processing (Clinical NLP), held at ACL 2023.

Clinical text offers unique challenges that differentiate it not only from open-domain data, but from other types of text in the biomedical domain as well. Notably, clinical text contains a significant number of abbreviations, medical terms, and other clinical jargon. Clinical narratives are characterized by non-standard document structures that are often critical to overall understanding. Narrative provider notes are designed to communicate with other experts while at the same time serving as a legal record. Finally, clinical notes contain sensitive patient-specific information that raise privacy and security concerns that present special challenges for natural language systems. This workshop focuses on the work that develops methods to address the above challenges, with the goal of advancing state-of-the-art in clinical NLP.

ClinicalNLP 2023 also hosted the MEDIQA-Chat 2023 shared tasks that promote research on effective solutions for clinical note generation from medical conversations. The shared tasks focused on the summarization of doctor-patient conversations and on the generation of synthetic dialogues from clinical notes for data augmentation. They introduced new benchmarks for training and evaluation and used an ensemble of evaluation metrics that highly correlate with human judgments. Further, the organizers added a new requirement to submit the code for a second evaluation of the outputs. The MEDIQA-Chat shared tasks attracted 120 registered teams with 17 teams submitting their codes and runs for official participation. The participating teams experimented with the recently released Large Language Models (LLMs) vs. older models and explored data augmentation, fine-tuning, and prompting methods. The results provided new insights on the best approaches and techniques for future research directions in clinical text generation.

This year, we received the total of 82 submissions, inclusive of shared task submissions, from which 58 were accepted for presentation.

Organizing Committee

General Chairs

Tristan Naumann, Microsoft Research
Asma Ben Abacha, Microsoft
Steven Bethard, University of Arizona
Kirk Roberts, UTHealth Houston
Anna Rumshisky, UMass Lowell

MEDIAQA-Chat 2023 Shared Task Organizers

Asma Ben Abacha, Microsoft
Wen-wai Yim, Microsoft
Griffin Adams, Columbia University
Neal Snider, Microsoft/Nuance
Meliha Yetisgen, University of Washington

Program Committee

Program Committee

Griffin Thomas Adams, Surabhi Adhikari, Ashwag Alasmari, Amal Abdullah Alqahtani, Emily Alsentzer, Ibtihel Amara, Hajer Ayadi

Steven Bethard, Sjaak Brinkkemper

Leonardo Campillos-Llanos, Salim Chemlal, Qingyu Chen, Shan Chen, Cheryl Clark

Hong-Jie Dai, Judith W Dexheimer, Dmitriy Dligach, Richard Dufour, Jocelyn Dunstan

Naome A Etori

Matúš Falis, Yadan Fan, Xavier Fontaine

Zelalem Gero, John Michael Giorgi, Natalia Grabar

Mei-Hua Hall, Sadid A. Hasan, Rim Helaoui, Ming Huang

Raphael Iyamu

Xiaoqian Jiang, Qiao Jin, Alistair Johnson

Yoshinobu Kano, Raghav Kapoor, Yejin Kim, Martin Krallinger, Shiba Kuanar

Yanis Labrak, Egoitz Laparra, Alberto Lavelli, Ulf Leser, Qiuhao Lu, Yuxing Lu

Diwakar Mahajan, Pranita Yogesh Mahajan, Gaetano Manzo, Sérgio Matos, George Michalopoulos, Kirill Milintsevich, Timothy A Miller, Ishani Mondal

Aakanksha Naik, Tristan Naumann, Aurélie Névéol

Kadir Bulut Ozler

Ankur Padia, Yifan Peng, Joseph M Plasek

Satyajeet Raje, Pavithra Rajendran, Giridhar Kaushik Ramachandran, Anku Rani, Fabio Rinaldi, Kirk Roberts, Matías Rojas, Frank Rudzicz

Guergana K Savova, Ashwyn Sharma, Atul Singh, Sonish Sivarajkumar, Sarvesh Soni, Arvind Krishna Sridhar, Dhananjay Srivastava, Karl Eric Swanson

Behrad Taghibeyglou, Neset Tan, Xiangru Tang, Khushboo Thaker, Augustin Toma

Byron C Wallace, Jinge Wu, Susmitha Wunnava

Dongfang Xu

Meliha Yetisgen, Wen-wai Yim, WonJin Yoon, Paul Youssef, Xiang Yue

Boya Zhang, Weipeng Zhou

Keynote Talk: Patient record summarization: tasks, approaches, evaluation, and open challenges

Noémie Elhadad
Columbia University

Abstract: The patient record contains an overwhelming large amount of information, too much for a clinician to make sense of it, and yet the information it contains may be critical for clinicians to care for their patients safely and effectively. In this talk, I will review two tasks to alleviate the information overload in clinical care: longitudinal patient record summarization and abstractive brief hospital course summarization. I will describe potential approaches, evaluation objectives, and current open questions. Finally, using the abstractive task of brief hospital course summarization as a grounding example, I will discuss large language models (LLMs) in the context of clinical NLP.

Bio: Noémie Elhadad is Chair of the department of Biomedical Informatics at Columbia University, affiliated with the department of Computer Science and the Data Science Institute. Elhadad's research lies at the intersection of artificial intelligence, human-centered computing, and medicine. She creates novel methods and tools to support patients and clinicians in their information needs, with particular focus on ensuring that the AI systems of the future are robust, safe, fair, and just.

Keynote Talk: The evolution of representations for clinical text and a few more thoughts about generative clinical models

Timothy Miller
Boston Children's Hospital, Harvard Medical School

Abstract: Large language models (LLMs) have excited the broader public like no previous NLP advance. This has led to predictions from all corners about the future of LLM-enabled NLP for clinical data and tasks. In this talk, I review several recent projects from my lab that did not use LLMs, and re-imagine these projects in an LLM-enabled context. The talk then synthesizes the lessons from those projects to propose some guidelines for optimal use of LLMs in clinical NLP research, imagine future directions that are now enabled, and to make some predictions about the future of our field.

Bio: Tim Miller is an Associate Professor in the Computational Health Informatics Program at Boston Children's Hospital, Department of Pediatrics at Harvard Medical School, and at the Harvard-MIT Center for Regulatory Science. He is the PI of the Machine Learning for Medical Language Lab, home of several federally funded projects, including projects focused on basic biomedical NLP research, as well as projects that are driven by biomedical use cases. His research focuses on domain adaptation/generalizability of ML-based NLP methods, as well as methods for learning universal patient representations.

Table of Contents

Clinical BERTScore: An Improved Measure of Automatic Speech Recognition Performance in Clinical Settings

Joel Shor* Verily Life Sciences, USA joelshor@verily.com

Ruyue Agnes Bi* MIT, USA ruyuebi@mit.edu

Subhashini Venugopalan Google Research, USA

Steven Ibara Verily Life Sciences, USA

Roman Goldenberg Verily Life Sciences, Israel

Ehud Rivlin Verily Life Sciences, Israel

Abstract

Automatic Speech Recognition (ASR) in medical contexts has the potential to save time, cut costs, increase report accuracy, and reduce physician burnout. However, the healthcare industry has been slower to adopt this technology, in part due to the importance of avoiding medically-relevant transcription mistakes. In this work, we present the Clinical BERTScore (CBERTScore), an ASR metric that penalizes clinically-relevant mistakes more than others. We collect a benchmark of 18 clinician preferences on 149 realistic medical sentences called the Clinician Transcript Preference benchmark (CTP) and make it publicly available[1] for the community to further develop clinically-aware ASR metrics. To our knowledge, this is the first public dataset of its kind. We demonstrate that our metric more closely aligns with clinician preferences on medical sentences as compared to other metrics (WER, BLUE, METEOR, etc), sometimes by wide margins.

1 Introduction

Clinicians in a number of disciplines work in an overburdened healthcare system that leads to difficult working environments and an epidemic of physician burnout (Dzau et al., 2018). AI-related technologies have the potential for improving efficiency on repetitive tasks, therefore increasing both patient throughput and decreasing physician burnout. For example, physicians in a number of disciplines spend as much time doing paperwork as with patients (Tai-Seale et al., 2017). However, the adoption of speech technology in the medical community has been slow (Latif et al., 2021), and there are a number of speech technologies that could improve efficiency.

Speech technology can be applied to a number of medical problems including transcribing patient-physician conversations (Shafran et al., 2020), helping dysarthric patients communicate (Shor et al., 2020), and diagnosing medical conditions from speech (Shor et al., 2022; Shor and Venugopalan, 2022; Peplinski et al., 2021; Venugopalan et al., 2021). In this work, we focus on the task of generating a report after a colonoscopy procedure.

One of many reasons for the lower adoption of time-saving speech transcription technologies is that the ASR systems often don't perform as well in real-world clinical settings as they do on evaluation benchmarks. The most common metric for measuring ASR performance, Word Error Rate (WER), has significant practical drawbacks (Wang et al., 2003; Morris et al., 2004; He et al., 2011). First, all mistakes are treated equally. In clinical settings, however, medical words are more important (e.g. "had complete resection" → "had complete **c-section**" is a worse mistake than → "**has** complete resection", but both have equal WER). Second, some mistakes affect the overall intelligibility more than others (e.g. "was no perforation" → "was no **puffer age**" vs "was **not any** perforation"). Although researchers have proposed alternatives to the WER, no metric combines medical domain knowledge with recent AI advances in language understanding.

In this work, we make the following contributions:

1. Generate a collection of realistic medical sentences and transcripts with plausible ASR errors and collect preferences from 18 clinicians on 149 sentences. We publicly released this dataset for reference and future studies. This is the first public dataset of its kind.
2. Present the Clinical BERTScore (CBERTScore) and demonstrate that it more closely matches clinician preferences on medical transcripts than other ASR metrics (WER, BLEU, METEOR, BERTScore).
3. Demonstrate that CBERTScore does not perform worse than other metrics on non-medical

*Authors contributed equally

[1] https://osf.io/tg492/

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 1–7
July 14, 2023 ©2023 Association for Computational Linguistics

transcripts.

2 Related work

There are a number of ways to evaluate transcript quality. The Word Error Rate (WER), is the simplest to compute and most common. It counts the number of insertions, deletions, and substitutions between two text strings, and normalizes by the length of the reference string. The Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002) measures the amount of n-gram overlap between two text strings (where *n* is often 4). It captures the intuition that groups of words are important in addition to individual words. METEOR (Banerjee and Lavie, 2005) focuses on unigrams, but computes an explicit alignment between two strings and takes both precision and recall into consideration. While these techniques are cheap to compute, they primarily focus on character or string similarity, not semantic similarity.

Our work most closely follows the BERTScore (Zhang et al., 2019). This metric computes a neural word embedding for each word in the reference and candidate. Embeddings are matched using cosine distance instead of string similarity, and the final score takes precision and recall into account (see Fig.1). This method takes semantic similarity into account, but not that some words are more important to preserve in clinical contexts.

Structured graphs are one way to encode real-world knowledge in a machine-readable format. The Knowledge Graph (KG) (Singhal, 2012) is a publicly available structure that encodes medical knowledge. Previous work has used the medical subset of the KG to learn medical entity extraction (Shafran et al., 2020). We primarily follow this approach to determine which words are clinically significant.

3 Methods

3.1 Clinical BERTScore

Our proposed metric, the Clinical BERTScore (CBERTScore), combines the BERTScore (Zhang et al., 2019) and the medical subset of the Knowledge Graph (Shafran et al., 2020).

BERTScore is a relatively novel language generation evaluation metric proposed in (Zhang et al., 2019) based on pre-trained BERT contextual embeddings. It is designed to capture semantic similarity between two sentences, instead of simple string matching. Given a reference sentence $x = \langle x_1, ..., x_k \rangle$ and a candidate sentence $\hat{x} = \langle \hat{x}_1, ..., \hat{x}_l \rangle$, we first represent each token by a contextual embedding, and then calculate the cosine similarities between the tokens. Each token in the reference sentence is matched to the most similar token in the candidate sentence, and vice versa. The former is used to compute the recall R_{BERT}, and the latter to compute the precision P_{BERT}. Precision and recall are then combined into a single score BERTScore as follows:

$$R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^T \hat{\mathbf{x}}_j,$$

$$P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^T \hat{\mathbf{x}}_j$$

$$\text{BERTScore} = 2\frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}$$

Building on this, we define CBERTScore:

$$\textbf{CBERTScore}(x, \hat{x}) = k \times \text{BERTScore}_{\text{medical}}(x, \hat{x}) + (1-k) \times \text{BERTScore}_{\text{all}}(x, \hat{x}), \text{ where } 0 \leq k \leq 1$$

$\text{BERTScore}_{\text{all}}$ is computed over all words in the sentences, and $\text{BERTScore}_{\text{medical}}$ is computed over a subset of them that are medically relevant. If there are no medical terms in either the reference or candidate sentence, we define the CBERTScore to be the standard BERTScore (on all words), i.e., k is set at 0.

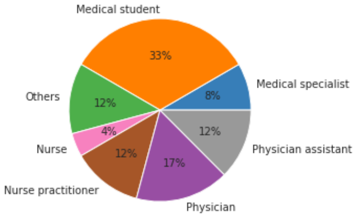
We inject medical information into this metric in two ways. First, we compute a weighted score on a subset of words involving medical terms, as determined by the Knowledge Graph (Shafran et al., 2020). Second, we tune the weight of the clinical term penalty to best match a clinician transcript dataset (CTP) that we collected. We describe our method for determining k in Sec. 3.1.2.

3.1.1 Medical Entities

Similar to (Shafran et al., 2020), we derive roughly 20K medically relevant words from Google's Knowledge graph (Singhal, 2012). These words come from entities with properties such as "/medicine/disease", "/medicine/drug", "/medicine/medical_treatment", and "/medicine/medical_finding". We also include numbers for the CBERTScore algorithm, since numerical accuracy is important in medical contexts.

3.1.2 Tuning the medical entities weight factor

CBERTScore has a parameter controlling the weight of the clinical component. To determine



Original	Transcript 1	Transcript 2	Clinician preference	WER	BLEU	METEOR	BERTScore	CBERTScore
We see a serrated five millimeter hyperplastic polyp during a u-turn.	We see a terraced five millimeter hyperplastic polyp during a u-turn.	We see a serrated five millimeter hyperplastic polyp doing a u-turn.	2	Same	Same	Same	2	2
A biopsy forceps is being used.	A biopsy force is being used.	A biopsy forceps was used.	2	1	1	2	2	2
Three clips were applied to close these defects.	Three clips were applied to close these side effect.	Three clips were replied to close these defects.	2	2	1	2	1	2
Coming down from the descending colon and to the sigmoid colon.	Came down from the descending colon and to the sigmoid colon.	Coming down from the ascending colon and to the sigmoid colon.	1	Same	1	1	2	1

Figure 1: **Left:** Background of the clinicians who were surveyed to create the Clinician Transcript Preference (CTP) dataset. **Right:** Some examples of triplet medical sentences, which transcript clinicians prefer, and which transcript scores better based on different metrics.

this factor, we picked the best performing k on the training subset of the Clinician Transcript Preference (CTP) dataset (Sec. 3.2). We evaluated k using 11 points evenly spaced between 0 and 1, and performed the evaluation methodology in Sec. 3.2 for each. We then used this value for all subsequent results and analyses.

3.2 Clinician Transcript Preference (CTP) Dataset

In order to compare CBERTScore's agreement with human preference, we sent out a Qualtrics survey to elicit judgment specifically from clinicians[2]. We call this dataset the Clinician Transcript Preference dataset (CTP), and we make it publicly available on the Open Science Framework (OSF). To our best knowledge, this is the first publicly available dataset with clinician preferences of transcript errors.

We collected data on 150 sentences. They were divided into three groups, each containing 50 trials. 18 subjects with clinical backgrounds responded to more than half the questions. Fig. 1 (left) describes clinician backgrounds. Each participant was randomly assigned to a group to ensure approximately uniform response coverage. For each trial, participants are given a ground truth sentence and two "transcripts" and asked to select the less useful one or to indicate the two are about the same. An example of such a triplet is as follows: *"Patient elects to go under Propofol sedation."*
#1: Patient elects to go under Prilosec sedation.
#2: Patient selects to go under Propofol sedation.
The survey was designed to take no more than 20 min to minimize the cognitive strain on participants. One sentence was malformed, resulting in 149 sentences for the final dataset.

3.2.1 Constructing the CTP triplets

To generate the triplets of (target, transcript #1, transcript #2) used in the survey, we started by downloading publicly available YouTube videos on colonoscopies created by GI physicians and educational institutes. The target sentences were transcribed by Google's publicly available Speech-to-Text medical dictation model (Soltau et al., 2021) and manually checked for accuracy. Filler words such as "uh" and repeated words were edited out. Sentences longer than 30 words or less than 5 were discarded.

For each target sentence, transcript #1 was generated by one of Google's other, non-medical, publicly available ASR models. Transcripts with an edit distance(edi) outside [1, 3] were discarded. This procedure generated 1220 candidate sentences.

To ensure that the two transcripts were roughly comparable in terms of fidelity, transcript #2 was generated synthetically. We used a publicly available English word frequency dictionary(Goldhahn et al., 2012) to select words in the target sentence that were candidates for synthetic errors. Candidate words were at least 5 characters, appeared in the 1M word dictionary fewer than 10 times, and were not proper nouns. 486 candidate sentences matched these criteria. Finally, transcript #2 was generated by deleting the candidate word or manually substituting it with a phonetically similar word or phrase[3]. We discarded similar sentences and selected 150 triplets for the final survey. The ordering of the two transcripts was randomized, and so were the sentences.

3.2.2 Evaluating metrics on the CTP

To compare the ability of different metrics to agree with rater preference from the CTP, we define a

[2]Broadly defined as a person with extensive clinical experience or from a clinical research background, for our purpose.

[3]A Python fuzz search algorithm based on CMU Pronouncing Dictionary was used for consistency.

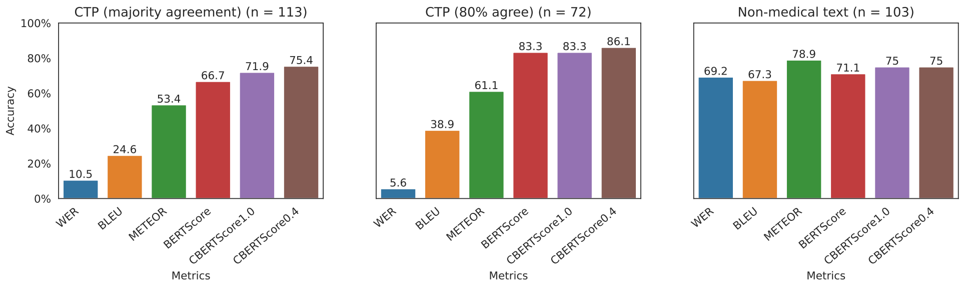


Figure 2: Comparison of different metrics' agreement with human rater transcript preferences. Process of deriving a prediction from metric values is described in Sec. 3.2. In all plots, "CBERTScore1.0“ is the performance from only the medical term component ($k = 1.0$ in Sec. 3.1). "CBERTScore0.4“ uses the optimal value of k according to the train set. **Left:** Agreements with clinicians on the CTP benchmark when labels are derived using majority voting. **Center:** Agreements with clinicians on the CTP benchmark when restricted to questions with unanimous answers. **Right:** Agreement with speech pathologist raters on the non-medical dataset, when restricting the data to cases where there is a fidelity difference between two candidate transcripts.

3-class classification problem as follows:

$$\text{Predicted better transcript}(M)(gt, t_1, t_2) = \begin{cases} t_1 & M(gt, s_1) - M(gt, s_2) > l \\ t_2 & M(gt, s_1) - M(gt, s_2) < -l \\ \text{same} & \text{else} \end{cases}$$

where M is an evaluation metric, gt is the ground truth sentence, and t_i are the transcripts. Note the predictions are reversed for the WER, since lower values indicate higher fidelity. l is a free variable, which we optimize separately for each metric. We split the data into two halves, choose the best performing l on one half, and report the accuracy using that l on the second half.

3.2.3 Non-medical sentences

To demonstrate that CBERTScore doesn't degrade on non-medical speech, we compare the metrics' agreement with rater preferences on a dataset with annotations similar to (Tobin et al., 2022). Part of this dataset consists of 5-tuples of (ground truth sentence, transcript 1, transcript 2, assessment 1, assessment 2), where the sentence assessments describe how much of the ground truth sentence's meaning is captured in the transcript. We used a subset of 103 utterances from our annotated data where the ratings were not the same, and at least one transcript was rated as having “Major errors". We report performance using a similar formulation as on the CTP evaluation in Sec. 3.2.2: we frame this as a 2-way classification problem (no cutoff is needed since we exclude tuples that have the same rating).

4 Results

4.1 Clinician responses

18 clinicians responded to a total number of 149 triplet questions. Each question had 5 or 6 responses. 78% of questions had more than half agreement on which transcript was less useful and 42% had more than 80% agreement. Clinicians thought transcripts were the same usefulness in 21% of cases.

4.2 Metric agreement on medical text

We report 3-way accuracy classification on the CTP dataset using two labeling schemes (Fig. 2). In the first, we only look at the questions where more than half the respondents agreed. In the second, we report accuracy on the questions where more than 4/5 of the respondents agreed. For both numbers, we determine the cutoff from one half the data and report accuracy on the second half.

First, the metric ordering by performance is the same using both labeling schemes, and the best CBERTScore medical weighting factor was the same using both label schemes. Second, BERTScore and CBERTScore are significantly more closely aligned with clinician preferences than other metrics. Third, CBERTScore weighted entirely toward medical terms outperforms or ties with BERTScore agreement. Fourth, the weighted combination of medical and non-medical terms outperforms other metrics in terms of clinician agreement. Fifth, the medical component meaningfully improves the performance of CBERTScore over BERTScore (75.9% vs 67.2% and 87.5% vs 84.4%).

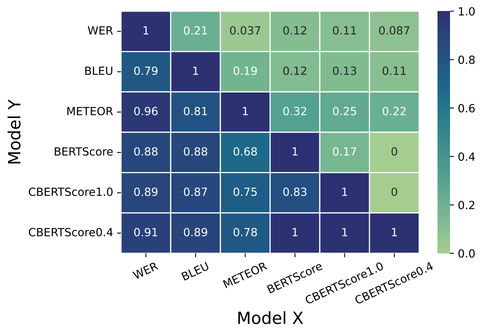


Figure 3: Fraction of cases where metric Y is correctly conditioned on metric X and Y disagreeing. An indicator of how similar the pattern of mistakes is between metrics.

4.3 Metric agreement on non-medical text

CBERTScore was the second best-performing metric on non-medical text. Importantly, the addition of the medical component did not degrade the performance compared to BERTScore.

5 Discussion

5.1 Knowledge Graph medical terms wins and losses on the CTP

The CTP (Sec. 3.2) had 127 distinct words that were the source of transcript errors, and 684 distinct other words. The medically-relevant terms used in the CBERTScore algorithm, identified primarily from the Knowledge Graph as described in Sec. 3.1.1, intersected with 99 of the 127 transcript error words. By manual inspection, 25 of the 28 transcript error words in the CTP not included in the CBERTScore word list were used in a medical context but were not only medical in meaning (ex. “surveillance", “tethered", and “longitudinal"). 3 of the 28 missed words did have a primarily medical meaning, but were not included in the CBERTScore list either due to errors in the KG or errors in the queries generating the list (“cologuard", “colonoscope", “protuberance"). Some of the words have a clear meaning in a medical context, and could be manually added to the list for future applications (“snare", “suctioning", etc.).

The CBERTScore word list included 100 words that weren’t selected for transcript errors. Many of these are medical in nature, but were not selected for synthetic transcript errors via the method described in Sec. 3.2 (ex. “endoscope", “hypoplastic", “lymphoma").

5.2 CBERTScore performance on the CTP

5.2.1 CBERTScore wins

Fig. 3 left shows the degree to which better-performing metrics subsume other metrics, or make a different pattern of mistakes. The plot shows the (Metric Y correct)/(Metric X and Y disagree). Metrics that have higher clinician agreement and a high fraction on this plot are strictly better, whereas metrics with higher agreement but a low value in this plot indicate that another metric might have an additional signal. We see that CBERTScore is nearly strictly better than the other metrics, with the possible exception of METEOR (when they differ, METEOR gives the correct rating in roughly a third of cases).

There were some triplets that CBERTScore got correct that no other metric did. The improvements over BERTScore always involved a medical term, and sometimes involved encouraging the metric to prioritize medical mistakes (ex. "Marked the site with 5 cc's of indigo carmine.“ $\rightarrow$ "Marked the site with 5 cici's of indigo carmine.“ vs "Marked the sight with 5 cc's of indigo carmine.“)

There were thirteen triplets that the neural word embeddings predicted correctly that other metrics did not. Many of these wins came from the strength of neural word embeddings penalizing less for semantically similar mistakes (ex. "Small burst of coagulation to create a darkish white ablation.“ $\rightarrow$ "Small burst of coagulation to create a darkish white oblation.“ vs "Small burst of coagulation to create a dark white ablation.“). Furthermore, BERTScore agreed with clinicians on some medical word mistakes, likely due to the BERT embedding somewhat understanding when a transcript error leads to a large semantic change in a medical term (ex. "No ongoing infection or coagulopathy.“ $\rightarrow$ "No on going infection or coagulopathy.“ vs "No ongoing infection or glomerulopathy.“).

5.2.2 CBERTScore mistakes

Fig. 3 shows that METEOR made the most correct predictions when CBERTScore was incorrect. Some mistakes are due to the KG medical list being incomplete. For example, "longitudinal“ was not included, but has medical meaning in clinical contexts (ex. "The longitudinal extent of the hot snare.“ $\rightarrow$ "The long eternal extent of the hot snare.“ vs "The longitudinal extend to the hot snare.“).

Another pattern of mistake is when a non-medical adjective contains an error, but the adjective modifies a medical term in an important way. For example, "vessel" is a medical term, but "feeding" is not (ex. "This polyp is at high risk of bleeding, with multiple feeding vessels." → "This polyp is at high risk of bleeding, with multiple seeding vessels." vs "This polyp is at high risking bleeding, with multiple feeding vessels."). This suggests that future work might include modifications and dependencies when calculating clinical importance.

Finally, a third pattern of mistake involves the fact that METEOR penalizes complex correspondences between candidate and reference sentences, while CBERTScore only considers the best pairwise word matches. One example in the CTP preserves most of the words, but reorders them (ex. "Inject into the head of the polyp, another 1 to 2 cc." → "Injectant the head of the polyp, another 1 to 2 cc." vs "Inject into the head of the polyp, another 1 2 to cc.").

6 Conclusions

We present CBERTScore, a novel metric that combines medical domain knowledge and recent advances in neural word embeddings. We collect and release a benchmark of clinician rater preferences on transcript errors, demonstrate that CBERTScore is more closely aligned with clinician preferences, and release the benchmark for the research community to continue to improve ASR in medical contexts.

References

Edit distance. https://en.wikipedia.org/wiki/Edit_distance. Accessed: 2023-03-03.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Victor J Dzau, Darrell G Kirch, Thomas J Nasca, et al. 2018. To care is human—collectively confronting the clinician-burnout crisis. *N Engl J Med*, 378(4):312–314.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Xiaodong He, Li Deng, and Alex Acero. 2011. Why word error rate is not a good metric for speech recognizer training for the speech translation task? In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5632–5635.

Siddique Latif, Junaid Qadir, Adnan Qayyum, Muhammad Usama, and Shahzad Younis. 2021. Speech technology for healthcare: Opportunities, challenges, and state of the art. *IEEE Reviews in Biomedical Engineering*, 14:342–356.

Andrew Cameron Morris, Viktoria Maier, and Phil Green. 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Eighth International Conference on Spoken Language Processing*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Jacob Peplinski, Joel Shor, Sachin Joglekar, Jake Garrison, and Shwetak Patel. 2021. FRILL: A non-semantic speech embedding for mobile devices. In *Interspeech 2021*. ISCA.

Izhak Shafran, Nan Du, Linh Tran, Amanda Perry, Lauren Keyes, Mark Knichel, Ashley Domin, Lei Huang, Yu-hui Chen, Gang Li, Mingqiu Wang, Laurent El Shafey, Hagen Soltau, and Justin Stuart Paul. 2020. The medical scribe: Corpus development and model performance analyses. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2036–2044, Marseille, France. European Language Resources Association.

Joel Shor, Aren Jansen, Wei Han, Daniel Park, and Yu Zhang. 2022. Universal paralinguistic speech representations using self-supervised conformers. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Joel Shor, Aren Jansen, Ronnie Maor, Oran Lang, Omry Tuval, Félix de Chaumont Quitry, Marco Tagliasacchi, Ira Shavitt, Dotan Emanuel, and Yinnon Haviv. 2020. Towards Learning a Universal Non-Semantic Representation of Speech. In *Proc. Interspeech 2020*, pages 140–144.

Joel Shor and Subhashini Venugopalan. 2022. TRILLsson: Distilled universal paralinguistic speech representations. In *Interspeech 2022*. ISCA.

Amit Singhal. 2012. Introducing the knowledge graph: things, not strings. https://blog.google/products/search/introducing-knowledge-graph-things-not. Accessed: 2023-03-03.

Hagen Soltau, Mingqiu Wang, Izhak Shafran, and Laurent El Shafey. 2021. Understanding medical conversations: Rich transcription, confidence scores & information extraction. In *Interspeech*.

Ming Tai-Seale, Cliff W Olson, Jinnan Li, Albert S Chan, Criss Morikawa, Meg Durbin, Wei Wang, and Harold S Luft. 2017. Electronic health record logs indicate that physicians split time evenly between seeing patients and desktop medicine. *Health affairs*, 36(4):655–662.

Jimmy Tobin, Qisheng Li, Subhashini Venugopalan, Katie Seaver, Richard Cave, and Katrin Tomanek. 2022. Assessing ASR Model Quality on Disordered Speech using BERTScore. In *Proc. 1st Workshop on Speech for Social Good (S4SG)*, pages 26–30.

Subhashini Venugopalan, Joel Shor, Manoj Plakal, Jimmy Tobin, Katrin Tomanek, Jordan R. Green, and Michael P. Brenner. 2021. Comparing Supervised Models and Learned Speech Representations for Classifying Intelligibility of Disordered Speech on Selected Phrases. In *Proc. Interspeech 2021*, pages 4843–4847.

Ye-Yi Wang, A. Acero, and C. Chelba. 2003. Is word error rate a good indicator for spoken language understanding accuracy. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*, pages 577–582.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. *arXiv e-prints*, page arXiv:1904.09675.

Medical Visual Textual Entailment for Numerical Understanding of Vision-and-Language Models

Hitomi Yanaka*, **Yuta Nakamura***, **Yuki Chida, Tomoya Kurosawa**
the University of Tokyo
{hyanaka,yuki.chida,kurosawa-tomoya}@is.s.u-tokyo.ac.jp
yutanakamura-tky@umin.ac.jp

Abstract

Assessing the capacity of numerical understanding of vision-and-language models over images and texts is crucial for real vision-and-language applications, such as systems for automated medical image analysis. We provide a visual reasoning dataset focusing on numerical understanding in the medical domain. The experiments using our dataset show that current vision-and-language models fail to perform numerical inference in the medical domain. However, the data augmentation with only a small amount of our dataset improves the model performance, while maintaining the performance in the general domain.

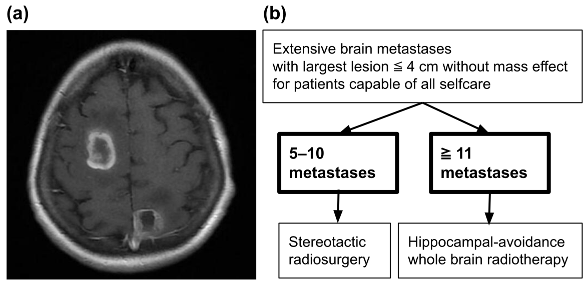


Figure 1: The practical example of the need for visual reasoning in the medical domain. (a) A magnetic resonance imaging (MRI) image showing two brain metastases[1]. (b) Treatment strategy depending on the lesion number of brain metastases (modified from Gondi et al. (2022), not for medical use).

1 Introduction

Vision-and-language models have made great progress on complex tasks, going beyond image recognition and towards reasoning over images and texts (Antol et al., 2015; Xie et al., 2019; Suhr et al., 2019). Following the success of pre-trained language models (Devlin et al., 2019, inter alia), recent advances in vision-and-language models have been made by the introduction of large-scale pretraining (Li et al., 2019; Kim et al., 2021; Singh et al., 2022). However, as with pre-trained language models, it is unclear what information pretrained vision-and-language models learn and use in their predictions, and what their limitations are.

While a large body of research (Naik et al., 2018; Rozen et al., 2019; Ravichander et al., 2019; Richardson et al., 2020) has provided challenging reasoning tasks to probe the reasoning ability of pre-trained language models, such work has been more limited for vision-and-language models. Furthermore, previous visual reasoning datasets are usually provided by the general domain of images, and analysis across different domains is desirable.

Our focus is to investigate whether current vision-and-language models have the ability to infer numerical relationships between images and texts in the medical domain, which is crucial for real vision-and-language applications such as systems for automated medical image analysis. Consider the example of images and textual descriptions in a medical article presented in Figure 1. The lesion number affects the treatment strategy for diseases such as brain metastasis. If systems can automatically judge whether the lesion number in given images matches that in arbitrary query texts, they can support medical decision-making. Recently, a vision-and-language model focusing on the medical domain (Delbrouck et al., 2022) has begun to be provided but is not yet fully developed.

With this motivation, we provide a visual reasoning dataset focusing on numerical inference in the medical domain by adding annotations to the previous medical image and caption dataset MedICaT (Subramanian et al., 2020). We call our dataset MedVTE, which will be publicly available at `https://github.com/ynklab/MedVTE`. Using MedVTE, we investigate the extent to which current pre-trained vision-and-language models have the ability of numerical

*Equal Contribution.

[1] https://radiopaedia.org/cases/haemorrhagic-intracranial-metastases-from-breast-cancer

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 8–18
July 14, 2023

understanding on visual reasoning tasks across images and texts in the medical domain. The experiments show that current models have much room to perform numerical inference in the medical domain.

2 Background

2.1 Vision-and-language understanding

Regarding standard vision-and-language understanding tasks, SNLI-VE (Xie et al., 2019) is a large general domain dataset for the Visual Textual Entailment (VTE) task. The dataset consists of image-sentence pairs annotated with a three-class label (*entailment*, *contradiction*, or *neutral*), indicating whether a premise image entails a hypothesis sentence. There have been studies investigating the counting ability of vision-and-language models on visual question-answering tasks and object detection tasks (Chattopadhyay et al., 2017; Zhang et al., 2018; Song and Qiu, 2018; Trott et al., 2018; Acharya et al., 2019; Parcalabescu et al., 2021). However, since previous studies only use datasets in the general domain, it is unclear the extent to which models can maintain the ability to understand numerical expressions in the medical domain.

For visual reasoning in the medical domain, Li et al. (2020) have compared the performance of four pre-trained vision-and-language models and traditional CNN-RNN models on two datasets of thoracic findings classification tasks: MIMIC-CXR (Johnson et al., 2019) and OpenI datasets. The results showed that the pre-trained models outperformed the traditional models. Our VTE dataset gives a fine-grained analysis of the capacity of the pre-trained vision-and-language models for numerical understanding in the medical domain.

2.2 Clinical NLP

Clinical NLP is one of the practical fields of NLP, and various reasoning tasks in the medical domain have been provided. For sentence-level language understanding tasks, emrQA (Pampari et al., 2018) is a large-scale QA dataset on electronic medical records, and MedSTS (Wang et al., 2020) is a resource for Semantic Textual Similarity (STS) tasks in the medical domain. The most related dataset to ours is MedNLI (Romanov and Shivade, 2018), a physician-annotated Natural Language Inference (NLI) dataset with premises extracted from clinical notes. However, a recent study has reported annotation artifacts in MedNLI (Herlihy and Rudinger, 2021). To avoid such undesired artifacts, we cover a variety of numerical expressions.

3 MedVTE Datasets

We introduce MedVTE, visual reasoning datasets in the medical domain involving numerical expressions. MedVTE is composed of pairs of medical images, captions, and three-class entailment labels (*entailment*, *contradiction*, or *neutral*). MedVTE focuses on the relationship between the number of lesions, such as cancer in an image and the numerical expression in a text.

We created MedVTE by selecting examples involving numerical expressions from MedICaT dataset (Subramanian et al., 2020). MedICaT contains 217,060 figure-caption pairs in medical articles, whose captions sometimes refer to the number of the depicted lesions (e.g., tumors or nodules). The selection is conducted by one medical expert.

3.1 Premise–hypothesis collection

In MedVTE, a premise is a MedICaT figure, and a hypothesis is one complete sentence containing one or more lesion numbers. We created 409 examples for the MedVTE dataset in total.

Step 1. Cleaning We removed 58 MedICaT figure–caption duplicate pairs. We also mitigated occasional errors in MedICaT captions, such as missing letters or interrupted sentences. Some MedICaT captions are provided in two versions, the one by the MedICaT authors and the other from the S2ORC dataset (Lo et al., 2020). In such cases, we always chose the longer one to avoid including incomplete sentences.

Step 2. Figure collection We collected MedICaT figure–caption pairs whose captions include lesion numbers in a rule-based approach. We assigned Penn Treebank part-of-speech (POS) tags (Marcus et al., 1993) to all captions. We then applied a spaCy rule-based matcher to accept only captions having a numeral followed by a noun suggesting lesions. This step left us 431 figure–caption pairs. See Appendix A for details.

Step 3. Hypothesis collection Every MedVTE hypothesis is a single sentence including one or more lesion numbers. We collected hypotheses by splitting captions into sentences and selecting sentences containing at least one lesion number.

Sentence selection was performed in a rule-based approach as in Step 2 followed by manual

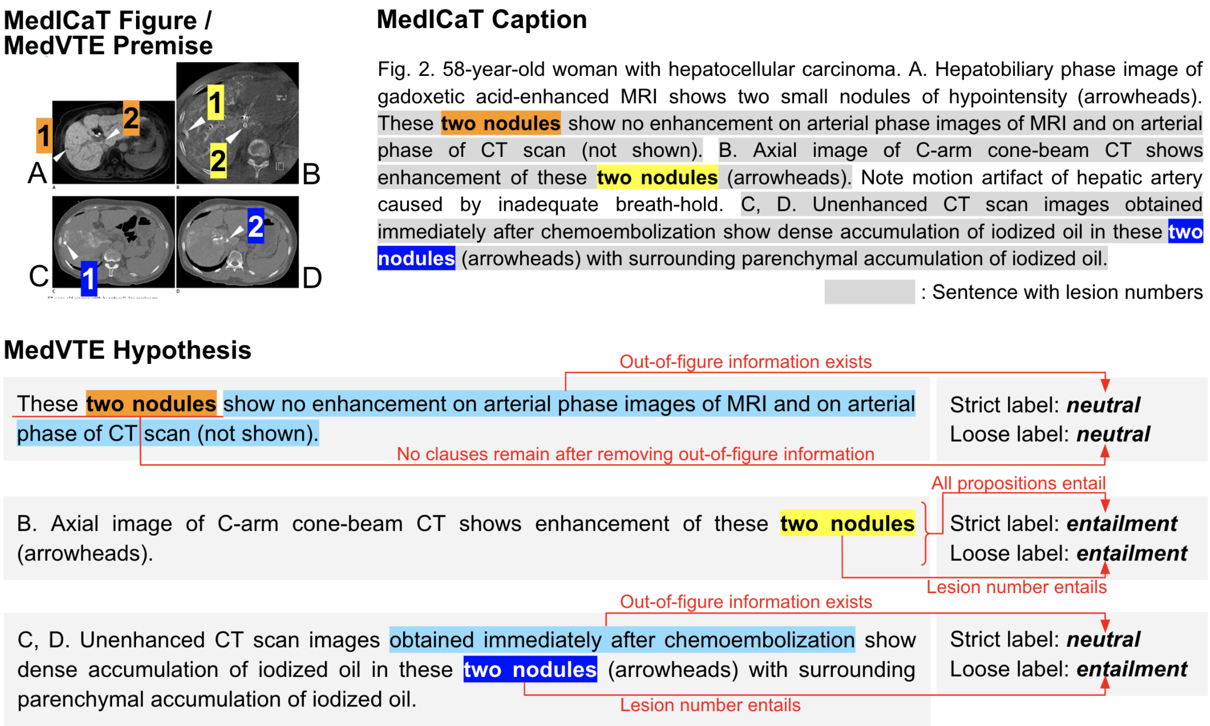


Figure 2: MedVTE examples. Premises are MedICaT figures and hypotheses are MedICaT caption sentences containing numerical expressions of lesions. For each hypothesis, the strict label considers all information, and the loose label is only determined by comparing lesion numbers. Corresponding lesion numbers are colored in orange, yellow, and dark blue. Light blue spans indicate out-of-figure information, which is beyond the figure's scope and deemed unverifiable by the medical expert based on the figure alone.

reviews. In manual reviews, we removed erroneous lesion numbers where integers do not actually count lesions, such as cell line names *Walker 256 tumor*. We also excluded invalid premise figure-hypothesis sentence pairs meeting the below criteria:

- the figure file contains multiple article figures
- the hypothesis is not a single sentence
- the hypothesis does not make sense due to ungrammaticality.

When multiple hypothesis sentences corresponded to a single premise figure, we treated each premise figure-hypothesis sentence pair as an independent sample. We obtained 409 premise-hypothesis pairs for 373 premise figures, where 430 lesion numbers appear in total.

3.2 Labeling

We assigned two types of entailment labels, *strict labels* and *loose labels*, to premise-hypothesis pairs on MedVTE.

Strict labels follow the common practice of annotating visual reasoning datasets to compare all the information, not only numerical one but also medical background knowledge, of a premise figure and a hypothesis sentence. However, we found that the considerable number of strict labels became *neutral* under given images because out-of-figure information in hypothesis sentences (i.e., information that is not acquired from images), such as "this image was obtained six months after surgery," is necessary to judge their labels as *entailment*.

To realize separate assessments of the numerical reasoning abilities of models under only given

Models	Train \ Test	SNLI-VE	MVTEl	MVTEs
ViLT	SNLI-VE	0.757	0.243	0.290
	+MVTEl	0.757	0.443	0.366
	+MVTEs	0.745	0.371	0.416
FLAVA	SNLI-VE	0.790	0.236	0.281
	+MVTEl	0.791	0.428	0.356
	+MVTEs	0.791	0.355	0.408

Table 1: F1-macro scores of each baseline model and dataset. MVTEl and MVTEs indicate MedVTE annotated with loose labels and strict labels, respectively. +MVTEl indicates SNLI-VE mixed with MVTEl.

images, we add loose label annotations rather than editing hypothesis sentences. Loose labels only compare numerical information of a premise figure and a hypothesis sentence. This approach provides an option to focus on numerical reasoning abilities with loose labels, or to fully measure medical reasoning abilities with strict labels, which requires expert knowledge to recognize out-of-figure information.

The following is the definition of loose labels. Details are available in Appendix C.

- *entailment*: All lesion numbers are consistent with the premise figure
- *contradiction*: One or more lesion numbers are smaller than those depicted in the premise figure
- *neutral*: Either of the following is satisfied: (i) one or more lesion numbers are larger than those depicted in the premise figure although the others are consistent, (ii) the number of lesion numbers cannot be determined only from the premise figure, or (iii) no clauses remain after removing out-of-figure information from the hypothesis.

Figure 2 shows MedVTE examples. In the top and middle examples, their loose labels are the same as their strict labels. In the bottom example, its loose label is different from its strict label with the consideration of out-of-figure information. The distribution of loose labels in MedVTE is $(entailment, neutral, contradiction) = (310, 95, 4)$, and that of strict labels is $(entailment, neutral, contradiction) = (208, 197, 4)$.

4 Experiments and Analysis

4.1 Experimental setup

Models Vision-and-language models are categorized into three broad types based on their encoding style, fusion encoder, dual encoder, and a combination of both. We used two vision-and-language models for our experiments: a Vision-and-Language Transformer model (ViLT) (Kim et al., 2021) and a Foundational Language And Vision Alignment model (FLAVA) (Singh et al., 2022). ViLT is a fusion-encoder style model which has 112M parameters. FLAVA is a fusion-encoder plus dual-encoder style model which has 243M parameters. See details of pre-training datasets for each model in Appendix D.

Training For baseline models, we use vision-and-language models fine-tuned with the training set of SNLI-VE. We split the MedVTE dataset as train:test=306:103 and evaluate the performance of the models on the MedVTE test set. To investigate whether a small portion of additional training data in the medical domain contributes to knowledge transfer for visual reasoning, we evaluate models fine-tuned with the SNLI-VE training set mixed with the MedVTE training set. We fine-tune the models for three epochs for each dataset and use F1-macro scores for evaluation metrics. Details on the hyperparameters can be found in Appendix D.

4.2 Baseline results

Table 1 shows baseline results. While both ViLT and FLAVA models trained with SNLI-VE achieved around 75% on in-domain SNLI-VE, their performance was very low on MedVTE.

When we evaluated models trained with SNLI-VE mixed with a subset of MedVTE, the performance on MedVTE was improved while maintaining the performance on SNLI-VE. However, the overall performance on MedVTE was still lower than 50%. This indicates that numerical inference in the medical domain is challenging for vision-and-language models even when they train with a subset of MedVTE. Regarding the difference between loose labels and strict labels with a subset of MedVTE, the performance improvement on MedVTE strict labels was lower than that on loose labels. This suggests that the ability to use out-of-figure information is difficult to obtain from the data augmentation.

5 Conclusion

We created the visual reasoning dataset MedVTE, focusing on numerical understanding in the medical domain. The experiments using MedVTE showed that current vision-and-language models struggled with performing numerical inference in the medical domain. However, the data augmentation with only a small amount of our MedVTE dataset improved the model performance, while maintaining the performance in the general domain. In future work, we increase the size of our MedVTE dataset and make further analysis of vision-and-language models to investigate the extent to which the size of a fine-tuning dataset affects the performance of numerical inference in the medical domain. Improving automated numerical vision-and-language understanding in the medical domain could aid therapeutic decision-making that depends on lesion numbers.

6 Limitation

Since hypothesis sentences were created and labeled by medical experts, the size of our current dataset is small. In particular, the number of examples of contradiction is small because the hypothesis sentences were created based on captions to efficiently construct our dataset. However, we can increase the number of examples of contradiction by rewriting phrases in the hypothesis sentences. The claim of this study is that we can relatively efficiently create a VTE dataset in the medical domain from the existing image caption dataset, and can empirically demonstrate the challenges of current vision-and-language models on the VTE dataset. Although increasing the data size is an important next step, it is beyond the scope of this paper.

Acknowledgements

We thank the three anonymous reviewers for their helpful comments and suggestions, which improved this paper. This work was supported by JST, PRESTO grant number JPMJPR21C8, Japan.

References

Manoj Acharya, Kushal Kafle, and Christopher Kanan. 2019. TallyQA: Answering complex counting questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8076–8084.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Prithvijit Chattopadhyay, Ramakrishna Vedantam, Ramprasaath R. Selvaraju, Dhruv Batra, and Devi Parikh. 2017. Counting everyday objects in everyday scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jean-benoit Delbrouck, Khaled Saab, Maya Varma, Sabri Eyuboglu, Pierre Chambon, Jared Dunnmon, Juan Zambrano, Akshay Chaudhari, and Curtis Langlotz. 2022. ViLMedic: a framework for research at the intersection of vision and language in medical AI. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 23–34.

Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. 2021. RedCaps: Web-curated image-text data created by the people, for the people. In *Proceedings of the NeurIPS Datasets and Benchmarks*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Vinai Gondi, Glenn S. Bauman, Lisa Bradfield, Stuart H. Burri, Alvin R. Cabrera, Danielle A. Cunningham, Bree R Eaton, Jona A. Hattangadi-Gluth, Michelle M. Kim, Rupesh R. Kotecha, Lianne Kraemer, Jing Li, Seema Nagpal, Chad G Rusthoven, John H. Suh, Wolfgang A. Tomé, Tony J.C. Wang, Alexandra S. Zimmer, Mateo Ziu, and Paul D. Brown. 2022. Radiation therapy for brain metastases: An astro clinical practice guideline. *Practical radiation oncology*, 12(4):265–282.

Christine Herlihy and Rachel Rudinger. 2021. MedNLI is not immune: Natural language inference artifacts in the clinical domain. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1020–1027, Online. Association for Computational Linguistics.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(317):2052–4463.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. ViLT: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 5583–5594.

Nikita Kitaev, Steven Cao, and Dan Klein. 2019. Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505, Florence, Italy. Association for Computational Linguistics.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings* *of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li an Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Yikuan Li, Hanyin Wang, and Yuan Luo. 2020. A comparison of pre-trained vision-and-language models for multimodal representation learning across medical images and reports. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1999–2004.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. 2011. Im2Text: Describing images using 1 million captioned photographs. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, page 1143–1151, Red Hook, NY, USA. Curran Associates Inc.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium. Association for Computational Linguistics.

Letitia Parcalabescu, Albert Gatt, Anette Frank, and Iacer Calixto. 2021. Seeing past words: Testing the cross-modal capabilities of pretrained V&L models on counting tasks. In *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*, pages 32–44, Groningen, Netherlands (Online). Association for Computational Linguistics.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. Connecting vision and language with localized narratives. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 647–664.

Abhilasha Ravichander, Aakanksha Naik, Carolyn Rose, and Eduard Hovy. 2019. EQUATE: A benchmark evaluation framework for quantitative reasoning in natural language inference. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 349–361, Hong Kong, China. Association for Computational Linguistics.

Kyle Richardson, Hai Hu, Lawrence Moss, and Ashish Sabharwal. 2020. Probing natural language inference models through semantic fragments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8713–8721.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium. Association for Computational Linguistics.

Ohad Rozen, Vered Shwartz, Roee Aharoni, and Ido Dagan. 2019. Diversify your datasets: Analyzing generalization via controlled variance in adversarial datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 196–205, Hong Kong, China. Association for Computational Linguistics.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. FLAVA: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15617–15629.

Zichen Song and Qiang Qiu. 2018. Learn to classify and count: A unified framework for object classification and counting. In *Proceedings of the 2018 International Conference on Image and Graphics*

Processing, page 110–114. Association for Computing Machinery.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. WIT: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2443–2449, New York, NY, USA. Association for Computing Machinery.

Mitchell Stern, Jacob Andreas, and Dan Klein. 2017. A minimal span-based neural constituency parser. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 818–827, Vancouver, Canada. Association for Computational Linguistics.

Sanjay Subramanian, Lucy Lu Wang, Ben Bogin, Sachin Mehta, Madeleine van Zuylen, Sravanthi Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. 2020. MedICaT: A dataset of medical images, captions, and textual references. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2112–2120, Online. Association for Computational Linguistics.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy. Association for Computational Linguistics.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl S. Ni, Douglas N. Poland, Damian Borth, and Li-Jia Li. 2015. YFCC100M: the new data in multimedia research. *Commun. ACM*, 59:64–73.

Alexander Trott, Caiming Xiong, and Richard Socher. 2018. Interpretable counting for visual question answering. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Yanshan Wang, Naveed Afzal, Sunyang Fu, Liwei Wang, Feichen Shen, Majid Rastegar-Mojarad, and Hongfang Liu. 2020. MedSTS: a resource for clinical semantic textual similarity. *Language Resources and Evaluation*, 54(1):57–72.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*.

Yan Zhang, Jonathon Hare, and Adam Prügel-Bennett. 2018. Learning to count objects in natural images for visual question answering. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

A Sample selection rules

This section explains detailed MedVTE sample selection rules.

We employed rule-based approaches to select figure-caption pairs from the MedICaT dataset so that all sampled captions refer to the number of lesions.

We selected sentences in the MedICaT captions containing LESION-NUMBER-EXPRESSIONS. We defined a LESION-NUMBER-EXPRESSION as any token subsequence of a single sentence of a caption that satisfies all of the following *Rules 1* to *3*:

- *Definition 1*. COMPARATIVE is a string whose lowercase form is either *at least*, *at most*, *more than*, or *less than*.
- *Definition 2*. NUMBER is a single token whose Penn Treebank part-of-speech (POS) tag (Marcus et al., 1993) is CD (cardinal number).
- *Definition 3*. LESION-NOUN is a single token whose POS tag is either NN (noun, singular or mass) or NNS (noun, plural).
- *Rule 1*. A LESION-NUMBER-EXPRESSION must be a concatenation of COMPARATIVE, NUMBER, and LESION-NOUN in this order, or a concatenation of NUMBER and LESION-NOUN in this order.
- *Rule 2*. The lemma of LESION-NOUN must be either *cancer*, *lesion*, *mass*, *metastasis*, *nodule*, or *tumor*.
- *Rule 3*. A LESION-NUMBER-EXPRESSION must not appear immediately after a token whose lowercase form is either *figure, figures, fig, figs, patient, case, day, sample, type, category, group, grade, level, stage, rads, pirads, birads, cin, score, likert, c, t, l, s, segment, gs, suv, +, +1, +2, +3, +4, +5, mm, cm, mm2, cm2, mm3,* or *cm3*.

In our implementation, we first assigned POS tags to all MedICaT captions using Berkeley Neural Parser (Stern et al., 2017; Kitaev and Klein, 2018; Kitaev et al., 2019). We then built a spaCy rule-based matcher and applied it to all parsing results.

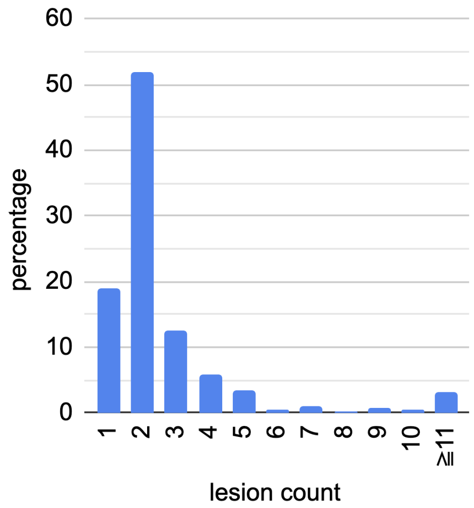


Figure 3: Distribution of the quantity of 424 of the 430 lesion numbers in the MedVTE hypotheses. Note that the remaining six lesion numbers are excluded because they appear immediately after a comparative expression such as “at least” or “more than.”

B Dataset statistics

Of the 409 MedVTE premise-hypothesis pairs, 300 (73.3%) have radiological premise figures, twelve (2.9%) have scopic premise figures, and the remaining have other various types of premise figures including histopathological images.

MedVTE contains 430 lesion numbers in total because three of the 409 hypotheses (0.7%) contain three lesion numbers, fifteen hypotheses (3.7%) contain two lesion numbers, while the remaining 391 hypotheses (95.6%) contain one lesion number.

Six of the 430 lesion numbers (1.4%) include comparative expressions, four of which are associated with “at least” and the others are accompanied by “more than.” Figure 3 shows the distribution of the remaining 424 lesion numbers. The most frequent lesion number is two, occurring 223 times in the dataset (52.6%). 398 lesion numbers (92.6%) are between one and five, and fourteen lesion numbers (3.3%) are greater than ten.

C Details of labeling

C.1 Loose labels

Each MedVTE premise image consists of one or more subfigures that are often excerpts of a vast series of radiological, pathological, or endoscopic images. Therefore, it must be considered that the premise image may not reflect the entire patient and may contain only a subset of the lesions that are actually present, or conversely, the same lesion

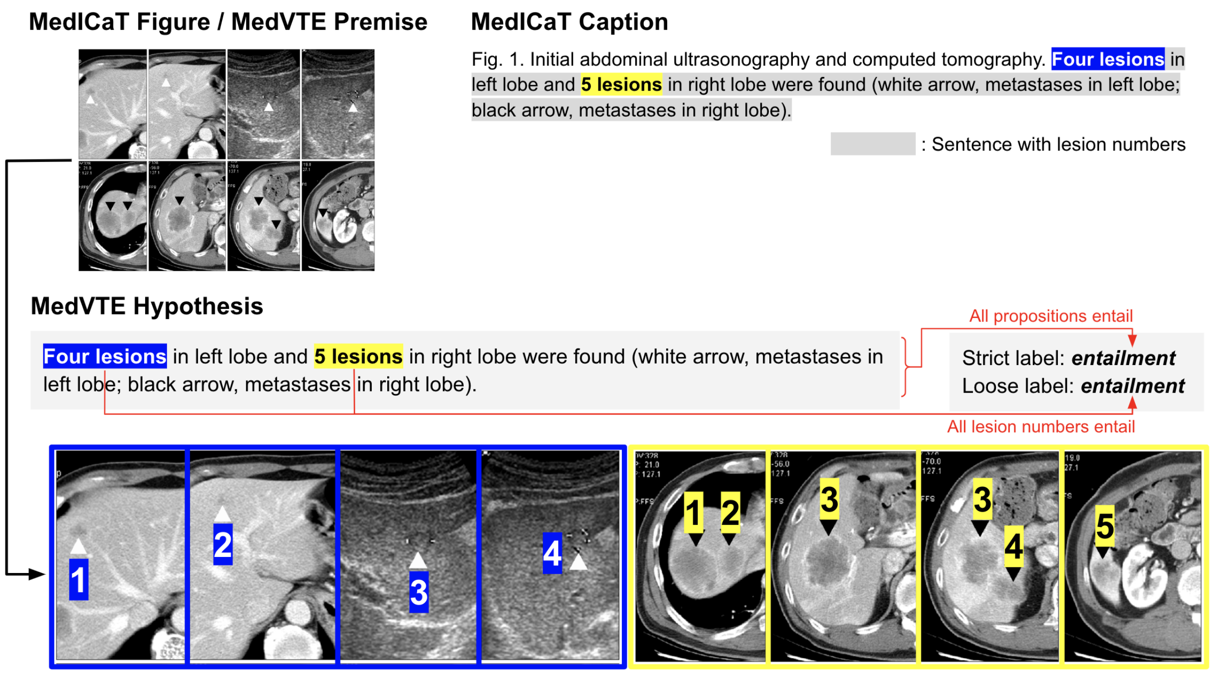


Figure 4: Another example of MedVTE. The four subfigures outlined in yellow apparently have *six* lesions. However, the medical expert has determined that the yellow subfigures demonstrate *five* lesions and assigned *entailment* label because it is explainable that the lesion numbered “3” repeatedly appears in the second and third subfigures at the different levels.

may repeatedly appear across multiple subfigures as in Figure 4. This phenomenon is prevalent not only in the medical articles from which MedVTE originates but also in the real-world clinical practice that we target for application.

We regard each hypothesis as a set of propositions. For each proposition addressing the lesion number in the hypothesis sentence, the following procedure was employed to determine the veracity or falsity.

(a) If the medical expert determines that the quantities are equal in the premise figure and the hypothesis sentence, the proposition is supported.

(b) When the lesion number in the hypothesis sentence apparently exceeds that in the premise figure, the medical expert is requested to carefully review the premise figure and determine if the gap can be explained by the following reason:

- The original caption is correct, but the medical expert initially missed some lesions due to subtle image findings.

If so, the hypothesis is supported. Otherwise, the loose label is *neutral* because it is impossible to judge which of the following is happening:

- The original caption is correct, but the premise figure does not show all the lesions
- The original caption has overcounted the lesions.

(c) When the lesion number in the hypothesis sentence appears to be smaller than the premise figure, the medical expert is asked to examine the premise figure again and determine which of the following is the most convincing:

- The original caption is correct, but the medical expert initially overcounted the lesions due to equivocal image findings
- The original caption is correct, but the medical expert initially overcounted the lesions due to the same lesion repeatedly appearing across multiple subfigures
- The original caption has undercounted the lesions.

In the first or second case, the hypothesis is supported. In the last case, the loose label is *contradiction*.

C.2 Strict labels

When a hypothesis contains propositions that cannot be judged true or false from the premise image alone, we consider it out-of-figure information. The following are examples of propositions that we regard as out-of-figure information:

- Mention to other figures than the premise figure (e.g., "show no enhancement on arterial phase images of MRI and on the arterial phase of CT scan (not shown)")
- Numerical values for elapsed time, such as days, months, or years (e.g., "Axial contrast-enhanced CT *six weeks* pre-RF ablation (a) demonstrates two lesions")
- Specific lesion size numbers (e.g., "The two nodules were *1.2 cm* in diameter").

If the hypothesis sentence includes out-of-figure information, we set the strict label to *neutral* regardless of the loose label. Otherwise, the strict label is the same as the loose label.

D Model details

ViLT is pre-trained on MSCOCO (Lin et al., 2014)+VG (Krishna et al., 2017)+CC (Sharma et al., 2018)+SBU (Ordonez et al., 2011). FLAVA is pre-trained on filtered YFCC100M (Thomee et al., 2015)+CC12M (Changpinyo et al., 2021)+WIT (Srinivasan et al., 2021)+RedCaps (Desai et al., 2021)+LN (Pont-Tuset et al., 2020)+MSCOCO+VG+CC+SBU.

We basically adopted models and parameters implemented in transformers[2]. We attached a 2-layer classifier head ourselves for FLAVA since there was no model implementation for classification tasks in the library. Table 2 and Table 3 show hyperparameters in ViLT and FLAVA models, respectively.

Hyperparameter	Value
Encoder	
hidden size	768
number of heads	12
number of layers	12
intermediate size	3072
dropout probability	0
patch size	32×32
input image size	384×640
Classifier Head	
hidden size	768
Others	
text vocabulary size	30522
Training	
epochs	3
gradient accumulation steps	3
per device batch size	48
learning rate	5e-05
AdamW weight decay	0
AdamW β_1	0.9
AdamW β_2	0.999

Table 2: Hyperparameters in ViLT

[2]https://huggingface.co/docs/transformers/v4.20.1/en/index

Hyperparameter	Value
Image Encoder	
hidden size	768
number of heads	12
intermediate size	3072
number of layers	12
dropout probability	0
patch size	16×16
input image size	224×224
Text Encoder	
hidden size	768
number of heads	12
intermediate size	3072
number of layers	12
dropout probability	0
Multimodal Encoder	
hidden size	768
number of heads	12
intermediate size	3072
number of layers	6
dropout probability	0
Classifier Head	
hidden size	1536
Others	
text vocabulary size	30522
image dVAE codebook size	8192
Training	
epochs	3
gradient accumulation steps	3
per device batch size	24
learning rate	1e-05
learning rate schedule	linear
warmup updates	2000
AdamW weight decay	1e-02
AdamW β_1	0.9
AdamW β_2	0.999

Table 3: Hyperparameters in FLAVA

Privacy-Preserving Knowledge Transfer through Partial Parameter Sharing

Paul Youssef Jörg Schlötterer Christin Seifert
University of Duisburg-Essen
University of Marburg
{firstname.lastname}@uni-marburg.de

Abstract

Valuable datasets that contain sensitive information are not shared due to privacy and copyright concerns. This hinders progress in many areas and prevents the use of machine learning solutions to solve relevant tasks. One possible solution is sharing models that are trained on such datasets. However, this is also associated with potential privacy risks due to data extraction attacks. In this work, we propose a solution based on sharing parts of the model's parameters, and using a proxy dataset for complimentary knowledge transfer. Our experiments show encouraging results, and reduced risk to potential training data identification attacks. We present a viable solution to sharing knowledge with data-disadvantaged parties, that do not have the resources to produce high-quality data, with reduced privacy risks to the sharing parties. We make our code publicly available.[1]

1 Introduction

NLP research in many areas (e.g., healthcare) is hindered by the unavailability of publicly-available datasets. Even though such datasets might be available for some researchers, sharing them with the community is problematic in many cases due to privacy and copyright concerns (Liu et al., 2021).

De-identifying sensitive information in such datasets is a potential option. However, depending on the nature of the data, the utility of the data might be negatively affected (Jordon et al., 2021) when de-identifying the data. Sharing a model that is trained on the data instead of directly sharing the data itself is another option (Lehman et al., 2021). The shared model transfers knowledge gained from raw data and is beneficial in many cases (e.g., when an institute is interested in solving the same task, but lacks enough data). However, sharing the model is also associated with potential re-identification risks (Carlini et al., 2021).

Instead of directly sharing models or data, data-free knowledge distillation (DF-KD) aims to transfer the knowledge from a large teacher model to a smaller student model without relying on any task-specific data, i.e., data that has been used to train the teacher model. Instead, many approaches make use of a proxy dataset (Krishna et al., 2020) to facilitate the knowledge transfer.

In this work, we propose a solution to the problem of sharing knowledge between models in a privacy-preserving manner. Our solution depends on sharing parts of the model, and using a proxy dataset for complementary knowledge transfer. Partially sharing the model mitigates potential privacy risks. Further training on a proxy dataset helps compensating the loss caused by the absence of the non-shared parts of the model.

We experiment on two datasets for text classification from the clinical domain, **AP** (Gao et al., 2023) for relation classification and **MedNLI** (Romanov and Shivade, 2018) for natural language inference, and show that our approach substantially improves the performance of a student model trained only on a proxy dataset. Additionally, we show that the resulting model cannot be leveraged to reliably identify the original training data.

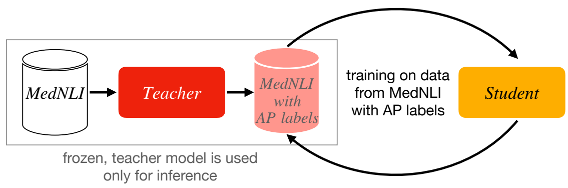


Figure 1: The process of using the proxy dataset, **MedNLI**, to indirectly train the student model on the target task, **AP**. Left: **MedNLI** is labeled with **AP** labels using a teacher model that was trained on **AP** before. Right: Training the student model with the proxy dataset, i.e., **MedNLI** inputs and **AP** labels.

[1] https://github.com/paulyoussef/ppkt/

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 19–23
July 14, 2023 ©2023 Association for Computational Linguistics

2 Related Work

Knowledge distillation (KD). The goal of knowledge distillation is to transfer knowledge from a large teacher model to a student model of a smaller size. Hinton et al. (2015) propose training the student model such, that its output distribution matches the output distribution of the teacher. In order to distill knowledge from BERT (Devlin et al., 2019) into a smaller transformer architecture, Sanh et al. (2019) additionally use the masked language modeling loss used to pre-train BERT and a cosine embedding loss in order to make the hidden representations from both models more similar on the original pre-trainig corpus of BERT. Haidar et al. (2022) randomly choose two intermediate layers from the teacher and the student and train the student's layer to produce similar representations to that of the teacher. In our method, we make use of the teacher's hard predictions, and do not assume access to its outputs distribution.

Data-free knowledge distillation (DF-KD). Even though the teacher's training data can be used in KD, the DF-KD setting assumes the unavailability of such data. Lopes et al. (2017) aim to reconstruct the teacher's training set using the teacher's activation records on the same data. Rashid et al. (2021) use an adversarial generator to generate out-of-domain data, on which the teacher and student disagree the most, and then use this data to train the student. Krishna et al. (2020) show that it is possible to extract a model using its predictions on nonsensical data, but put no restrictions on the size of the model. Our work assumes the availability of a proxy dataset from a related task and that the teacher and the student share the same architecture.

Data extraction from language models. Carlini et al. (2021) show that it is possible to extract training data from GPT-2 (Radford et al., 2019). Huang et al. (2022) experiment on GPT-Neo (Gao et al., 2020) and show that it could leak sensitive information, but the chances of extracting information about a specific user are small because of the model's weak association abilities. Similar work that targets BERT (Vakili and Dalianis, 2021; Lehman et al., 2021) suggests that extracting sensitive information from BERT is unlikely, but robustness against more sophisticated attacks cannot be guaranteed. Membership inference attacks, that aim to identify whether certain data instances have been used to train the model, show some success against BERT (Shejwalkar et al., 2021). We conduct a membership inference attack, in order to inspect if the student models we produce can be used to identify the teacher's training examples.

3 Problem Statement

Let T be a teacher model, trained for a specific task $target$ on training data D_{target} and S be a student model with the same architecture, but untrained. We are interested in transferring the knowledge captured by T on D_{target} to S without providing S any access to D_{target}. Ideally, S cannot be used to identify any data from $target$. S can be trained on any data that does not come from the same distribution as D_{target}. We refer to such data as D_{proxy}. T can provide predictions on D_{proxy} based on what it has learned on D_{target}. We measure the performance of both, T and S, using a held-out test set from $target$, which we refer to as D'_{target}.

4 Method

Our method for transferring knowledge from T to S without using any task-specific data, consists of two parts: 1) partial parameter sharing, 2) finetuning on a proxy dataset.

Partial parameter sharing. Since T and S have the same architecture, we copy parameters from N non-adjacent layers of T, and use them directly in the corresponding position in S, in order to facilitate knowledge transfer from T to S. We consider sharing only non-adjacent layers from T, since having several consecutive layers in their initialized state might result in representations of lower quality. We keep the parameters from T fixed during the later finetuning step to avoid degrading to parameters of lower quality. Since the parameters from T reflect a compressed version of the data, we conjecture that partially sharing them provides S only with a distorted and partial view of D_{target}.

Finetuning on a proxy dataset. Sharing parameters in the first step only affects N layers from S, the rest of the layers in S are kept in their state from pre-training, and the task-specific parameters are randomly initialized. In order to make these layers contribute to the knowledge transfer as well, we finetune the model using the proxy dataset D_{proxy}. Note that D_{proxy} contains data that are not part of $target$, but that are artificially labeled using T. Hence, D_{proxy} can be unlabeled. This process is depicted in Figure 1. We only use hard predictions from T, i.e., we only use the class with the highest probability as label and do not use T's probability distribution over all classes. We leave experimenting with T's probability distribution over all classes for future work. To train the student model, we use the cross-entropy loss:

$$L_{CE} = -\sum_{c=1}^{C} y_{t,c} \log(y_{s,c}) \quad (1)$$

where C is the number of classes, $y_{t,c} \in \{0, 1\}$ is the teacher's prediction, indicating if the input belongs to the c-th class or not, and $y_{s,c}$ is the students' model probability for class c.

5 Experimental Setup

In this section, we describe the data and the experiments we design to evaluate our proposed method for knowledge transfer.

5.1 Data

We use two datasets in our experiments. The first one, AP, acts as the target task, whose data should be kept private. The second dataset, MedNLI, is larger and we use it as a proxy dataset to transfer knowledge from the teacher model. Table 1 provides statistics on both datasets, and Table 2 shows an example from each dataset.

The **Assessment and Plan Relation Labeling (AP)** (Gao et al., 2023) dataset is based on clinical notes from MIMIC-III v1.4 (Johnson et al., 2016). Each instance consists of an assessment that describes the current state of the patient and her active health problems, a plan that handles a specific problem, and a label that describes the relation between the assessment and the plan (direct, indirect, neither or irrelevant). We set the training and test sets of AP to be D_{target} and D'_{target} respectively, i.e., AP is our target task.

The **Medical Natural Language Inference (MedNLI)** (Romanov and Shivade, 2018) is a dataset for medical language inference. Each instance consists of a premise, a hypothesis and a label belonging to one of three classes (entailment, neutral and contradiction) depending on whether the hypothesis can be entailed from the premise or not. The premise sentences are taken from MIMIC-III v1.3 (Johnson et al., 2016), whereas the hypothesis sentences were generated by clinicians. We set MedNLI to be D_{proxy}, i.e., MedNLI is the proxy dataset, that we label with the teacher, and use for complementary knowledge transfer.

	Training	**Dev**	**Test**	**len_1**	**len_2**
AP	4633	467	667	40	51.0
MedNLI	11232	1395	1422	20	5.8

Table 1: Dataset statistics. len_i refers to the average length of the i-th input in tokens. Note that we do not use the test set of MedNLI, the evaluation is done on AP's test set. We report the size of the test set for completeness.

	AP	
$Input_1$	64M with EtOH cirrhosis, Afib, admit with upper GI bleed...	**Label:** Direct
$Input_2$	Anemia. Predominary acute blood loss	
	MedNLI	
$Input_1$	She has cough with sputum, occasional blood streaks but no gross blood.	**Label:** Contradiction
$Input_2$	The patient has normal lungs	

Table 2: Examples from AP and MedNLI

5.2 Target Task Performance

The goal of this experiment is to compare the performance of the **teacher** model with the performance of several student models:

- **student-none:** a student that depends only on the proxy dataset, MedNLI, to learn the target task.
- **student-3:** a student model with 3 non-adjacent layers from the teacher. We select the first 3 layers with even indices.
- **student-6:** the same as student-3, but with 6 layers instead of 3.

We use BERT base-cased (Devlin et al., 2019), which consists of 12 encoder layers, as a base model for both the teacher and the student. Note that other domain-specific BERT-based models (e.g., BioClinicalBERT (Alsentzer et al., 2019)) perform better on both tasks. However, these models are pre-trained on data from MIMIC, and we wanted to avoid confounding our results by this factor. We initially train the teacher model on the AP training set for 3 epochs, with a learning rate of 5×10^{-5}, store a model checkpoint every 20 steps and select the checkpoint with the highest Macro-F1 on the validation set. Similarly, we finetune the student model for 1 epoch using the proxy train and validation sets after substituting some layers (in the case of **student-3** and **student-6**).

5.3 Training Data Identification

The goal of this experiment is to evaluate to what extent the different student models can be used to re-identify training data from the target task, AP, compared to the teacher model.

We create a synthetic dataset of positives (real training data from AP), and negatives (other data). To keep the task challenging, we create negatives by identifying medical entities in the positive example, and replacing these by other randomly chosen entities of the same type. We use a clinical NER model (Zhang et al., 2021) to annotate the entities of type: problem (e.g., diseases), treatment (e.g., medications), and test (e.g., diagnostic tests). We restrict the number of replacements to 4 in each instance (2 in each input part). Our final dataset consists of 100 positive and 100 negative examples.

We evaluate the capability of the models to identify training data after finetuning on the proxy dataset in case of the student models, and after finetuning on the AP dataset in case of the teacher model. We use the positive and negative examples as input to all models, and extract their respective representations of the `[CLS]` token from the last layer. This representation is often used as an input to a linear layer, which outputs the final predictions in classification tasks in BERT.

After extracting the representations for the positive and negative examples, we train a logistic regression model using 4-fold cross validation to predict whether the provided representations constitute real training data or not. Note, that this setting assumes the availability of labeled data to train the logistic regression model, i.e., access to original training data of the model under attack. However, this data should be difficult to acquire in practice. We follow other authors (e.g., (Shejwalkar et al., 2021)) in assuming the availability of such data.

	AP Performance (Macro-F1)	**Identification** (Accuracy)
majority	11.2	50.00
teacher	**76.9**	**67.40**
student-none	27.1	56.35
student-3	39.0	54.65
student-6	<u>59.3</u>	<u>56.89</u>

Table 3: Performance of all models on AP's test set (section 5.2), and the training data identification task (section 5.3). Majority refers to a majority baseline. The best performing model overall is **bold**. The best performing among the student models is <u>underlined</u>.

6 Results and Discussion

The results for the experiments explained in Sections 5.2 and 5.3 are shown in Table 3. The results show that the teacher model performs the best on AP's test set. This is not surprising, given that the teacher is trained on data that is quite similar to the test data. The gains in performance from training only on the proxy dataset from MedNLI, without sharing any parameters, are limited (see **student-none**). This might be attributed to the fact that the datasets are still different, even though they come from similar tasks (e.g., AP's inputs are much longer than MedNLI's, cf. Table 1). Grafting the student models with parameters from the teacher substantially improves the performance. This is especially apparent as the number of shared layers is increased to six.

However, the good performance of the teacher model on AP makes it more susceptible to the training data identification attack. Indeed, the results in the second column show that the representations from the teacher model are more helpful in identifying the training data than the representations extracted from the student models. The student models in general perform poorly in identifying the real training examples from AP, and their performance is close to that of the majority baseline. This suggests that sharing parameters with student models is harmless, as the representations we extract from them cannot be reliably used to identify the original training data of the teacher.

7 Conclusion

In this work, we presented an approach to tackle knowledge transfer between two parties: a teacher, that is trained on sensitive data, and a student model, that lacks enough data to be trained, but is interested in learning the same task. Our solution depends on the teacher partially sharing some of its parameters with the student, and providing it with predictions on an unlabeled proxy dataset that is different from the target dataset. Our experimental results indicate that the proposed solution is effective in knowledge transfer, and associated with reduced risks to potential training data identification attacks. In future work, we will look into using other model architectures, use more tasks for evaluation, take into account more advanced privacy attacks and consider cross-lingual settings, where the teacher and student use different languages.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *USENIX Security Symposium*, volume 6.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, John Caskey, Brihat Sharma, Matthew M. Churpek, and Majid Afshar. 2023. DR.BENCH: Diagnostic reasoning benchmark for clinical natural language processing. *Journal of Biomedical Informatics*, 138:104286.

Md Akmal Haidar, Nithin Anchuri, Mehdi Rezagholizadeh, Abbas Ghaddar, Philippe Langlais, and Pascal Poupart. 2022. RAIL-KD: RAndom intermediate layer mapping for knowledge distillation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1389–1400, Seattle, United States. Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. 2022. Are large pre-trained language models leaking your personal information? In *Findings of the Association for Computational Linguistics: EMNLP 2022*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

James Jordon, Daniel Jarrett, Evgeny Saveliev, Jinsung Yoon, Paul Elbers, Patrick Thoral, Ari Ercole, Cheng Zhang, Danielle Belgrave, and Mihaela van der Schaar. 2021. Hide-and-seek privacy challenge: Synthetic data generation vs. patient re-identification. In *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *Proceedings of Machine Learning Research*, pages 206–215. PMLR.

Kalpesh Krishna, Gaurav Singh Tomar, Ankur P Parikh, Nicolas Papernot, and Mohit Iyyer. 2020. Thieves on sesame street! model extraction of bert-based apis. In *International Conference on Learning Representations*.

Eric Lehman, Sarthak Jain, Karl Pichotta, Yoav Goldberg, and Byron Wallace. 2021. Does BERT pretrained on clinical notes reveal sensitive data? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 946–959, Online. Association for Computational Linguistics.

Yuang Liu, Wei Zhang, Jun Wang, and Jianyong Wang. 2021. Data-free knowledge transfer: A survey. *arXiv preprint arXiv:2112.15278*.

Raphael Gontijo Lopes, Stefano Fenu, and Thad Starner. 2017. Data-free knowledge distillation for deep neural networks. *ArXiv*, abs/1710.07535.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Ahmad Rashid, Vasileios Lioutas, Abbas Ghaddar, and Mehdi Rezagholizadeh. 2021. Towards zero-shot knowledge distillation for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6551–6561. Association for Computational Linguistics.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Virat Shejwalkar, Huseyin A Inan, Amir Houmansadr, and Robert Sim. 2021. Membership inference attacks against NLP classification models. In *NeurIPS 2021 Workshop Privacy in Machine Learning*.

Thomas Vakili and Hercules Dalianis. 2021. Are clinical bert models privacy preserving? the difficulty of extracting patient-condition associations. In *HUMAN@ AAAI Fall Symposium*.

Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D Manning, and Curtis P Langlotz. 2021. Biomedical and clinical English model packages for the Stanza Python NLP library. *Journal of the American Medical Informatics Association*.

Breaking Barriers: Exploring the Diagnostic Potential of Speech Narratives in Hindi for Alzheimer's Disease

Kritesh Rauniyar[1], **Shuvam Shiwakoti**[1], **Sweta Poudel**[2], **Surendrabikram Thapa**[3], **Usman Naseem**[4], **Mehwish Nasim**[5,6]
[1]Delhi Technological University, India
[2]Kathmandu Engineering College, Tribhuvan University, Nepal
[3]Virginia Tech, USA [4]University of Sydney, Australia
[5]University of Western Australia [6]Flinders University, Australia

Abstract

Alzheimer's Disease (AD) is a neurodegenerative disorder that affects cognitive abilities and memory, especially in older adults. One of the challenges of AD is that it can be difficult to diagnose in its early stages. However, recent research has shown that changes in language, including speech decline and difficulty in processing information, can be important indicators of AD and may help with early detection. Hence, the speech narratives of the patients can be useful in diagnosing the early stages of Alzheimer's disease. While the previous works have presented the potential of using speech narratives to diagnose AD in high-resource languages, this work explores the possibility of using a low-resourced language, i.e., Hindi language, to diagnose AD. In this paper, we present a dataset specifically for analyzing AD in the Hindi language, along with experimental results using various state-of-the-art algorithms to assess the diagnostic potential of speech narratives in Hindi. Our analysis suggests that speech narratives in the Hindi language have the potential to aid in the diagnosis of AD. Our dataset and code are made publicly available at `https://github.com/rkritesh210/DementiaBankHindi`.

1 Introduction

Alzheimer's Disease (AD) is the most typical kind of dementia, characterized by a specific pattern of cognitive and functional deterioration brought on by aging that may eventually lead to death (Soria Lopez et al., 2019). This condition is mostly seen in adults over 60. Hampel et al. (2011) predicted that by 2040, more than 80 million people would be affected by dementia globally, up from an estimated 24 million in 2001.

In the early stages of AD, it is common to experience subtle language impairments such as problems with word finding and comprehension, the use of incorrect words, ambiguous referents, loss of verbal fluency, speaking too much or too loudly, repeating ideas, straying from the topic, which worsens in the moderate and severe stages (Meghanani et al., 2021). This shows that the temporal aspects of spontaneous speech are impacted by this disease. With the advancement of technology, machine learning approaches have been widely applied in the early diagnosis of AD utilizing neuroimaging scans such as Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) (Thapa et al., 2020b). However, this technique for identifying AD patients from Control Normal (CN) is limited to medical personnel (Thapa et al., 2020b). Szatloczki et al. (2015) showed that linguistic analysis could be used to identify AD more accurately than other types of cognitive testing. The temporal features of spontaneous speech, such as speech pace, frequency, and length of pauses, are sensitive detectors of the early stage of the illness, allowing an early and straightforward linguistic screening for AD. Thus, speech might be a straightforward but crucial characteristic that can be utilized to create potent AI models for AD diagnosis.

Through groundbreaking advancements in NLP, machines are now able to comprehend human language with unprecedented accuracy, unlocking a new realm of possibilities for data analysis and knowledge extraction (Naseem et al., 2021). Due to its ability to analyze language patterns and detect small alterations that may signal cognitive deterioration, NLP has grown in prominence in identifying AD (Thapa et al., 2022). NLP has been used to diagnose AD largely in high-resource languages like English. However, there is potential for this approach to be adapted to the low-resource languages in developing countries, including those spoken in India. Therefore the motivation behind this work is to promote the use of automated NLP-based tools for detecting AD in a low-resource language. Such a method will result in an accurate, quick, and economical AD diagnosis. Our contri-

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 24–30
July 14, 2023

butions are as follows:

- A new dataset for the low-resource language, *Hindi*, is created. The dataset (DementiaBankHindi) includes transcripts from 168 patients with Alzheimer's disease (AD) and 98 healthy control normal (CN) participants. The scientific community is expected to benefit from this original dataset.

- Using Hindi transcripts, an NLP-based methodology is given for the early diagnosis of AD patients. We have also explored how various machine translation-based systems perform in diagnosing AD using Hindi transcripts.

2 Related Works

In recent times, there has been significant research in the field of detecting AD using data-driven approaches (Adhikari et al., 2022). The rapid growth of NLP techniques has led to increased utilization of speech and linguistic features for detecting AD. Consequently, machine learning (ML) techniques are extensively employed in this domain. Classical ML-based methods require manual feature engineering. Such feature extraction methods can vary widely for different languages and can get outdated easily for evolving languages (Thapa et al., 2020a). Driven by the limitation of manual feature engineering of classical ML methods for such a diverse and complex task, in more recent times researchers have leveraged deep learning methods for the detection of AD. Karlekar et al. (2018) applied three neural models Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and a stronger CNN-LSTM for detection of AD against CN on transcripts of Dementia bank's dataset of cookie theft picture description. With their initial approach, they were able to achieve an accuracy of 82.8%, 83.7%, and 84.9% for the CNN, LSTM, and CNN-LSTM architecture respectively. But when fed with POS-tagged data, their best-performing model CNN-LSTM model achieved an accuracy of 91.1%. Wang et al. (2021) took a multimodal approach where they also leveraged acoustic features for dementia detection. They used a CNN-attention Network and explored audio-based, text-based, and multimodal approaches using both audio and text-based features. They reported that a multimodal approach (C-Attention-Unified model) using Linguistic features and X-vector (acoustic) features performed best and could detect AD with an accuracy of 77.2% and an F1 score of 0.763.

Work has also been done for the detection of AD in languages other than English. Guo et al. (2020) proposed an autoencoder-based method to augment the Mandarin corpus (Liu et al., 2019) with a larger English dataset from DementiaBank and used a contrastive learning method based on BERT embeddings. With the data augmentation method, they achieved an accuracy of 81.6% in AD prediction. Rentoumi et al. (2017) used a dataset of transcripts of Boston cookie theft picture descriptions in the Greek language. The samples were obtained from native Greek Speakers diagnosed with Alzheimer's and normal controls. They extracted a total of 10 features based on Lexical and Syntactic measures and employed Naive Bayes (NB) and SVM with SMO (Sequential Minimal Optimization) classifiers.

Though extensive work has been done in the field of dementia detection for widely used languages such as English and Mandarin (Chinese) as per WHO's report, 58% of dementia patients worldwide belong to low-income, middle-income countries (Chen et al., 2019). This highlights the importance of building NLP-based diagnostic tools for lesser-known and low-resource languages. India as one of the lower-middle-income (Review, 2023) countries, is estimated to have dementia prevalence in 7.4% of the population, for ages 60 and above, resulting in about 8.8 million Indians older than 60 years living with dementia (Lee et al., 2023). Hindi is the most spoken language in India and to the best of our knowledge, no work has been done in the Hindi language for the detection of Alzheimer's disease. Thus, we believe our annotated dataset could serve as a stepping stone toward the detection of AD in the Hindi language and would contribute to further research in this field.

3 Dataset

DementiaBank's Pitt Corpus is utilized in this study. DementiaBank results from an experiment conducted by Becker et al. (1994) that contains audio recordings and transcripts for the Boston Cookie Theft picture description task. The task required the participants to describe a scene, as shown in Figure 1.

The transcription for the recordings was done

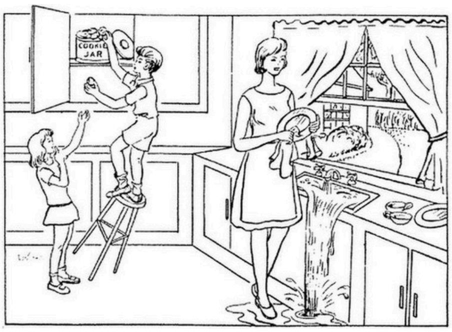


Figure 1: Boston cookie theft picture. This picture is widely used in the diagnosis of AD where patients are made to describe the scene.

manually using the CHAT (Codes for the Human Analysis of Transcripts) protocol (MacWhinney, 2017). The experiment consisted of 292 participants, with 194 having some sort of dementia. This resulted in 309 transcripts for the dementia category, as some participants had multiple recording sessions. This study deals with AD diagnosis. Thus, only the 255 transcripts from 168 AD patients and 244 transcripts from 98 CN participants were used in the study. Table 1 gives the demographics of the participants of the experiment.

Attributes	AD	CN
No. of subjects	168	98
Sex	113F / 55M	67F / 31M
Avg. Age	71.2	64.7
Avg. MMSE	19.9	29.1

Table 1: Demographics of the participants.

The transcriptions of recordings are accessible in English, which were translated into Hindi by three fluent Hindi speakers. We decided on manual translation since it is more likely to capture the social subtleties needed for the language.

We also created four more datasets using neural machine translation. BLEU (Bilingual Evaluation Understudy) score (Papineni et al., 2002) was used to compare machine translation against manual translation. mBART-50 (Tang et al., 2020), Google Translate, M2M-100 (Fan et al., 2020), and OPUS-MT (Tiedemann and Thottingal, 2020) were used as translation models. The BLEU score for each translation model is shown in Table 2. The table reflects that the text translated through the neural models is inaccurate. This shows that although deep learning models have been very prominent for various tasks, they still lack human-level performance that requires capturing niche social subtleties. Thus, manual translation was used in the study.

Translation models	BLEU-score
mBART-50	0.342
Google Translate	0.503
M2M-100	0.350
OPUS-MT	0.267

Table 2: BLEU score of different translation models

3.1 Exploratory Data Analysis

The top 10 important words from the entire dataset, along with their corresponding translation and TF-IDF (Term Frequency - Inverse Document Frequency) scores, are displayed in Table 3. TF-IDF scores are used to determine which words are more important to the document. A high TF-IDF score denotes the high significance of the word in the document.

Word	Translation	TF-IDF
कुकी	Cookie	0.3109
पानी	Water	0.2969
बर्तन	Utensils	0.2630
सिंक	Sink	0.2416
स्टूल	Stool	0.2414
माँ	Mother	0.2362
जार	Jar	0.2294
लड़की	Girl	0.2124
लड़का	Boy	0.2109
लगता	Seems	0.1942

Table 3: Top-10 most frequent words in the dataset with corresponding TF-IDF scores

3.2 Data Preprocessing

A crucial phase of data preprocessing in English is to convert all the words to uppercase or lowercase. Unlike English, case insensitivity is a feature of the Hindi language. As a result, no such modification is necessary. The punctuation marks, such as commas, semicolons, etc. that do not contribute any substantive significance to the content are eliminated in this study. Eliminating stop words from classification jobs while using NLP is another common practice that often enhances the model's performance. However, stop words like “and", “therefore," and others were frequently repeated by AD patients so, for this reason, stop words were not removed as they maintain the language traits of AD people (Khodabakhsh et al., 2014; Adhikari et al., 2021). Furthermore, Khodabakhsh et al. (2014) also suggested that pause words such as ‘um,’ ‘uh,’ and ‘ah’ were more frequently used by AD patients; as a result, they were not removed in the preprocessing phase and were translated as is.

Model	Manual Translation			OPUS-MT			M2M-100			Google Translate			mBART-50		
	Acc↑	**MMAE**↓	**F1**$_{macro}$↑	**Acc**↑	**MMAE**↓	**F1**$_{macro}$↑	**Acc**↑	**MMAE**↓	**F1**$_{macro}$↑	**Acc**↑	**MMAE**↓	**F1**$_{macro}$↑	**Acc**↑	**MMAE**↓	**F1**$_{macro}$↑
RF	0.727	0.287	0.702	0.672	0.327	0.662	0.727	0.267	0.717	0.696	0.304	0.690	0.654	0.343	0.639
NB	0.732	0.237	0.718	0.593	0.386	0.527	0.684	0.277	0.651	0.660	0.310	0.624	0.551	0.460	0.456
LR	0.726	0.269	0.721	0.666	0.332	0.653	0.690	0.307	0.680	0.709	0.292	0.704	0.654	0.343	0.639
SVC	0.732	0.267	0.731	0.690	0.309	0.683	0.690	0.309	0.683	0.703	0.298	0.697	0.648	0.353	0.637
XGB	0.709	0.291	0.702	0.666	0.335	0.659	0.690	0.307	0.680	0.690	0.310	0.684	0.690	0.309	0.683
ADA	0.715	0.285	0.708	0.678	0.321	0.669	0.684	0.317	0.682	0.721	0.281	0.718	0.696	0.302	0.688
LSTM	0.836	0.146	0.836	0.727	0.256	0.726	0.727	0.274	0.723	0.781	0.212	0.781	0.727	0.270	0.725
Bi-LSTM	0.872	0.127	0.869	0.800	0.200	0.797	0.745	0.272	0.730	0.763	0.242	0.758	0.745	0.263	0.738
BERT (Hindi)	0.842	0.125	0.840	0.820	0.199	0.807	0.717	0.276	0.717	0.700	0.300	0.694	0.740	0.260	0.734
ALBERT	0.829	0.169	0.828	0.714	0.283	0.711	0.794	0.208	0.791	0.800	0.200	0.797	0.743	0.267	0.735
XLM-RoBERTa	**0.880**	**0.120**	**0.879**	0.800	0.200	0.798	0.820	0.180	0.819	0.760	0.240	0.753	0.820	0.180	0.816
RoBERTa	0.860	0.140	0.859	0.740	0.260	0.739	0.780	0.220	0.779	0.720	0.280	0.719	0.780	0.220	0.779

Table 4: Baseline results with different algorithms for multiple translation models. The DementiaBank was translated manually and also using various machine translation algorithms.

4 Experimental Results and Discussion

We developed benchmarks using a variety of approaches, including traditional machine learning methods, deep learning, and transformer-based models. To evaluate the results for each baseline, we used accuracy, macro-mean-squared-error (MMAE), and F1-score (macro) as assessment metrics. Accuracy is a trivial evaluation metric in classification tasks. However, we use macro MAE and macro F1 to account for imbalanced datasets. Using macro MAE and macro F1 score gives equal weight to each class, regardless of size.

4.1 Benchmark Algorithms

We performed benchmarks with various machine learning and deep learning algorithms.

Machine Learning Algorithms: We employed Random Forest (RF) (Svetnik et al., 2003), Naive Bayes (NB) (Rish et al., 2001), Logistic Regression (LR), Support Vector Classification (SVC) (Hsu et al., 2003), XGBoost (XGB) (Chen et al., 2015), and AdaBoost (ADA) (Schapire, 2013) as our classical machine learning techniques. The vectorization of the corpus was done using the TF-IDF vectorizer.

Deep Learning Algorithms: LSTM (Hochreiter and Schmidhuber, 1997) and bidirectional LSTM were used as deep learning algorithms. Word embedding was done using the TensorFlow tokenizer. For transformer-based models, we used FillMask models for RoBERTa and ALBERT. We also implemented BERT (Doiron, 2023), ALBERT (Joshi, 2022), XML-RoBERTa (Pandya et al., 2021), and RoBERTa (Huang et al., 2021) for the benchmark evaluations.

4.2 Results and Analysis

The comprehensive classification results for diagnosing AD and CN are shown in Table 4. With an F1-score and accuracy of 0.879 and 0.880, respectively, the XML-Roberta model performed the best among all algorithms. SVC and NB beat the other ML methods with an accuracy of 0.732. DL and ML models did not perform as good as transformer-based models. The requirement for more sophisticated and reliable algorithms for text identification is highlighted by the model's substantially lower F1 score of ML models compared to transformer-based models. Similarly, we run the benchmark evaluations for the translations using neural machine translation. The translation made using OPUS-MT showed a f1-score of 0.807. This shows that the machine translation-based transcripts can also capture the nuances in speech which are necessary to delineate AD patients from CN groups. The benchmark evaluations done on manual translation have shown remarkable performance compared to the translations done by machine translation algorithms. This shows that automated translations may not account for social subtleties.

5 Conclusion

Our work presents a novel dataset in the Hindi language that classifies the speech of AD patients against CN individuals. AD cannot be cured, so its detection and management become crucial. Speech impairment is one of the most common symptoms of AD. Hence, we have created this dataset that has the potential to significantly aid in the development of automated, speedy, and cost-effective systems for detecting AD. We also performed benchmarks on the created dataset and achieved the highest accuracy of 0.880 and an F1 score of 0.879 with the XLM-Roberta model. Such considerable benchmark results encourage further research in the field by extending the dataset and creating more sophisticated and domain-specific models. Results show there is room for improvement in constructing superior models.

Limitations

Some potential limitations of this work include a relatively small sample size, which may limit the generalizability of the results. Hindi is spoken differently across India, hence the translations made by the three translators may not be representative. Study did not examine the potential impact of regional dialects or variations in Hindi, on the accuracy of the diagnosis. Finally, the study focused solely on the use of speech narratives and did not explore the other types of data, e.g., imaging or genetic data, which could be important for the diagnosis of AD.

References

Surabhi Adhikari, Surendrabikram Thapa, Usman Naseem, Priyanka Singh, Huan Huo, Gnana Bharathy, and Mukesh Prasad. 2022. Exploiting linguistic information from Nepali transcripts for early detection of Alzheimer's Disease using natural language processing and machine learning techniques. *International Journal of Human-Computer Studies*, 160:102761.

Surabhi Adhikari, Surendrabikram Thapa, Priyanka Singh, Huan Huo, Gnana Bharathy, and Mukesh Prasad. 2021. A comparative study of machine learning and NLP techniques for uses of stop words by patients in diagnosis of Alzheimer's Disease. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

James T Becker, François Boiler, Oscar L Lopez, Judith Saxton, and Karen L McGonigle. 1994. The natural history of Alzheimer's Disease: description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594.

Jia-Hung Chen, Chien-Tai Hong, Dean Wu, Wen-Chou Chi, Chia-Feng Yen, Hua-Fang Liao, Lung Chan, and Tsan-Hon Liou. 2019. Dementia-related functional disability in moderate to advanced parkinsons disease: assessment using the World Health Organization disability assessment schedule 2.0. *International Journal of Environmental Research and Public Health*, 16(12):2230.

Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, et al. 2015. XGBoost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4.

N. Doiron. 2023. Hindi-Bert. `https://huggingface.co/monsoon-nlp/hindi-bert`. [Online; accessed 12-Apr-2023].

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond English-Centric Multilingual Machine Translation.

Zhiqiang Guo, Zhaoci Liu, Zhenhua Ling, Shijin Wang, Lingjing Jin, and Yunxia Li. 2020. Text classification by contrastive learning and cross-lingual data augmentation for Alzheimers Disease detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6161–6171.

Harald Hampel, David Prvulovic, Stefan Teipel, Frank Jessen, Christian Luckhaus, Lutz Frölich, Matthias W. Riepe, Richard Dodel, Thomas Leyhe, Lars Bertram, Wolfgang Hoffmann, and Frank Faltraco. 2011. The future of Alzheimer's Disease: The next 10 years. *Progress in Neurobiology*, 95(4):718–728. Biological Markers for Neurodegenerative Diseases.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al. 2003. A practical guide to support vector classification.

Xixuan Huang, Nankai Lin, Kexin Li, Lianxi Wang, and Suifu Gan. 2021. Hinplms: Pre-trained language models for hindi. In *2021 International Conference on Asian Language Processing (IALP)*, pages 241–246. IEEE.

Raviraj Joshi. 2022. L3Cube-HindBERT and DevBERT: Pre-Trained BERT Transformer models for Devanagari based Hindi and Marathi Languages. *arXiv preprint arXiv:2211.11418*.

Sweta Karlekar, Tong Niu, and Mohit Bansal. 2018. Detecting linguistic characteristics of Alzheimer's Dementia by interpreting neural models. *arXiv preprint arXiv:1804.06440*.

Ali Khodabakhsh, Serhan Kuxuolu, and Cenk Demirolu. 2014. Natural language features for detection of Alzheimer's Disease in conversational speech. In *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 581–584.

Jinkook Lee, Erik Meijer, Kenneth M Langa, Mary Ganguli, Mathew Varghese, Joyita Banerjee, Pranali Khobragade, Marco Angrisani, Ravi Kurup, Sankha Shubhra Chakrabarti, et al. 2023. Prevalence of dementia in India: National and state estimates from a nationwide study. *Alzheimer's & Dementia*.

Zhaoci Liu, Zhiqiang Guo, Zhenhua Ling, Shijin Wang, Lingjing Jin, and Yunxia Li. 2019. Dementia detection by analyzing spontaneous Mandarin speech. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 289–296. IEEE.

Brian MacWhinney. 2017. Tools for analyzing talk part 1: The chat transcription format. *Carnegie.[Google Scholar]*, 16.

Amit Meghanani, CS Anoop, and Angarai Ganesan Ramakrishnan. 2021. Recognition of Alzheimers Dementia from the transcriptions of spontaneous speech using fastText and CNN models. *Frontiers in Computer Science*, 3:624558.

Usman Naseem, Imran Razzak, Shah Khalid Khan, and Mukesh Prasad. 2021. A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models. *Transactions on Asian and Low-Resource Language Information Processing*, 20(5):1–35.

Hariom A. Pandya, Bhavik Ardeshna, and Dr. Brijesh S. Bhatt. 2021. Cascading adaptors to leverage english data to improve performance of question answering for low-resource languages.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Vassiliki Rentoumi, George Paliouras, Eva Danasi, Dimitra Arfani, Katerina Fragkopoulou, Spyridoula Varlokosta, and Spyros Papadatos. 2017. Automatic detection of linguistic indicators as a means of early detection of Alzheimer's Disease and of related dementias: A computational linguistics analysis. In *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 000033–000038. IEEE.

World Population Review. 2023. Middle-Income Countries 2023. https://worldpopulationreview.com/country-rankings/middle-income-countries. [Online; accessed 12-Apr-2023].

Irina Rish et al. 2001. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46.

Robert E Schapire. 2013. Explaining Adaboost. *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, pages 37–52.

Jose A. Soria Lopez, Hector M. González, and Gabriel C. Léger. 2019. Chapter 13 - Alzheimer's Disease. In Steven T. Dekosky and Sanjay Asthana, editors, *Geriatric Neurology*, volume 167 of *Handbook of Clinical Neurology*, pages 231–255. Elsevier.

Vladimir Svetnik, Andy Liaw, Christopher Tong, J Christopher Culberson, Robert P Sheridan, and Bradley P Feuston. 2003. Random Forest: a classification and regression tool for compound classification and qsar modeling. *Journal of Chemical Information and Computer Sciences*, 43(6):1947–1958.

Greta Szatloczki, Ildiko Hoffmann, Veronika Vincze, Janos Kalman, and Magdolna Pakaski. 2015. Speaking in Alzheimers Disease, is that an early sign? importance of changes in language abilities in Alzheimers Disease. *Frontiers in aging neuroscience*, 7:195.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Surendrabikram Thapa, Surabhi Adhikari, Usman Naseem, Priyanka Singh, Gnana Bharathy, and Mukesh Prasad. 2020a. Detecting Alzheimers Disease by exploiting linguistic information from nepali transcript. In *Neural Information Processing: 27th International Conference, ICONIP 2020, Bangkok, Thailand, November 18–22, 2020, Proceedings, Part IV 27*, pages 176–184. Springer.

Surendrabikram Thapa, Awishkar Ghimire, Surabhi Adhikari, Akash Kumar Bhoi, and Paolo Barsocchi. 2022. Cognitive internet of things (IoT) and computational intelligence for mental well-being. In *Cognitive and Soft Computing Techniques for the Analysis of Healthcare Data*, pages 59–77. Elsevier.

Surendrabikram Thapa, Priyanka Singh, Deepak Kumar Jain, Neha Bharill, Akshansh Gupta, and Mukesh Prasad. 2020b. Data-driven approach based on feature selection technique for early diagnosis of Alzheimers Disease. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Ning Wang, Yupeng Cao, Shuai Hao, Zongru Shao, and KP Subbalakshmi. 2021. Modular multi-modal attention network for Alzheimer's Disease detection using patient audio and language data. In *Interspeech*, pages 3835–3839.

A Appendix

The example of translated transcripts for control normal (CN) participants and AD patients are shown in Table 5 and Table 6 respectively. The translations show that google translate was able to generate more similar translations as manual translation than other translation algorithms. As the evaluation of translation can be subjective, we use BLEU scores as mentioned in Table 2.

	CHAT ID: 002-2.cha
Transcript from DementiaBank	a boy and a girl are in the kitchen with their mother. and the little boy is getting a cookie for the little girl but he's on a stool it's about to fall down. the mother is washing dishes. and she's obviously thinking of something else because the water's pouring out over the sink. she's finished with some of the dishes. it seems to be summertime because there are bushes. the window is open. there seems to be some kind of a breeze because the curtains on the sill there blowing. must be fairly hot. the mother's in a sleeveless dress. the children are in short sleeve tops and have sandals. the little boy has tennis shoes on. the mother obviously is unaware of what the children are doing. she's going to be aware of it very shortly. I mean how much more do you want hm okay.
Manual Translation	एक लड़का और एक लड़की रसोई में अपनी माँ के साथ हैं। और छोटा लड़का छोटी लड़की के लिए एक कुकी ले रहा है लेकिन वह एक स्टूल पर है जो नीचे गिरने वाला है। माँ बर्तन धो रही है। और वह स्पष्ट रूप से कुछ और सोच रही है क्योंकि सिंक के ऊपर से पानी बह रहा है। उसने कुछ बर्तन समाप्त कर लिए हैं। ऐसा लगता है कि गर्मियों का समय है क्योंकि झाड़ियाँ हैं। खिड़की खुली है। ऐसा लगता है कि किसी तरह की हवा चल रही है क्योंकि वहाँ की सिल पर लगे पर्दे उड़ रहे हैं। काफी गर्म होना चाहिए। माँ एक बिना आस्तीन की पोशाक में है।
Google TM	एक लड़का और एक लड़की अपनी माँ के साथ रसोई में हैं। और छोटे लड़के को छोटी लड़की के लिए एक कुकी मिल रही है, लेकिन वह एक स्टूल पर है जो नीचे गिरने वाला है। माँ व्यंजन धो रही है। और वह स्पष्ट रूप से कुछ और के बारे में सोच रही है क्योंकि पानी सिंक पर बाहर निकल रहा है। वह कुछ व्यंजनों के साथ समाप्त हो गई है। यह गर्मियों में लगता है क्योंकि वहाँ झाड़ियों हैं। खिड़की खुली है । वहाँ कुछ प्रकार की हवा लगती है क्योंकि वहाँ पर पर्दे बहते हैं। काफी गर्म होना चाहिए। एक स्लीवलेस ड्रेस में माँ की। बच्चे छोटी आस्तीन में होते हैं और सैंडल होते हैं। छोटे लड़के के पास टेनिस के जूते हैं। मां स्पष्ट रूप से इस बात से अनजान है कि बच्चे क्या कर रहे हैं। वह बहुत जल्द ही इसके बारे में जागरूक हो रही है। Imean कितना अधिक आप hm ठीक चाहते हैं।
mBART-50 TM	एक लड़का और एक लड़की मां के साथ रसोईघर में हैं। और छोटा लड़का छोटे लड़की के लिए एक कुकी ले रहा है लेकिन वह एक कुर्सी पर है यह नीचे गिरने के लिए जा रहा है। मां बर्तन धो रही है। और वह स्पष्ट रूप से कुछ और के बारे में सोच रहा है क्योंकि पानी Sink के ऊपर बह रहा है। वह कुछ बर्तनों के साथ समाप्त हो गया है। लगता है यह गर्मी का समय है क्योंकि वहाँ झाड़ियाँ हैं। विंडो खुला है। ऐसा लगता है कि हवा कुछ तरह से चल रही है क्योंकि सील पर परदे उड़ रहे हैं। काफी गर्म होना चाहिए। मां बेहिश्त पोशाक में है। बच्चों को कमज़ोर सूती के टोप पहने जाते हैं और सैंडल पहने जाते हैं। छोटे लड़के पर टेनिस जूते हैं। मां स्पष्ट रूप से नहीं जानती कि बच्चे क्या कर रहे हैं। वह बहुत जल्दी यह जानने के लिए जा रहा है। imean कितना अधिक आप चाहते हैं hm ठीक है।
M2M-100 TM	एक लड़का और एक लड़की अपनी मां के साथ रसोई में हैं। और छोटा लड़का छोटी लड़की के लिए एक कुकी मिल रहा है लेकिन वह एक मल पर है यह गिरने के लिए तैयार है। मां ने खाना धोया है। और वह स्पष्ट रूप से कुछ और के बारे में सोच रहा है क्योंकि पानी स्नान पर बह रहा है। वह कुछ डिश के साथ खत्म हो गया है। ऐसा लगता है कि यह गर्मियों का समय है क्योंकि वहाँ झाड़ियाँ हैं। खिड़की खुली है। ऐसा लगता है कि वहाँ कुछ तरह का एक बर्फ है क्योंकि वहां ब्लेड पर पर्दे फेंक रहे हैं। काफी गर्म होना चाहिए। मम्मी के कपड़े बेकार हैं। बच्चों के पास छोटी–छोटी बूंदें हैं और सैंडल हैं। बच्चे के पास टेनिस जूते हैं। माता–पिता स्पष्ट रूप से यह नहीं जानते हैं कि बच्चे क्या कर रहे हैं। वह जल्द ही इसके बारे में जान लेगी। इमेन कितना अधिक आप चाहते हैं ठीक है।
OPUS-MT TM	एक लड़का और लड़की रसोई में अपनी माँ के साथ हैं. और छोटा लड़का छोटी लड़की के लिए एक कुकी हो रहा है...... लेकिन वह यह नीचे गिर करने के बारे में है. माँ बरतन धो रही है. और वह स्पष्ट रूप से कुछ और के बारे में सोच रहा है क्योंकि पानी डूबने पर बाहर बहा रहा है. वह कुछ व्यंजनों के साथ समाप्त हो गया है. ऐसा लगता है कि गर्मियों का समय है क्योंकि वहाँ झाड़ी हैं. विंडो खुला है. ऐसा लगता है कि एक प्रकार की हवा है क्योंकि वहाँ के परदे वहाँ बाढ़ की बाढ़ से गुज़र रहे हैं । काफी गर्म होना चाहिए. माँ एक बेकार पोशाक में है. बच्चे छोटी सीटों पर हैं और जूते हैं. छोटे लड़के के जूते. स्पष्ट रूप से माँ को पता नहीं कि बच्चे क्या कर रहे हैं । वह इसके बारे में बहुत जल्द पता करने जा रहा है. Dmm ठीक है चाहते हैं कितना अधिक.

Table 5: Example of the original and translated transcripts of control normal (CN) participants

	CHAT ID: 051-1.cha
Transcript from DementiaBank	the boy reaching for the cookies is gonna fall down on the um what uh shes saying shho i think shho but give me a cookie too um mother is wiping the dish sink is running over the uh im to tell everything i see all the action yeah splashing the water gram well xxx and the girl saying shho give me a cookie too thats all i see
Manual Translation	लड़का कुकीज़ के लिए पहुँच रहा है उम पर गिरने वाला है वह क्या कह रही है उह क्या कह रही है शो मुझे लगता है शो लेकिन मुझे एक कुकी भी दे उम चने के पानी के छींटे अच्छी तरह से और लड़की कह रही है कि मुझे एक कुकी भी दो वह सब मैं देख रहा हूँ।
Google TM	कुकीज़ के लिए पहुंचने वाला लड़का उम पर गिरने वाला है जो उह शेस कह रहा है कि मुझे लगता है कि मुझे लगता है कि मुझे एक कुकी भी है, उम माँ डिश सिंक को पोंछ रही है, उह इम के ऊपर चल रही है सब कुछ बताने के लिए मैं सभी एक्शन को देख रहा हूं पानी चने की अच्छी तरह से xxx और शॉ कहती है कि मुझे एक कुकी भी है जो मैं देख रहा हूं।
mBART-50 TM	लड़का कुकी के लिए पहुंचने के लिए नीचे गिरने के लिए है um क्या uh वह कह रही है shho मैं सोचता हूँ shho लेकिन मुझे एक कुकी भी दे रहा हूँ um माँ बर्तन धो रहा है धोने के लिए कूल्हे पर चल रहा है uh im सब कुछ बताने के लिए मैं देख रहा हूँ सभी क्रिया हाँ छिड़का हुआ पानी ग्राम अच्छी तरह से xxx और लड़की कह रही है shho मुझे एक कुकी भी दे रहा है कि सब मैं देख रहा हूँ।
M2M-100 TM	कुकीज़ के लिए पहुंचने वाला लड़का ओम पर गिरने जा रहा है कि ओह क्या कहता है shho मैं सोचता हूं shho लेकिन मुझे एक कुकी दें ओम माँ डिश स्नान पर चल रही है ओह im सब कुछ बताने के लिए मैं देखता हूं सभी कार्रवाई हाँ पानी gram अच्छी xxx और लड़की कहती है shho मुझे एक कुकी देना भी यही सब मैं देखता हूं।
OPUS-MT TM	कुकीज़ के लिए प्रवेश करने वाला लड़का उमर पर गिर रहा है क्या उह वह कह रहा है hhhh मैं लगता है पर नीचे गिर रहा है लेकिन मुझे एक कुकी भी एक उमर पानी सिंक बंद कर रहा है सब कुछ बताने के लिए उह, मैं सब कुछ अच्छी तरह से देखने के लिए जा रहा हूँ और लड़की कहते हैं कि मुझे एक कुकी है कि मैं भी देख रहा हूँ कि मैं देख रहा हूँ

Table 6: Example of the original and translated transcripts of patients with AD

Investigating Massive Multilingual Pre-Trained Machine Translation Models for Clinical Domain via Transfer Learning

Lifeng Han[1], Gleb Erofeev[2], Irina Sorokina[2], Serge Gladkoff[2], and Goran Nenadic[1]
[1] The University of Manchester, UK
[2] Logrus Global, Translation & Localization
lifeng.han, g.nenadic@manchester.ac.uk
gleberof, irina.sorokina, serge.gladkoff@logrusglobal.com

Abstract

Massively multilingual pre-trained language models (MMPLMs) are developed in recent years demonstrating superpowers and the pre-knowledge they acquire for downstream tasks. This work investigates whether MMPLMs can be applied to clinical domain machine translation (MT) towards entirely unseen languages via transfer learning. We carry out an experimental investigation using Meta-AI's MMPLMs "wmt21-dense-24-wide-en-X and X-en (WMT21fb)" which were pre-trained on 7 language pairs and 14 translation directions including English to Czech, German, Hausa, Icelandic, Japanese, Russian, and Chinese, and the opposite direction. We fine-tune these MMPLMs towards English-Spanish language pair which did not exist at all in their original pre-trained corpora both implicitly and explicitly. We prepare carefully aligned clinical domain data for this fine-tuning, which is different from their original mixed domain knowledge. Our experimental result shows that the fine-tuning is very successful using just 250k well-aligned in-domain EN-ES segments for three sub-task translation testings: clinical cases, clinical terms, and ontology concepts. It achieves very close evaluation scores to another MMPLM NLLB from Meta-AI, which included Spanish as a high-resource setting in the pre-training. To the best of our knowledge, this is the first work on using MMPLMs towards clinical domain transfer-learning NMT successfully for totally unseen languages during pre-training.

1 Introduction

Multilingual neural machine translation (MNMT) has its root from the beginning of NMT era (Dong et al., 2015; Firat et al., 2016) but only made its first milestone when Google's end-to-end MNMT arrived (Johnson et al., 2017) where the artificial token was introduced for the first time for translation task at the beginning of the input source sentence to indicate the specified target language, e.g. "2en" as translating into English. This model used a shared word-piece vocabulary and enabled multilingual NMT through a single encoder-decoder model training. Google's MNMT also demonstrated the possibility of "zero-shot" translation as long as the languages to be translated from or to have been seen during the training stage, even though not explicitly. However, as the authors mentioned, Google's MNMT only allows translating between languages that have been seen individually as "source and target languages during some point, not for entirely new ones" in their many-to-many model, which was tested using the WMT14 and WMT15 data (Johnson et al., 2017). This set an obstacle to translating freshly new languages that do not exist in their pre-training stage. Then using the later developed NMT structure Transformer and BERT (Devlin et al., 2019; Vaswani et al., 2017), Facebook AI extended the coverage of multilingual translation into 50, 100, and 200+ languages via mBERT-50 (Tang et al., 2020), M2M-100 (Fan et al., 2021), and NLLB (NLLB Team et al., 2022) models. However, these models never address the issue of translating entirely new languages that do not exist in their pre-training stage, which sets an obstacle for MT applications in serving an even broader community.

In this work, we move one step forward towards domain-specific transfer-learning (Zoph et al., 2016) for NMT via fine-tuning an entirely new language pair that does not exist in the deployed multilingual pre-trained language models (MPLMs). The MPLMs we used are from Facebook AI (Meta-AI)' s submission to the WMT21 news translation

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 31–40
July 14, 2023 ©2023 Association for Computational Linguistics

task, i.e. “wmt21-dense-24-wide-en-X” and “wmt21-dense-24-wide-X-en” which were pre-trained for 7 languages Hausa (ha), Icelandic (is), Japanese (ja), Czech (cs), Russian (ru), Chinese (zh), German (de) to English (en), and backward (Tran et al., 2021). We use a well-prepared 250k pairs of English-Spanish (en-es) clinical domain corpus and demonstrate that not only it is possible to achieve successful transfer-learning on this explicit new language pair, i.e. the Spanish language is totally unseen among the languages in the MPLM, but also the domain knowledge transfer from general and mixed domain to the clinical domain is very successful. In comparison to the massively MPLM (MMPLM) NLLB which covers Spanish as a high-resource language at its pre-training stage, our transfer-learning model achieves very close evaluation scores in most sub-tasks (clinical cases and clinical terms translation) and even wins NLLB in ontology concept translation task by the metric COMET (Rei et al., 2020) using ClinSpEn2022 testing data at WMT22. This is a follow-up work reporting further findings based on our previous shared task participation (Han et al., 2022a) and pre-print (Han et al., 2022b).

2 Related Work

Regarding the early usage of special tokens in NMT, Sennrich et al. (2016) designed the token T from Latin Tu and V from Latin Vos for familiar and polite indicators attached to the source sentences towards English-to-German NMT. Yamagishi et al. (2016) designed tokens <all-active>, <all-passive>, <reference> and <predict> to control of voice of Japanese-to-English NMT; either they are active, passive, reference aware or prediction guided. Subsequently, Google's MNMT system designed target language indicators, e.g. <2en> and <2jp> controlling the translation towards English and Japanese respectively (Johnson et al., 2017). Google's MNMT also designed mixed target language translation control, e.g. (1-α)<2ko> + α<2jp> tells a mixed language translation into Korean and Japanese with a weighting mechanism. We take one step further to use an existing language controller token from a MPLM as a pseudo code to fine-tune an external language translation model, which was entirely not seen during the pre-training stage.

Regarding transfer-learning applications for downstream NLP tasks other than MT, Muller et al. (2021) applied transfer learning from MPLMs towards unseen languages of different typologies on dependency parsing (DEP), named entity recognition (NER), and part-of-speech (POS) tagging. Ahuja et al. (2022) carried out zero-shot transfer learning for natural language inference (NLI) tasks such as question answering.

In this paper, we ask this research question (RQ): Can Massive Multilingual Pre-Trained Language Models Create a Knowledge Space Transferring to Entirely New Language (Pairs) and New (clinical) Domains for Machine Translation Task via Fine-Tuning?

3 Model Settings

To investigate into our RQ, we take Meta-AI's MNMT submission to WMT21 shared task on news translation, i.e. the MMPLM “wmt21-dense-24-wide-en-X” and “wmt21-dense-24-wide-X-en” as our test-base, and we name them as WMT21fb models (Tran et al., 2021)[1]. They are conditional generation models from the same structure of massive M2M-100 (Fan et al., 2021) having a total number of 4.7 billion parameters which demand high computational cost for fine-tuning. WMT21fb models were trained on mixed domain data using “all available resources” they had, for instances, from historical WMT challenges, large-scale data mining, and their in-domain back-translation. Then these models were fine-tuned in news domain for 7 languages including Hausa, Icelandic, Japanese, Czech, Russian, Chinese, German from and to English.

The challenging language we choose is Spanish, which did not appear in the training stage of WMT21fb models. The fine-tuning corpus we use is extracted from MeSpEn (Villegas et al., 2018) clinical domain data, of which we managed to extract 250k pairs of English-Spanish segments after data cleaning. They are from IBECS-descriptions, IBECS-titles, MedlinePlus-health_topics-titles, MedlinePlus-health_topics-descriptions,

[1] https://github.com/facebookresearch/fairseq/tree/main/examples/wmt21

Рис. 1: (Figure:) Difference of Google's Multi-lingual NMT Bridge Models (left) and Our Transfer-Learning Model (right).

Pubmed-descriptions, Scielo-descriptions, and Scielo-titles.

To implement the fine-tuning, we use the <2en> token for translating from Spanish to English, and <2ru> (originally to Russian) pseudo token for translating towards English-to-Spanish (en2es) [2]. The difference between our transfer-learning NMT model and Google's MNMT can be shown in Figure 1, right vs left. In Google's MNMT model, it can only translate "new language pairs" that are not explicitly seen but implicitly seen, e.g. bridging language pairs (Ukrainian-to-English and English-to-Russian ⇒ Ukrainian-to-Russian), or language pairs that have been seen as source (Korean) and target (Portuguese) somewhere. In our transfer-learned NMT, Spanish was not among the trained languages at all.

In comparison, we deploy another MMPLM from Meta-AI, i.e. the "No-Language-Left-Behind (NLLB)" which was trained on 204 languages including Spanish as one of their high-resource ones (NLLB Team et al., 2022). NLLB full model is a massive size Transformer having 55 billion parameters and we use its distilled version NLLB-200-distilled [3], which still has 1.3 billion parameters. Fine-tuning is carried out on NLLB using the same 250K ES-EN corpus.

3.1 Model Parameters in Detail

Some fine-tuning parameters for NLLB-200-distilled (NLLB Team et al., 2022) are listed below:

- batch size = 24
- gradient accumulation steps = 8
- weight decay = 0.01
- learning rate = 2e-5
- number of training epochs = 1
- encoder-decoder layers = 24+24
- Activation function (encoder/decoder) = ReLU

The Parameters for fine-tuning WMT21fb model are the same as for the NLLB-200, except for the batch size which is set as 2, which is because the model is too large and we got an OOM error if the batch size is set above 2. More details on M2M-100 for Conditional Generation structure (Fan et al., 2021) we used can be find in Figure 2.

4 Model Evaluations

4.1 Testing Corpus from Clinical Domain

We used the official testing corpus from ClinSpEn2022 shared task affiliated to Biomedical-MT at WMT22. ClinSpEn2022 aims at developing clinical domain machine translation on Spanish-English language pair[4], which is hosted in CodaLab (Pavao et al., 2022) [5].

[2] using <2es> token will result into errors since Spanish was actually not used in the WMT21fb PLMs

[3] https://huggingface.co/facebook/nllb-200-distilled-1.3B

[4] https://temu.bsc.es/clinspen/

[5] https://codalab.lisn.upsaclay.fr/competitions/6696

```
M2M100ForConditionalGeneration(
  (model): M2M100Model(
    (shared): Embedding(128009, 2048, padding_idx=1)
    (encoder): M2M100Encoder(
      (embed_tokens): Embedding(128009, 2048, padding_idx=1)
      (embed_positions): M2M100SinusoidalPositionalEmbedding()
      (layers): ModuleList(
        (0): M2M100EncoderLayer(
          (self_attn): M2M100Attention(
            (k_proj): Linear(in_features=2048, out_features=2048, bias=True)
            (v_proj): Linear(in_features=2048, out_features=2048, bias=True)
            (q_proj): Linear(in_features=2048, out_features=2048, bias=True)
            (out_proj): Linear(in_features=2048, out_features=2048, bias=True)
          )
          (self_attn_layer_norm): LayerNorm((2048,), eps=1e-05, elementwise_affine=True)
          (activation_fn): ReLU()
          (fc1): Linear(in_features=2048, out_features=16384, bias=True)
          (fc2): Linear(in_features=16384, out_features=2048, bias=True)
          (final_layer_norm): LayerNorm((2048,), eps=1e-05, elementwise_affine=True)
        )
        (1): M2M100EncoderLayer(
```

Рис. 2: (Figure:) M2M-100 Model Structure For Conditional Generation Encoder: Samples and Parameters

Task-I: Clinical Cases (CC) EN→ES						
MT fine-tuning	in.es?	SacreBLEU	METEOR	COMET	BLEU-HF	ROUGE-L-F1
Clnical-NLLB	Yes	37.74	0.6273	0.4081	0.3601	0.6193
Clinical-WMT21fb	No	34.30	0.5868	0.3448	0.3266	0.5927
Task-II: Clinical Terms (CT) EN←ES						
MT fine-tuning	in.es?	SacreBLEU	METEOR	COMET	BLEU-HF	ROUGE-L-F1
Clinical-NLLB	Yes	28.57	0.5873	1.0290	0.2844	0.6710
Clinical-WMT21fb	No	24.39	0.5840	0.8584	0.2431	0.6699
Task-III: Ontology Concept (OC) EN→ES						
MT fine-tuning	in.es?	SacreBLEU	METEOR	COMET	BLEU-HF	ROUGE-L-F1
Clinical-NLLB	Yes	41.63	0.6072	0.9180	0.3932	0.7477
Clinical-WMT21fb	No	40.71	0.5686	0.9908	0.3859	0.7199

Таблица 1: (Table:) Evaluation Scores using Five Official Metrics from ClinSpEn2022 Benchmark on Two Models. The column "in.es" means if the original pre-trained model included the Spanish language before fine-tuning/transfer-learning.

There are three sub-tasks: 1) Clinical Cases (CC): on 202 COVID-19 clinical case reports; 2) Clinical Terms (CT): using more than 19K parallel terms extracted from biomedical literature and electric health records (EHRs); 3) Ontology Concepts (OC): using more than 2K parallel concepts from biomedical ontology. The translation direction on these three sub-tasks are EN→ES, EN←ES, and EN→ES respectively.

4.2 Evaluation Metrics

The official evaluation metrics used by ClinSpEn2022 shared task are METEOR (Banerjee and Lavie, 2005), SacreBLEU (Post, 2018), COMET (Rei et al., 2020), BLEU-HF (HuggingFace) (Papineni et al., 2002), and ROUGE-L-F1 (Lin, 2004). Among these, METEOR is a metric using both precision and recall not only on word surface level but also introducing paraphrasing features. COMET was proposed recently by taking advantage of cross-lingual PLMs using knowledge from both source and target languages. ROUGE was originally designed for text summarisation evaluation using n-gram co-occurrences, while ROUGE-L added the Longest Common Subsequence (LCS) feature from translation study.

The reporting of BLEU metric scores has certain uncertainty, which is caused by some parameter settings when using BLEU metric including number of references, length penalty computation on multi-references, maximum n-gram, and smoothing applied to 0-count n-grams. To address these issues, SacreBLEU added some constrains while using BLEU metric. These include the applying of its own metric-internal pre-processing for detokenised system outputs, the avoiding of user handling reference set via automatically downloading from WMT, and the export of a summary on settings used.

4.3 Evaluation Scores

We present the MT evaluation scores using five official metrics from ClinSpEn2022 shared task on the three sub-tasks in Table 1, for translating clinical cases, clinical terms, and clinical concepts. The two fine-tuned models are clinic-NLLB which is achieved by domain fine-tuning and clinic-WMT21fb which is a domain fine-tuning plus transfer-learning model to a new language space.

On Task 1 and 2, Clinical-WMT21fb has very comparable evaluation scores to clinical-NLLB, even though it only used 250k pairs es-en sentences for fine tuning without seeing any en-es or Spanish language at all during pre-training. In contrast, clinical-NLLB used a large amount of Spanish data for its pre-training phase. On Task 3, the evaluation scores of these two models are even closer on BLEU and SacreBLEU, especially the clinical-WMT21fb wining COMET metric over clinical-NLLB (0.9908 vs 0.9180).

This experimental result shows that with a carefully prepared certain amount of fine-tuning data, e.g. 250k pair of sentences, the MMPLMs are capable to create a semantic knowledge space transferring to an entirely new (external) language pair for NMT task in a new domain, i.e. clinical domain. This answers our RQ set up in the beginning of this investigation.

4.4 Human Evaluation

We looked into three sub-task translation outputs from the model clinical-WMT21fb. It shows that for the EN←ES translation task, i.e. the sub-task 2 clinical term translation, the output file is totally file with only English tokens. On the other two sub-tasks, i.e. the clinical cases and ontology concept translation, which have the translation direction EN→ES, there are some Russian tokens in the output, not only Spanish tokens. However, the Russian tokens in the Spanish sentences are not nonsense, instead proper translations of entities and words. The entire test set of these two sub-tasks is very large around 300K sentences/segments, and there are only 12K lines of them (4%) have Russian tokens. So we have fine-tuned the model in EN-RU direction on EN-SP data, and it translates well into Spanish! But if there isn't a suitable Spanish token in the generation model, it takes a Russian token.

We also looked into the translation outputs from clinic-NLLB model for error analysis using two native Spanish speakers, one of them having a PhD degree in biomedical NLP field and the other having a Master degree in translation studies. The error analysis shows that some of the translation errors come from very literal translation, and others come from gender related mistakes. This suggests that the massively pre-trained MLM is still not there to capture the differences of linguistic features among pre-trained languages.

5 Discussion

5.1 On Automatic Metrics

We had more thoughts on the automatic evaluation settings and outputs, especially on the COMET metric in comparison to others.

Firstly, the closeness of most automatic metric scores does not necessarily mean that the translation outputs are very good. Most metrics only measure the linguistic proximity of outputs to the "gold standard of reference".

Secondly, COMET is a reference-less metric taking advantage of cross-lingual PLMs using knowledge from both source and target languages. This has pros and cons: a) it might be able to capture the semantic relatedness without seeing the same language tokens, even in the same sequence/sentence; b) also due to this, it is not able to distinguish foreign language tokens in the translation output, which normally shall receive a penalty in evaluation scores. This also inspires another research topic, i.e. shall we really punish the foreign or mixed-language tokens in the translation output in all evaluation conditions, or it shall depend on the situation of the output applications? This has an echo to Google's zero-shot MNMT model (Johnson et al., 2017) when the mixed language tokens are used for translation model, e.g. (1-α)<2KO>+α<2JP> resulting in mixed tokens of Korean and Japanese in the output translation but they are semantically correct tokens.

In a situation when users want only the Spanish translation output, 4% of Russian tokens in the Spanish translation should surely receive a penalty in the quality evaluation setting. The COMET metric will fail this mission, and professional human evaluation is always much needed for trustworthiness. However, in a situation to measure the models' cross-lingual capability on semantic preservation for direct output, or as input into other ML models, is it better to generate NULL or meaningless tokens or random translations in the target language, or to choose semantically correct foreign tokens when the model does not know how to predict the exact correct target tokens? This inspires us to think again about the evaluation setting on different tasks.

5.2 MT System Output Examples

We present the MT system output examples from both clinical-WMT21fb and clinical-NLLB-200 for three tasks in Figure 3, 4, and 5. In these figures, the green colour is for the "preferred translations" while the orange colour is for "both sounds good". The annotations were firstly marked by one of the two human evaluators we have, and then verified by the second native Spanish speaker.

From these sampled MT outputs, the model clinical-WMT21fb sometimes outperforms clinical-NLLB-200, and vice versa. For instance, in the concept translation (Figure 5), the English concept "Abnormality of body height" (ont_1) is better translated by transfer-learned model into "Anomalía de la altura corporal" than "Anomalías de la talla corporal" by clinical-NLLB, since "altura" means "height" while "talla" actually means "size" which is not accurate. We will carry out a systematic human evaluation in a larger sample size.

Regarding rare Russian tokens from the language-transferred model, in Task-1, "Вскоре" from clinical-WMT21fb in line_n 4 means "soon", even though it is a Russian token, i.e. non-Spanish token. In Task-3, "Тип" in "Тип autosómico dominante" means "type of" from ont_11 which is a meaningful Russian token.

6 Conclusion and Future Work

We investigated if real transfer-learning NMT is possible using massive multilingual pre-trained LMs (MMPLMs) to translate external languages that are unseen at all in the training phase. We used Meta-AI's mixed domain multilingual PLMs (WMT21fb) as our test base, 250K well-prepared EN-ES clinical data as fine-tuning corpus, and <2ru> as pseudo-code for new language (out-of-en) fine-tuning. We tested the fine-tuned model on ClinSpEn2022 clinical domain shared task data, and the results show that this fine-tuning is successful, which achieves very comparable scores to Meta-AI's MMPLM NLLB model, which had Spanish in the training phase as a high-resource setting. We think this demonstrates that the Hyper-Transformer model from WMT21fb does build a language-independent "semantic space" that allows one to understand a different language and correctly construct a totally different language model when fine-tuned on the language which was absent and different from the languages it was trained upon. This finding can be very useful for future clinical knowledge transformation, e.g. from existing high-resource languages to low-resource languages, such that clinicians from low-resource language speakers can also benefit from AI-supported decision-making. The well-trained clinical models based on properly translated resources can also potentially support patients' self-diagnoses and self-care in originally scarce resource settings.

There are many future works that can be carried out based on the findings from this work. Firstly, we plan to carry out an extensive human-expert-based evaluation, e.g. using HOPE metric (Gladkoff and Han, 2022), looking into the differences between

doc_n	line_n	**Transfer-learning**: clinical-WMT21fb:en2es
doc_15976	0	Hombre de 58 años de edad, de raza caucásica, con diagnóstico de EP predominante en temblor a los 44 años de edad.
doc_15976	1	Agonistas dopaminérgicos y tratamiento con levodopa permitieron un buen control sintomático.
doc_15976	2	A los 48 años de edad fue diagnosticado VIH en una prueba **rutinaria**.
doc_15976	3	Seis años después, aunque permaneció asintomático, el recuento de CD4 alcanzó 209 células/µl y se inició TARGA.
doc_15976	4	Вскоре, después, aparecieron síntomas gastrointestinales severos (náuseas, vómitos y diarrea) y discinesias a dosis **pico**, que se atribuyeron a las interacciones farmacocinéticas entre levodopa y TARGA.
doc_15976	5	Inicialmente, la levodopa se redujo a costa de un control subóptimo de la **EP**, pero posteriormente el **tratamiento antirretroviral** ha de suspenderse debido a discinesias intolerables.
doc_15976	6	Tras 3 años de buen control sintomático de la EP y infección por VIH asintomática, el paciente comenzó a sufrir fuertes fluctuaciones motrices con distonía de mañana y discinesias de dosis máxima.
doc_15976	7	En el momento de considerarse STN-DBS, estaba en tratamiento con levodopa de liberación inmediata y controlada y ropinirol, totalizando una dosis diaria de levodopa equivalente de 1.250 mg.
doc_15976	8	PD estuvo en estadio 3 de Hoehn-Yahr durante la medicación, y la puntuación UPDRS-III fue 78 sin medicación y 18 tras la ingesta de levodopa.

doc_n	line_n	**Fine-tuning**: clinical-NLLB:en2es
doc_15976	0	Un hombre de 58 años de edad, de raza caucásica, fue diagnosticado de EP predominante en temblor a los 44 años.
doc_15976	1	Los agonistas de dopamina y el tratamiento con levodopa permitieron un buen control sintomático.
doc_15976	2	A los 48 años, fue diagnosticado **de VIH** en una prueba de rutina.
doc_15976	3	Seis años después, aunque **permanecía** asintomático, el recuento de CD4 había alcanzado 209 células/µl, y se inició la TARGA.
doc_15976	4	Poco después, se presentaron síntomas gastrointestinales graves (náuseas, vómitos y diarrea) y discinesias de dosis máxima, atribuidas a interacciones farmacocinéticas entre levodopa y TARPA.
doc_15976	5	Inicialmente, la levodopa se redujo a costa de un control subóptimo de la **EP**, pero posteriormente se tuvo que suspender la TARHA por las discinesias intolerables.
doc_15976	6	Tras 3 años de buen control sintomático de la EP y infección asintomática por el VIH, la paciente comenzó a sufrir de fluctuaciones motoras **severas** con distonía **matinal** y discinesias de dosis máxima.
doc_15976	7	Para el momento de la consideración de STN-DBS, estaba en levodopa y ropinirol de liberación inmediata y controlada, con una dosis equivalente diaria de 1.250 mg.
doc_15976	8	La EP se encontraba en estadio 3 de Hoehn-Yahr mientras estaba en tratamiento, y la puntuación UPDRS-III fue de 78 fuera de tratamiento y de 18 tras el consumo de levodopa.

doc_n	line_n	text:src:English
doc_15976	0	A 58-year-old Caucasian man was diagnosed with tremor-predominant PD at the age of 44 years.
doc_15976	1	Dopamine agonists and levodopa therapy allowed a good symptomatic control.
doc_15976	2	By the age of 48 years, he was diagnosed with HIV on a routine testing.
doc_15976	3	Six years later, although he remained asymptomatic, the CD4 count had reached 209 cells/µl, and HAART was started.
doc_15976	4	Soon after, severe gastrointestinal symptoms (nausea, vomiting, and diarrhea) and peak-dose dyskinesias emerged, which were attributed to pharmacokinetic interactions between levodopa and HAART.
doc_15976	5	Initially, levodopa was reduced at the cost of suboptimal control of PD, but afterwards HAART had to be discontinued because of intolerable dyskinesias.
doc_15976	6	After 3 years of good symptomatic PD control and asymptomatic HIV infection, the patient began to suffer from severe motor fluctuations with morning off dystonia and peak-dose dyskinesias.
doc_15976	7	By the time STN-DBS was considered, he was on immediate and controlled-release levodopa and ropinirole, totaling a daily levodopa equivalent dose of 1,250 mg.
doc_15976	8	PD was in Hoehn-Yahr stage 3 while on medication, and the UPDRS-III score was 78 off medication and 18 after suprathreshold levodopa intake.

Рис. 3: (Figure:) Task-1 Cases/Sentences EN-ES Translation Examples: clinic-WMT21fb vs clinic-NLLB

term_n	**Transfer-learning**: clinical-WMT21fb:es2en	**Fine-tuning**: clinical-NLLB-200:es2en	Source: Spanish
term_1	Infantile paralysis	infantile paralysis	parálisis infantil
term_2	convulsive seizures	seizures	crisis convulsivas
term_5	deletion in chromosome 5 in the q15-q22 region	chromosome 5 deletion in the q15-q22 region	deleción en el cromosoma 5 en la región q15-q22
term_6	Familial adenomatous polyposis	familial adenomatous polyposis	poliposis adenomatosa familiar
term_9	Chromosomopathy	chromosomal disease	cromosomopatía
term_12	arterial hypertension	hypertension	hipertensión arterial
term_15	pT2bN0Mo clear cell renal adenocarcinoma	Renal clear cell adenocarcinoma pT2bN0Mo	adenocarcinoma renal de células claras pT2bN0Mo
term_17	hepatic lesions	liver lesions	lesiones hepáticas
term_18	Hepatic metastases	liver metastases	metástasis hepáticas
term_19	Metastatic renal cancer	metastatic renal cancer	cáncer renal metastásico
term_22	Deep vein thrombosis	deep vein thrombosis	trombosis venosa profunda
term_23	Asterixis	asterixis	asterixis
term_24	Aortic atheromatosis	aortic atheromatous disease	ateromatosis aórtica
term_29	hypothyroidism grade 2	grade 2 hypothyroidism	hipotiroidismo grado 2
term_30	Grade 3 hypertension	grade 3 hypertension	hipertensión arterial grado 3
term_31	Grade 3 diarrhea with secondary hypomagnesemia	grade 3 diarrhea with secondary hypomagnesemia	diarrea grado 3 con hipomagnesemia secundaria
term_32	Thrombocytopenia	thrombopenia	trombopenia
term_33	gastrointestinal toxicity	digestive toxicity	toxicidad digestiva
term_35	Recurrent respiratory tract infection	recurrent infectious respiratory	respiratoria infecciosa recurrente
term_36	Pulmonary nodule located in the upper lobe	pulmonary nodule located in the upper lobe	nódulo pulmonar localizado en el lóbulo superior
term_37	Loculated cystic lesion in LSD	Cystic lesion loculated in LSD	lesión quística loculada en LSD
term_38	Multicystic lesion	Multi-cystic lesion	lesión multiquística
term_43	MCVAP type I of LSD	LSD type I MCVAP	MCVAP tipo I del LSD
term_0	mild mental retardation	mild mental retardation	retraso mental leve
term_3	urinary tract infections	urinary tract infections	infecciones del tracto urinario
term_4	ITU) of repetition	ITU) of repetition	ITU) de repetición
term_7	deletion of this gene	deletion of this gene	deleción de dicho gen
term_8	deletion in chromosome 5	deletion in chromosome 5	deleción en el cromosoma 5
term_10	drug allergies	drug allergies	alergias medicamentosas
term_11	smoker	smoker	fumador
term_13	dyslipidemia	dyslipidemia	dislipemia
term_14	atrial fibrillation	atrial fibrillation	fibrilación auricular
term_16	macroscopic hematuria	macroscopic hematuria	hematuria macroscópica
term_20	hypothyroidism	hypothyroidism	hipotiroidismo
term_21	dehydration	dehydration	deshidratación
term_25	Cardiomegaly	Cardiomegaly	Cardiomegalia
term_26	anemia	anemia	anemia
term_27	hyponatremia secondary to diarrhea	hyponatremia secondary to diarrhea	hiponatremia secundaria al cuadro diarreico
term_28	sepsis	sepsis	sepsis
term_34	smoker	smoker	fumadora
term_39	cyst	cyst	quiste
term_40	microcytic anemia	microcytic anemia	anemia microcítica
term_41	ectopic pregnancy	ectopic pregnancy	embarazo ectópico
term_42	adenopathies	adenopathies	adenopatías

Рис. 4: (Figure:) Task-2 Clinical Term ES-EN Translation Examples: clinic-WMT21fb vs clinic-NLLB

the outputs of these two MMPLMs, such as on translating multi-word expressions in the clinical domain (Bhatia et al., 2023; Han, 2022). We also designed corresponding measurements on the evaluation of uncertainty and inter-rater reliability (IRR) levels (Gladkoff et al., 2022, 2023). Secondly, we think it is valuable to integrate more high-performance automatic metrics into the comparison such as hLEPOR (Han et al., 2021). Finally, we will try more external languages from different typologies in future work.

ont_n	**Transfer-learning**: clinical-WMT21fb (en2es)	**Fine-tuning**: clinical-NLLB (en2es)	Source:English
ont_0	Todos	Todos	All
ont_1	Anomalía de la altura corporal	Anomalías de la talla corporal	Abnormality of body height
ont_2	Displasia renal multiquística	Displasia renal multicística	Multicystic kidney dysplasia
ont_3	Displasia renal multiquística	Riñón displásico multicístico	Multicystic dysplastic kidney
ont_4	Riñón multiquístico	Riñones multicísticos	Multicystic kidneys
ont_5	Displasia renal multiquística	Displasia renal multicística	Multicystic renal dysplasia
ont_6	Modo de herencia	Modos de herencia	Mode of inheritance
ont_7	Herencia	Herencia	Inheritance
ont_8	Herencia autosómica dominante	Herencia autosómica dominante	Autosomal dominant inheritance
ont_9	autosómica dominante	Autosomal dominante	Autosomal dominant
ont_10	Forma autosómica dominante	Forma autosómica dominante	Autosomal dominant form
ont_11	Тип autosómico dominante	Tipo autosómico dominante	Autosomal dominant type
ont_12	Herencia autosómica recesiva	Herencia autosómica recesiva	Autosomal recessive inheritance
ont_13	autosómica recesiva	Autosomal recesivo	Autosomal recessive
ont_14	Forma autosómica recesiva	Forma autosómica recesiva	Autosomal recessive form
ont_15	Predisposición autosómica recesiva	Predisposición autosómica recesiva	Autosomal recessive predisposition
ont_16	Morfología anormal de los genitales internos femeninos	Morfología anormal de los genitales internos femeninos	Abnormal morphology of female internal genitalia
ont_17	Anomalía de los genitales internos femeninos	Anomalías de los genitales internos femeninos	Abnormality of female internal genitalia
ont_18	Anomalía funcional de la vejiga	Anomalías funcionales de la vejiga	Functional abnormality of the bladder
ont_19	Mal función vesical	Función vesical deficiente	Poor bladder function
ont_20	Infecciones urinarias de repetición	Infecciones urinarias recurrentes	Recurrent urinary tract infections
ont_21	Infecciones del tracto urinario frecuentes	Infecciones frecuentes del tracto urinario	Frequent urinary tract infections
ont_22	ITU recidivante	ITU recurrentes	Recurrent UTIs
ont_23	Infecciones vesicales de repetición	Infecciones vesiculares repetidas	Repeated bladder infections
ont_24	Infecciones urinarias de repetición	Infecciones urinarias repetidas	Repeated urinary tract infections
ont_25	Infecciones del tracto urinario	Infecciones del tracto urinario	Urinary tract infections
ont_26	Infecciones del tracto urinario, recurrentes	Infecciones del tracto urinario, recurrentes	Urinary tract infections, recurrent
ont_27	vejiga neurogénica	Vejícula neurogénica	Neurogenic bladder
ont_28	Falta de control vesical por lesión del sistema nervioso	Falta de control vesical por lesión del sistema nervioso	Lack of bladder control due to nervous system injury
ont_29	Urgencia urinaria	Urgencia urinaria	Urinary urgency
ont_30	vejiga hiperactiva	Vejícula hiperactiva	Overactive bladder
ont_31	Síndrome de vejiga hiperactiva	Síndrome de vejiga hiperactiva	Overactive bladder syndrome
ont_32	Síndrome de frecuencia de urgencia	Síndrome de frecuencia de urgencia	Urgency frequency syndrome
ont_33	Hipoplasia del útero	Hipoplasia del útero	Hypoplasia of the uterus
ont_34	Utero hipoplásico	Útero hipoplásico	Hypoplastic uterus
ont_35	Utero rudimentario	Útero rudimentario	Rudimentary uterus
ont_36	Utero pequeño	Útero pequeño	Small uterus
ont_37	Utero subdesarrollado	Útero subdesarrollado	Underdeveloped uterus
ont_38	Anomalía vesical	Anomalía vesical	Abnormality of the bladder
ont_39	Divertículo vesical	Divertículo vesical	Bladder diverticulum
ont_40	Divertículos vesicales	Divertículos vesiculares	Bladder diverticula
ont_41	Retención urinaria	Retención urinaria	Urinary retention
ont_42	Aumento del volumen residual de orina post-vacío	Aumento del volumen de orina residual post-vacío	Increased post-void residual urine volume
ont_43	La nicturia	Nocturia	Nocturia

Рис. 5: (Figure:) Task-3 Concept EN-ES Translation Examples: clinic-WMT21fb vs clinic-NLLB

Limitations

1) On PLM Capability for Transferring to New Language, in this work, we used Meta-AI's WMT21 multilingual pre-trained language models as our test-base for the knowledge transfer into an external language fine-tuning and translation. This new-language ability is much dependent on the MPLMs we used, such as WMT21fb (Tran et al., 2021) as a huge size model, a conditional generation from Meta-AI's massive M2M-100 model (Fan et al., 2021). If we try to fine-tune a bilingual model on an external language that the PLM did not see, it will not be that good because for smaller-sized models such fine-tuning would be too much of a change, and the model will lose generalisation which leads to problems. For huge multilingual PLM models, the 250K of fine-tuning data is a small set of numbers, and that's why the model does not lose generalisation and captures new data well without losing linguistic knowledge of other languages that it was trained on.

2) On the Impact of Language Families, the MMPLM WMT21fb we deployed has both alphabetic languages and CJK (Chinese, Japanese, Koran) character languages, as well as Slavic language (Russian). This might make it easier to transfer to a new language, e.g. alphabetic language. However, in situations when the MPLMs did not include any of the language scripts that belong to the language family of the target one, it can be much harder for it to transfer to the new target language. This needs further investigation. One possible extension of this work is using the dynamic vocabulary method proposed by Lakew et al. (2018).

Acknowledgements

The authors thank the ClinSpEn2022 shared task organisers for preparing the data set and evaluation platforms, and thank Darryl Estrada for communicating with us during the competition. We thank Dr. Alfredo Madrid Garcia and Ms. Cristina Sánchezfor carrying out the human evaluation for this work and thank anonymous reviewers for their valuable comments on our work. We thank the open research projects Meta-AI's wmt21.dense-24-wide.En-X (2021) and NLLB (2022) we used. This work has been partially supported by grant EP/V047949/1 "Integrating hospital outpatient letters into the healthcare data space"(funder: UKRI/EPSRC).

References

Kabir Ahuja, Shanu Kumar, Sandipan Dandapat, and Monojit Choudhury. 2022. Multi task learning for zero shot performance prediction of multilingual models. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5454–5467, Dublin, Ireland. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In Proceedings of the ACL.

Archna Bhatia, Kilian Evang, Marcos García, Voula Giouli, Lifeng Han, and Shiva Taslimipoor. 2023. Proceedings of the 19th workshop on multiword expressions (mwe 2023). In Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1723–1732, Beijing, China. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. J. Mach. Learn. Res., 22(1).

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 866–875, San Diego, California. Association for Computational Linguistics.

Serge Gladkoff and Lifeng Han. 2022. HOPE: A task-oriented and human-centric evaluation framework using professional post-editing towards more effective MT evaluation. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 13–21, Marseille, France. European Language Resources Association.

Serge Gladkoff, Lifeng Han, and Goran Nenadic. 2023. Student's t-distribution: On measuring the inter-rater reliability when the observations are scarce. arXiv preprint arXiv:2303.04526.

Serge Gladkoff, Irina Sorokina, Lifeng Han, and Alexandra Alekseeva. 2022. Measuring uncertainty in translation quality evaluation (TQE). In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 1454–1461, Marseille, France. European Language Resources Association.

Lifeng Han. 2022. An investigation into multi-word expressions in machine translation. Ph.D. thesis, Dublin City University.

Lifeng Han, Gleb Erofeev, Irina Sorokina, Serge Gladkoff, and Goran Nenadic. 2022a. Examining large pre-trained language models for machine translation: What you don't know about it. In Proceedings of the Seventh Conference on Machine Translation (WMT), pages 908–919, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Lifeng Han, Gleb Erofeev, Irina Sorokina, Serge Gladkoff, and Goran Nenadic. 2022b. Using massive multilingual pre-trained language models towards real zero-shot neural machine translation in clinical domain.

Lifeng Han, Irina Sorokina, Gleb Erofeev, and Serge Gladkoff. 2021. cushLEPOR: customising hLEPOR metric using optuna for higher agreement with human judgments or pre-trained language model LaBSE. In Proceedings of the Sixth Conference on Machine Translation, pages 1014–1023, Online. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics, 5:339–351.

Surafel M. Lakew, Aliia Erofeeva, Matteo Negri, Marcello Federico, and Marco Turchi. 2018. Transfer learning in multilingual neural machine translation with dynamic vocabulary. In Proceedings of the 15th International Conference on Spoken Language Translation, pages 54–61, Brussels. International Conference on Spoken Language Translation.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In Text Summarization Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 448–462, Online. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Xavier Baró, Hugo Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. 2022. Codalab competitions: An open source platform to organize scientific challenges. Technical report.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2685–2702, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 35–40, San Diego, California. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. CoRR, abs/2008.00401.

Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. 2021. Facebook ai's wmt21 news translation task submission. In Proc. of WMT.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Conference on Neural Information Processing System, pages 6000–6010.

Marta Villegas, Ander Intxaurrondo, Aitor Gonzalez-Agirre, Montserrat Marimon, and Martin Krallinger. 2018. The mespen resource for english-spanish medical machine translation and terminologies: census of parallel corpora, glossaries and term translations. LREC MultilingualBIO: multilingual biomedical text processing.

Hayahide Yamagishi, Shin Kanouchi, Takayuki Sato, and Mamoru Komachi. 2016. Controlling the voice of a sentence in Japanese-to-English neural machine translation. In Proceedings of the 3rd Workshop on Asian Translation (WAT2016), pages 203–210, Osaka, Japan. The COLING 2016 Organizing Committee.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

Tracking the Evolution of Covid-19 Symptoms through Clinical Conversations

Ticiana L. Coelho da Silva, José Antônio F. de Macêdo, Régis Pires Magalhães
Insight Data Science Lab
ÍRIS Lab, Ceará, Brazil
{ticianalc, jose.macedo, regis}@insightlab.ufc.br

Abstract

The Coronavirus pandemic has heightened the demand for technological solutions capable of gathering and monitoring data automatically, quickly, and securely. To achieve this need, the Plantão Coronavirus chatbot has been made available to the population of Ceará State in Brazil. This chatbot employs automated symptom detection technology through Natural Language Processing (NLP). The proposal of this work is a symptom tracker, which is a neural network that processes texts and captures symptoms in messages exchanged between citizens of the state and the Plantão Coronavirus nurse/doctor, i.e., clinical conversations. The model has the ability to recognize new patterns and has identified a high incidence of altered psychological behaviors, including anguish, anxiety, and sadness, among users who tested positive or negative for Covid-19. As a result, the tool has emphasized the importance of expanding coverage through community mental health services in the state.

1 Introduction

The Covid-19 pandemic required efficient and agile measures from governments to mitigate the effects caused by the disease. Plantão Coronavírus (PC) is a chatbot, and one of the solutions developed around April 2020 in Ceará State, Brazil, to hold back the pandemic. PC is an automated tool to converse with patients via text and provide them guidelines on how to proceed based on their level of emergency. The Plantão Coronavirus project was specifically developed by ÍRIS Lab of Innovation and Data in Ceará state, Brazil, in collaboration with the Health Secretary of Ceará State. Its purpose is to create a chatbot with artificial intelligence capabilities. The platform incorporates chatbot technology, enabling users to interact with an artificial intelligence system, and also offers the option to redirect to a virtual service manned by healthcare professionals such as doctors or nurses.

The interactions between patients and healthcare professionals through the PC generated a lot of clinical conversation data that needed to be mined, analyzed, and transformed into valuable information. The Health Secretary was required to track the signs of the disease, and it was not feasible to perform this task manually by reading thousands of texts. In this way, an automated and intelligent solution to classify the Covid-19 symptoms was essential. Especially at the beginning of the pandemic, when little was known about Covid-19.

This work proposes a solution to address the issue of screening the symptoms reported by users in a chat dialogue box. Ultimately, the chatbot classifies the user's health status as mild, moderate, or severe. Based on this classification, the system recommends various services to the user, including medical appointments or Covid-19 tests.

Our solution approaches the identification of symptoms in text as a Named Entity Recognition (NER) problem. NER identifies named entities in documents and categorizes them into predefined classes based on the type of entity (Li et al., 2020). Typically, a neural network is utilized for entity recognition. In this study, the named entity is a symptom.

However, in order to train a NER model for extracting symptoms and diseases from clinical conversations in Plantão Coronavírus, automatic annotation was necessary due to the impracticality of manual annotation given the large size of the dataset. Additionally, there was no publicly available NER model for Brazilian Portuguese that could extract symptoms and diseases. Therefore, we utilized ScispaCy (Neumann et al., 2019), a disease-focused NER model trained on English-language texts, and employed transfer learning to build our model. During the training process, we initially translated the Portuguese texts into English and used the ScispaCy model to analyze each English text, extracting the symptoms identified and

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 41–47
July 14, 2023 ©2023 Association for Computational Linguistics

translating them back into Portuguese. The training set comprised the original text and the Portuguese symptoms extracted by the ScispaCy model.

The intelligence developed in this study was crucial in identifying patterns of disease indicators, moreover, new or rare symptoms that had not yet been documented by researchers and health professionals in the state of Ceará. This enabled the tracking of the evolution of Covid-19 findings over time.

The entity recognition process was performed automatically. Related works as (Tarcar et al., 2019) achieved an F1-score of 78.5% and (Neumann et al., 2019) reached 84.94% of F1-score for the symptom/disease discovery model. In contrast, this work achieved F1-score equal to 85.66%. The F1-score is a metric that quantifies the harmonic mean of precision and recall. In our case, we report the F1-score for the test set. Our approach has made significant advances in tackling the disease in Ceará, given the possibility of virus mutations and the consequent appearance of new symptoms. With the capability to recognize new patterns, our model identified a high frequency of altered psychological behaviors, such as anxiety, anguish, and sadness, in both Covid-19 positive and negative users. As a result, our tool highlighted the need for the state to expand its mental health care services to the population through various channels. Our work has enabled the government to develop a public policy to address this need.

Moreover, less research is available for clinical texts in low resource languages as Brazilian Portuguese (Schneider et al., 2020). One can argue whether or not Brazilian Portuguese is an low resource language, but we say that, at least for some tasks, there are fewer resources compared to English or other languages, as discussed by (Costa et al., 2020) and (Fischer et al., 2022).

2 Background

2.1 Plantão Coronavírus Dataset

During the COVID-19 pandemic, the Brazilian state government of Ceará introduced Plantão Coronavírus, a web-based system designed to facilitate online patient consultation[1].

To better understand how the Plantão Coronavírus dataset was built, it is necessary to explain some aspects of the Plantão Coronavírus system. When a user initiates an interaction with the system, a virtual screening protocol categorizes the user into one of three risk profiles: severe, moderate, or mild. In the severe risk profile, the user reports severe symptoms directly related to Covid-19, such as shortness of breath, fever above 39ºC for more than 48 hours, etc. The moderate risk profile is for users who do not report severe symptoms of Covid-19 but may be at an increased risk for getting very sick from Covid-19, such as elderly over 70 years old, those with diabetes, asthma or chronic lung disease, sickle cell disease, or those who are immunocompromised, pregnant women, etc. The low-risk or mild-risk profile is for users who report being asymptomatic or having mild symptoms and do not belong to high-risk groups for Covid-19, such as a stuffy or runny nose, headaches, pain, etc. After categorization, the user interacts with a nurse who answers questions on several topics related to Covid-19, primary care, testing locations, etc.

Since the launch of Plantão Coronavírus, numerous consultations have been conducted, resulting in the recording of several clinical dialogues between patients, nurses, and doctors. A subset of these recordings was utilized to create a dataset for training and testing a neural network that identifies Covid-19 symptoms in Brazilian Portuguese texts. To the best of the author's knowledge, no existing model was available to recognize symptoms in Brazilian Portuguese texts.

The criteria for classifying a user into one of the three categories have been subject to changes over time, following new guidelines from the Health State Secretary and the World Health Organization (WHO). Moreover, as the people's expertise in dealing with the pandemic has grown, the understanding of categorizing users has also evolved. Consequently, the recorded dialogues between users and healthcare professionals vary significantly depending on the period considered. This variability can introduce noise into the dataset. To minimize this effect, we used data from two months (April/20 and May/20) when the same Covid-19 protocol was followed in the state of Ceará. This was the longest continuous period we could identify when the same protocol was in place, allowing us to build a more robust dataset.

Table 1 presents an example of a clinical conversation between a patient and a doctor. The dataset, we used in this work, includes approximately 27,690 dialogues, with 577,814 utterances,

[1] https://coronavirus.ceara.gov.br/

an average of 21.48 turns per dialogue, and an average of 9.9 words per utterance. As a reminder, an utterance in a dialogue refers to a complete unit of speech produced by one speaker, which can be a sentence, phrase, or even a single word. In contrast, a turn refers to a sequence of utterances produced by one speaker before the other speaker takes a turn in the conversation. A turn may consist of one or more utterances. For privacy reasons, data is not publicly available.

Doctor	Hi, I'm Doctor Fabio. How can I help you?
Patient	I'm experiencing three days of fever and a dry cough.
Doctor	Did you get to take the temperature?
Patient	Yes.
Doctor	Do you feel shortness of breath?
Patient	During this period of fever, I took Dipirone...
Patient	A little, I think because of the cough
Doctor	right
Doctor	Do you feel shortness of breath when you walk?
Patient	only when I cough.
Doctor	I recommend taking a COVID test...
...	...

Table 1: Plantão Coronavírus Dialogue example

2.2 Related Works

Researchers have lately worked to create Covid-19-related chatbots due to the significant demand for patient follow-ups. Using research papers from the Covid-19 Open Research Dataset and CORD-19 (Wang et al., 2020), (Lei et al., 2021) trained a NER model. The research group used the papers to extract entities to identify symptoms in the patient's written sentences. The most prevalent symptoms in the articles are found using word clouds, and a knowledge graph was created using the chatbot NLU model to keep track of follow-up appointments with returning patients. (Fazzinga et al., 2021) uses argumentation graphs and natural language to create dialog systems explaining Covid-19 vaccination.

The paper (Miner et al., 2020) outlines issues and queries that a chatbot may handle during a pandemic like Covid-19. Initiatives like Clara [2] from the CDC in the United States aim to combat the proliferation of contradicting information brought on by ignorance and fake news, which can eventually make handling the pandemic crisis much more challenging.

(Schaeffer et al., 2022) provides a dataset based on three corpora: the first one contains 70 carefully annotated tweets and 10 transcriptions of YouTube videos. The second corpus comprises the same textual material with the named entities annotated. 100 YouTube transcriptions that were automatically tagged using NER models included the third corpus. The dataset provides geographic information, city names, and epidemiological data, such as diseases, symptoms, and virus entities. (Schaeffer et al., 2022) can be used to train and evaluate different NER models.

(Beltagy et al., 2019) proposes SciBERT, a pre-trained BERT-based language model for scientific data. To enhance performance on downstream scientific NLP tasks, SciBERT uses unsupervised pre-training on a large multi-domain corpus of scientific literature. The paper (Beltagy et al., 2019) evaluates the performance of SciBERT for different NLP tasks such as NER, Relation Classification, and Text Classification, among others. BioBERTpt (Schneider et al., 2020) was developed using clinical notes and scientific abstracts. BioBERTpt is NER model BERT-based for Portuguese texts. However, the paper does not clarify the quality of the model to extract the named entity disease. We might experiment BioBERTpt and maybe enrich our annotated data, instead of using only ScispaCy to annotated our training data. It might be a future study direction.

The authors (Lopes et al., 2019) manually collected and annotated a corpus of Portuguese clinical texts, identifying named entities such as characterization, test, evolution, genetics, additional observations, results, date and time, therapeutics, among others. They also evaluated the effectiveness of various state-of-the-art models for named entity recognition. While their work is relevant to ours, we are particularly interested in named entities related to diseases or symptoms, which were not included in the dataset used in (Lopes et al., 2019).

(Schäfer et al., 2022) explores two main approaches. Firstly, the authors investigate the application of English models to translated texts, followed by the transfer of predicted annotations back

[2]https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/testing.html

to the source language. This direction closely aligns with the approach adopted in our paper to create our training set. Secondly, the authors explore the possibility of utilizing existing high-quality annotations to train NLP models in the target language, going beyond mere translation. Given the scarcity of resources for low-resource languages, the ideia is to employ English models and external biomedical and clinical datasets as substitutes. The primary objective is to assess the potential benefits for low-resource languages by leveraging the existing resources available in English. The findings in (Schäfer et al., 2022) indicate that English language models can indeed be applied to other languages in clinical contexts. Translated training data can serve as a solid foundation in languages where resources are otherwise lacking. The success of the second approach depends on both the annotation standards and the similarity between English and the low-resource language in terms of grammar and morphology.

The study (Schäfer et al., 2022) further supports the suitability of our methodology. A potential future direction could involve evaluating our translation phase, similar to the approach described in (Schäfer et al., 2022), through word alignment using contextualized embeddings with the assistance of multilingual BERT.

3 Data and Methods

The symptom capture mechanism is the primary contribution of this work and is integrated into the entire triage process, from the chatbot to the teleservice with health professionals. Our solution is a technology that leverages Plantão Coronavírus data to process and identify symptoms contained in natural language texts. In the following sections, we describe the pipeline for building a tracker that monitors the evolution of Covid-19 symptoms.

Note that the sentences from Plantão Coronavírus were not annotated with symptoms. The main challenge at this point was to build a training set using this data. The pipeline from data collection to implementation of the NER model is shown in Figure 1 and described as follows.

3.1 Annotate the Brazilian Portuguese dataset with Covid-19 symptoms

The detection of symptoms in the Portuguese language was a challenge because, at the beginning of the pandemic, no publicly available model could

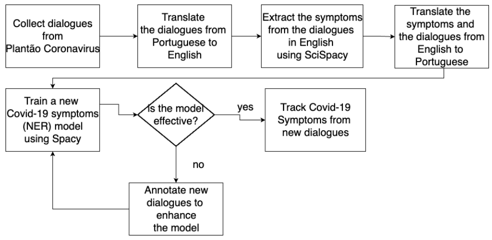


Figure 1: Pipeline to build and deploy the tracker of Covid-19 symptoms evolution.

perform this task, according to the authors' knowledge.

Our solution is based on Transfer Learning (Pan and Yang, 2009). The technological innovation provided by our solution is a pioneering neural model for recognizing symptoms in Brazilian Portuguese, mainly because the Portuguese language lacks NER models. The transfer learning technique uses the knowledge gained by solving one problem and applying it to a different but related problem, allowing for rapid progress and improved performance when modeling the second task. In other words, transfer of learning is the improvement of learning in a new task by transferring knowledge from a related task that has already been learned.

So after collecting the dialogues from Plantão Coronavírus, we chose to translate the texts that were initially in Portuguese into English. Then, submit each text (in English) to the ScispaCy model (Neumann et al., 2019) as an input parameter. For this work, the model used from ScispaCy was *en_ner_bc5cdr_md*. Then, we analyzed the result generated by this model and translated the symptoms captured by the ScispaCy model from English to Portuguese. All in all, the training set of our NER model comprises the original dialog text and the symptoms captured by the ScispaCy model in Portuguese.

The Google Translate was used in the text translation stages within our pipeline. Nowadays, these translation networks present very accurate results to the expected ones, making the noise insignificant when analyzed in the context of this work.

Another important aspect to consider is the alignment of words or tokens. When translating clinical conversations from Portuguese to English and annotating them using ScispaCy, all named entities are outputted in English by the NER model. Subsequently, we translate them back to Brazilian Portuguese and use a script to locate them in the original text. Our script also generates a training set in spaCy file format, indicating the occurrences of named entities in the original text from Plantão Coronavírus.

3.2 Train and Evaluate the NER model

As the NER component, we utilize spaCy NER [3]. Given its powerful neural network-based model's cutting-edge performance, this off-the-shelf NER technology is typically chosen for use in many industrial applications (Honnibal et al., 2020). The entity recognizer may be updated with new instances using an existing pre-trained statistical model thanks to SpaCy's support for online learning.

The embed, encode, attend, and forecast steps follow the basic four-step methodology used by the spaCy NLP models, particularly NER. The model first takes the text as input and converts the words into distinct number values. Prefix, suffix, shape, and lowercase characteristics are employed in the embedding step to extract the commonalities between the words. The values are sent through a CNN (Convolutional Neural Network) network to encode the context-free embeddings, creating a context-sensitive sentence matrix. The matrix must travel through the CNN Attention layer before transforming the prediction into a single vector. A common Multi-layer Perceptron (MLP) with a Softmax layer is then utilized as a tag decoder layer for class prediction. After training, the spaCy model is prepared for various NLP tasks.

Initially, we trained the NER model with a total of 27,690 dialogues in Portuguese. The dataset contains at least one symptom annotated per sentence. So, we split the dataset into training (22,152 dialogues) and test (5,538 dialogues) sets following the distribution of 80% and 20%, respectively. We kept training the NER model until it achieved an F1-score equivalent to ScispaCy, i.e., 85.02. In the end, it was possible to reach in terms of F1-score of 85,6 for our NER model.

3.3 Deployment of the NER model

A relevant aspect to point out of the NER model to extract Covid-19 symptoms in this work is the absence of manually annotating the data, usually performed by a human for entity recognition. In a scenario where there was a vast amount of data, and little time to process this information, the gain from optimizing this training step was crucial in

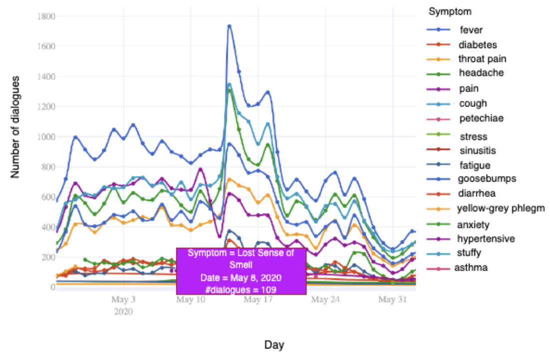


Figure 2: The Evolution of Covid-19 Symptoms and Related Diseases.

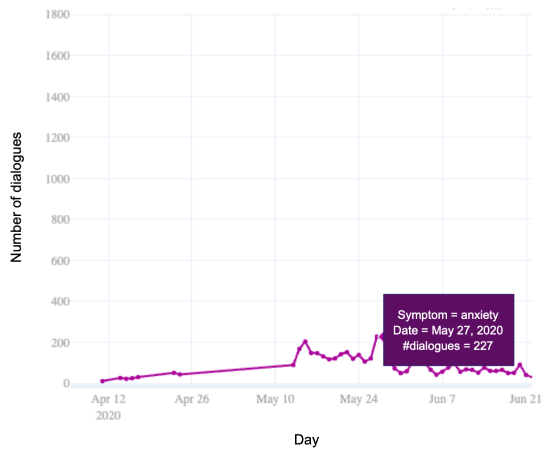


Figure 3: Repeatedly reports of anxiety along the days.

supporting decision-making.

Another innovative aspect of our approach is its ability to recognize mental health behaviors, enabling health professionals to develop and promote public policies to assist individuals affected by issues beyond the scope of epidemiology.

Currently, our solution is used in the Tele-Service platform of the State of Ceará, where it plays a pioneering role in the Health domain [4].

4 Experimental Results

Figure 2 demonstrates the evolution of symptoms in a time series. Not all symptoms could be shown in the figure's legend. World Health Organization (WHO), at the beginning of the pandemic, stated a set of symptoms commonly reported by people who got positive for Covid-19 very related to flu, like nasal flaring, runny nose, malaise, fever, cough, sore throat, diarrhea, headache, and no appetite.

However, new symptoms came along. Through Figure 2, it is possible to identify the detection

[3] https://spacy.io/api/entityrecognizer

[4] https://coronavirus.ceara.gov.br/

of a new symptom of loss of smell on the 8th of May. This symptom appeared and became quite characteristic of Covid-19 after a certain period. Still, on the time series, it is possible to see that the frequency of each symptom is seasonal during the analyzed period.

Diabetes is not tecnically a symptom, however our NER model considered as a symptom, possibly due to contextual factors or biases introduced by the scispaCy model. However, it is worth mentioning that there has been an observed rise in hyperglycemic conditions associated with COVID-19, particularly in patients with diabetes and those receiving steroid treatment (Lim et al., 2021). Nevertheless, this topic falls outside the scope of the current paper.

Another interesting point is that our NER model could capture some symptoms related to altered psychological behaviors, such as anxiety, mental confusion, neurological disorder, and disorientation among the symptoms, as highlighted in Figure 3. Figures 2 and 3 do not represent the entire population as a whole, but rather the number of dialogues on the Plantão Coronavírus platform that report specific symptoms.

As already reported in the previous session, our Covid-19 symptom tracker achieved an F1-score of 85.66, which is competitive compared to the SciSpacy English model, which has an F1-score of 85.02. To mitigate catastrophic forgetting of old knowledge as we update our NER model, we kept including new sentences from Plantão Coronavírus with the following symptoms annotated, such as breathing difficulty, mental confusion, loss of smell, loss of taste, tiredness, anxiety, anosmia, neurological disorder, and disorientation, so the model could learn not only from the frequent symptoms identified by *en_ner_bc5cdr_md* from SciSpacy and the symptoms commonly reported by people who got positive to Covid-19.

5 Conclusion and Future Works

This research provides a NER model to recognize Covid-19 symptoms in Portuguese textual conversations. At the start of the pandemic, no model could automatically identify the symptoms in a text written in Brazilian Portuguese; instead, we utilized ScispaCy, an English-language NER model for diseases, which through transfer learning, trained our NER model.

The texts were initially translated from Portuguese into English as part of the training procedure. The ScispaCy model then processes each English-language input text, and its identified symptoms are subsequently translated from English to Portuguese. The original text and the Portuguese symptoms determined by the ScispaCy model comprise the training set.

On the Plantão Coronavírus dataset, our NER model achieved an F1-score of 85.66, which is competitive with the English model of ScispaCy, which has an F1-score of 85.02. The NER model has brought to light the necessity for the state to increase its coverage of mental health services through the community mental health channel.

As a future research direction, we intend to extend the NER model to other diseases prevalent in Brazil, such as influenza, and explore various neural architectures. Another future work is to investigate the translation performance. One alternative might be manually translate a small sample of our dataset and then compute the BLEU score of the automatic translations on this sample for a more accurate estimate. As mentioned earlier, we selected data from April and May 2020, which corresponds to the early stages of the pandemic when Plantão Coronavírus followed the same protocol. As a future study, we can also examine the evolution of “altered physiological behaviors" in the dialogues that occurred after this time period.

Acknowledgments

The research reported in this work received support from the FUNCAP projects titled “Big Data Platform to Accelerate the Digital Transformation of Ceará State", “Citizen Platform" and “Digital Government" under the numbers 04772314/2020, 04772551/2020 and 04772420/2020, respectively. Part of the results presented in this work were obtained also through the project “Center of Excellence in Artificial Intelligence - AI4WELLNESS", funded by Samsung Eletrônica da Amazônia Ltda., under the scope of the Informatics Law no. 8,248/91”.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: Pretrained language model for scientific text. In *EMNLP*.

Felipe Almeida Costa, Thiago Castro Ferreira, Adriana Pagano, and Wagner Meira. 2020. Building the first english-brazilian portuguese corpus for automatic post-editing. In *Proceedings of the 28th international conference on computational linguistics*, pages 6063–6069.

Bettina Fazzinga, Andrea Galassi, and Paolo Torroni. 2021. An argumentative dialogue system for covid-19 vaccine information. In *International Conference on Logic and Argumentation*, pages 477–485. Springer.

Marcelo Fischer, Rejwanul Haque, Paul Stynes, and Pramod Pathak. 2022. Identifying fake news in brazilian portuguese. In *International Conference on Applications of Natural Language to Information Systems*, pages 111–118. Springer.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Hannah Lei, Weiqi Lu, Alan Ji, Emmett Bertram, Paul Gao, Xiaoqian Jiang, and Arko Barman. 2021. Covid-19 smart chatbot prototype for patient monitoring. *arXiv preprint arXiv:2103.06816*.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Soo Lim, Jae Hyun Bae, Hyuk-Sang Kwon, and Michael A Nauck. 2021. Covid-19 and diabetes mellitus: from pathophysiology to clinical management. *Nature Reviews Endocrinology*, 17(1):11–30.

Fábio Lopes, César Teixeira, and Hugo Gonçalo Oliveira. 2019. Contributions to clinical named entity recognition in portuguese. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 223–233.

Adam S Miner, Liliana Laranjo, and A Baki Kocaballi. 2020. Chatbots in the fight against the covid-19 pandemic. *NPJ digital medicine*, 3(1):1–4.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327.

Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Camille Schaeffer, Roberto Interdonato, Renaud Lancelot, Mathieu Roche, and Maguelonne Teisseire. 2022. Labeled entities from social media data related to avian influenza disease. *Data in Brief*, page 108317.

Henning Schäfer, Ahmad Idrissi-Yaghir, Peter Horn, and Christoph Friedrich. 2022. Cross-language transfer of high-quality annotations: Combining neural machine translation with cross-linguistic span alignment to apply ner to clinical texts in a low-resource language. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 53–62.

Elisa Terumi Rubel Schneider, João Vitor Andrioli de Souza, Julien Knafou, Lucas Emanuel Silva e Oliveira, Jenny Copara, Yohan Bonescki Gumiel, Lucas Ferro Antunes de Oliveira, Emerson Cabrera Paraiso, Douglas Teodoro, and Cláudia Maria Cabral Moro Barra. 2020. Biobertpt-a portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72.

Amogh Kamat Tarcar, Aashis Tiwari, Vineet Naique Dhaimodker, Penjo Rebelo, Rahul Desai, and Dattaraj Rao. 2019. Healthcare ner models using language model pretraining. *arXiv preprint arXiv:1910.11241*.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, et al. 2020. Cord-19: The covid-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Aligning Factual Consistency for Clinical Studies Summarization through Reinforcement Learning

Xiangru Tang♠ **Arman Cohan**♠ **Mark Gerstein**♠
♠ Yale University, New Haven, CT 06520, USA

{xiangru.tang, arman.cohan, mark.gerstein}@yale.edu

Abstract

In the rapidly evolving landscape of medical research, accurate and concise summarization of clinical studies is crucial to support evidence-based practice. This paper presents a novel approach to clinical studies summarization, leveraging reinforcement learning to enhance factual consistency and align with human annotator preferences. Our work focuses on two tasks: Conclusion Generation and Review Generation. We train a CONFIT summarization model that outperforms GPT-3 and previous state-of-the-art models on the same datasets and collects expert and crowd-worker annotations to evaluate the quality and factual consistency of the generated summaries. These annotations enable us to measure the correlation of various automatic metrics, including modern factual evaluation metrics like QAFactEval, with human-assessed factual consistency. By employing top-correlated metrics as objectives for a reinforcement learning model, we demonstrate improved factuality in generated summaries that are preferred by human annotators.

1 Introduction

Recently, the exponential growth of medical literature, specifically in the realm of clinical studies such as randomized controlled trials (RCTs), has underscored the necessity for efficient summarization techniques (Cohan et al., 2018; Sotudeh Gharebagh et al., 2020; Guo et al., 2021; Wang et al., 2020; Luo et al., 2022). Clinicians and researchers face the arduous task of sifting through vast amounts of information daily to remain abreast of the latest findings and advancements in their respective fields (Abacha et al., 2021; Chaves et al., 2022). Summarizing clinical studies enables healthcare professionals to access crucial information more rapidly, ensuring that their decisions and treatment plans are informed by the most recent, evidence-based knowledge. As a result, the development of effective and accurate summarization techniques for clinical studies has become an essential area of research in the medical domain (Shieh et al., 2019a; Wang et al., 2021; Wallace et al., 2021; DeYoung et al., 2021; Xie et al., 2022; Otmakhova et al., 2022a; Tang et al., 2023).

Automatic summarization of clinical studies is fundamental for systems that aim to interpret the vast array of available medical literature (Shieh et al., 2019a; Sotudeh Gharebagh et al., 2020; Otmakhova et al., 2022b; Tangsali et al., 2022). Randomized controlled trials (RCTs) are considered the gold standard of clinical evidence among various study types, including cohort studies, observational studies, and case studies (Concato et al., 2000; Katsimpras and Paliouras, 2022). The ability to efficiently process and summarize the massive volume of RCTs holds great potential for enhancing clinical decision-making (Meldrum, 2000; Ramprasad et al., 2023).

To delve deeper into clinical study summarization, we simultaneously explore single-document and multi-document summarization techniques. For single-document summarization, we propose an RCT conclusion generation task based on the PubMed 200k RCT sentence classification dataset (Dernoncourt and Lee, 2017). We utilize the PubMed RCT200k dataset (Dernoncourt and Lee, 2017), with original annotations for concluding sentences, meaning our summarization system's objective is to generate concluding sentences for a clinical study. In the case of multi-document summarization, we examine the challenge of automatically generating a narrative biomedical summary from multiple trial reports. Here inputs are titles and abstracts from systematic reviews previously conducted by members of the Cochrane collaboration[1] (Wallace et al., 2021), using the review abstract as our target, shown as Figure. 1.

Ensuring the factual consistency of summaries is vital in the medical field, as they must precisely

[1]https://www.cochrane.org/

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 48–58
July 14, 2023 ©2023 Association for Computational Linguistics

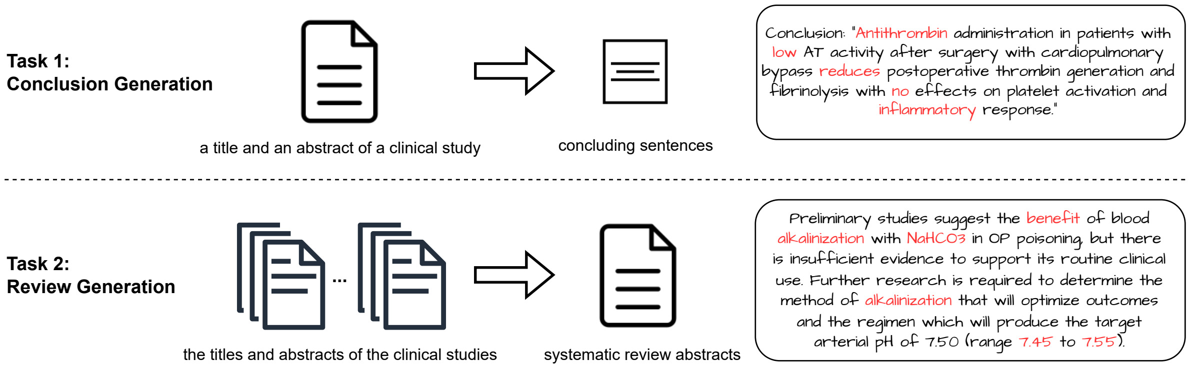


Figure 1: The diagram provides an overview of our tasks, which include single-document summarization and multi-document summarization. For the single-document summarization task, the input consists of a title and an abstract of a clinical study, with the goal being to generate concluding sentences. On the other hand, the input for the multi-document summarization task consists of the titles and abstracts from a corresponding review. Highlighted in red are various specialized medical concepts, logical reasoning, and numerical understanding, which introduce new challenges for clinical study summarization.

convey evidence to readers who make decisions for real patients. Wallace et al. demonstrated that modern summarization systems often struggle to create factually consistent summaries and tend to generate content with factual discrepancies compared to the input. At the same time, traditional automatic evaluation metrics have been deemed insufficient for assessing correctness, leading to a reliance on human evaluation for verifying generated summaries (Kryscinski et al., 2020; Maynez et al., 2020; Xie and Wang, 2023). However, such human evaluation demands medical expertise, which can be both expensive and challenging to scale. To tackle this issue, our work focuses on evaluating various automated metrics for their correlation with factual consistency and improving the factual consistency of clinical study summarization systems. In our approach, we utilize the top-correlated metrics from the previous experiment as the objective for a reinforcement learning (RL) model, like previous work (Paulus et al.). By doing so, we aim to guide the model toward generating more factually consistent summaries. Our results show that the RL-based models exhibit improved factuality and are preferred by human annotators, demonstrating the effectiveness of using RL for enhancing factual consistency in clinical study summarization systems.

Our main contributions: We emphasize our focus on clinical studies and discuss the unique challenges associated with their summarization. Our experiments feature comprehensive benchmarks and modern factual evaluation metrics, such as QAFactEval (Fabbri et al., 2022). We gathered annotations from both crowd workers and domain experts to assess the factual correctness of summaries generated by state-of-the-art models. By utilizing the top-correlated metrics as the objective for a reinforcement learning (RL) model, our results demonstrate improved factuality that is preferred by human annotators, showcasing the effectiveness of our approach.

2 Related Work

2.1 Clinical Trial Summarization

Clinical trial summarization has emerged as an important area of research due to the increasing volume of medical literature and the need for efficient information extraction. Early clinical trial summarization techniques often employed rule-based and template-based approaches, which relied on predefined templates and hand-crafted rules to generate summaries. For example, Demner-Fushman and Lin utilized a rule-based system to extract PICO elements from clinical trial abstracts. However, these methods were limited by their reliance on predefined templates and rules, which made them less adaptable to various domains and less effective in capturing the nuances of clinical trials. As machine learning gained traction, researchers began to explore feature-based approaches for clinical trial summarization. For instance, (Shieh et al., 2019b) worked towards understanding medical randomized controlled trials by conclusion genera-

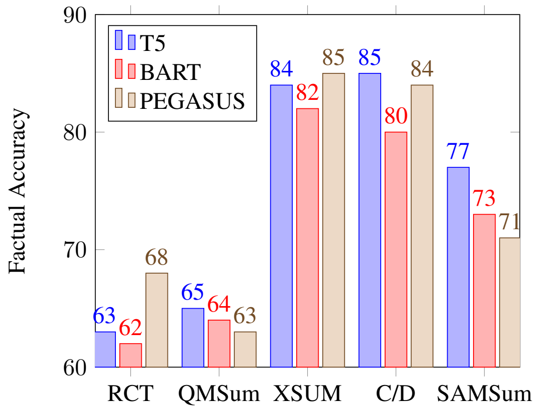


Figure 2: Bar chart illustrating the factual accuracy scores (higher is better) for different text summarization models (T5, BART, PEGASUS) on various datasets. The datasets are represented by the following abbreviations: RCT (RCT200K), QMSum (QMSum), XSUM (XSUM), C/D (CNNDM), and SAMSum (SAMSum).

tion. (Wallace et al., 2021) generated narrative summaries of RCTs with neural multi-document summarization. Although these methods showed promise, they still required significant manual feature engineering and were sensitive to the choice of features. The advent of deep learning has led to substantial improvements in clinical trial summarization. Neural network-based models, such as sequence-to-sequence models, have been employed for summarizing clinical trials. For example, (DeYoung et al., 2020) presented Evidence Inference 2.0, which focused on more data and better models in the biomedical domain. Additionally, (DeYoung et al., 2021) introduced MS2, a multi-document summarization approach for medical studies. These studies demonstrated superior performance compared to traditional machine learning methods.

2.2 Factual Consistency in Summarization

Factual consistency is a critical aspect of text summarization, as it ensures that generated summaries accurately represent the source content (Maynez et al., 2020). Previous works have discussed the challenges associated with achieving factual consistency, including issues like hallucination (Zhang et al., 2022a; Sridhar and Visser, 2022; van der Poel et al., 2022), and various techniques employed to address these challenges, such as reinforcement learning (Wan and Bansal, 2022a) and model fine-tuning (Zhang et al., 2022a; Wan and Bansal, 2022b; Tang et al., 2022b). Numerous existing models have attempted to address this issue, including extractive (Zhang et al., 2022b), abstractive (Ladhak et al., 2022; Chen et al., 2021; Wan and Bansal, 2022b). Additionally, several researchers have proposed better evaluation metrics to assess factual inconsistency, such as QAFactEval (Fabbri et al., 2022) and FactCC (Kryscinski et al., 2020). Promising avenues for future research may utilize high-quality negative examples (Wang et al., 2022), better evaluation metrics (Luo et al., 2023), and novel model architectures (Chaudhury et al., 2022) to improve the factual consistency of generated summaries. And recently, there has been a growing interest in the generation of surveys (Li et al., 2021). However, there is currently limited discussion concerning factual consistency issues in clinical survey or studies summarization.

3 Preliminarily

3.1 Datasets and Formulation of Tasks

We present two tasks:

Task 1: Conclusion Generation. We employ a modified version of the PubMed 200k RCT dataset, initially designed for sequential sentence classification. This dataset emphasizes medical abstracts, particularly randomized controlled trials (RCTs), which are considered the most reliable source of medical evidence. Each sentence in the dataset is labeled with a specific class that corresponds to the section it originates from Objective, Background, Conclusions, Methods, and Results. For inclusion, abstracts must (1) pertain to an RCT and (2) be structured. Among the 195,654 abstracts meeting both criteria, we allocate them into training (190,654), validation (2,500), and testing (2,500) sets. We assemble sentences from the Objective, Background, Methods, and Results classes of each abstract into a single paragraph, which serves as the input text for summarization. Sentences marked as Conclusions function as the reference summary for the medical abstract.

Task 2: Review Generation. Our dataset comprises systematic review abstracts as well as the titles and abstracts of the clinical trials summarized within these reviews. All data is sourced from PubMed, exclusively using abstracts. On average, each review encompasses ten trials, featuring an average abstract length of 245 words. We employ the "authors' conclusions" subsection of the systematic review abstract as our target summary (with an average of 75 words). The dataset is randomly partitioned into training (3,619 reviews), development

(455 reviews), and testing (454 reviews) sets.

Additionally, focusing on clinical trials in comparison to other summarization tasks, such as news, meetings, dialogues, movies, emails, sports, or games, is driven by the unique challenges presented in scientific and clinical summarization. Clinical trial summaries often require more logical reasoning, numerical understanding, and the accurate representation of a vast array of domain-specific terminology. The importance of correctly conveying specific drug values, concentrations, and scales adds significant complexity to summarization models. In an evaluation we conducted, we crowdsourced 100 different summarization examples (including CNNDM (Hermann et al., 2015), XSUM (Narayan et al., 2018), SAMSum (Gliwa et al., 2019), QMSum (Zhong et al., 2021), and RCT200K) and applied state-of-the-art summarization models, such as Pegasus (Zhang et al., 2020), BART (Lewis et al., 2020), and T5 (Raffel et al., 2020). We asked the participants to assess the factual consistency of the generated summaries and found that clinical summarization exhibited the lowest factuality accuracy (highest error rate). This highlights the challenge and importance of focusing on clinical trial summarization, ensuring accuracy and consistency in conveying critical information.

3.2 Meta evaluation

First, we assess the performance of automatic metrics using 200 test summaries for two tasks, in Table. 1 and Tabel. 2. We examine the results obtained using fine-tuned PEGASUS (Zhang et al., 2020), CONFIT (Tang et al., 2022b), BART models, as well as zero-shot GPT-3 summaries. Here, we use GPT-3 davinci model. ConFiT introduces a novel training approach that enhances the factual accuracy and overall quality of summaries through contrastive fine-tuning, emphasizing error identification. Although the original study focused on dialogue summarization, we adapt and fine-tune the approach for clinical summarization.

Additionally, we engage both domain experts and general crowd-workers to evaluate the generated summaries. Following the methodology of (Tang et al., 2022a), we employ a 10-point Likert scale for expert and crowd-worker annotators to assess factual consistency. For each summary, we have two crowd-workers and one expert providing scores, and we take the average of these three ratings. We also adopt a block design in our evaluation process: each crowd-worker evaluates 10 summaries, while we engage two MD students with medical bachelor's degrees, who each assess 100 samples. We employed the AWS Mechanical Turk (MTurk) platform to engage crowd-workers for our task. Each crowd-worker annotator received a $10 compensation. We recruited 40 MTurk workers with strong track records, using these qualifications: a HIT approval rate of at least 99%, a minimum of 200 approved HITs, and residence in one of the following English native-speaking countries: US, Australia, Canada, New Zealand, or UK.

Table. 1 displays the performance of the summarization systems evaluated using automatic reference-based evaluation metrics, such as ROUGE, METEOR, and BLEU, highlighting the differences in scores between the models for each task. The model with the highest scores across most metrics for Task 1 is `CONFIT`, while for Task 2, it is again `CONFIT`. Notably, `CONFIT`outperforms the other models in terms of ROUGE, METEOR, and EM and F1 of QAEval.

Table. 2 showcases the performance of the same summarization systems evaluated using automatic reference-free metrics, including SUPERT(Gao et al., 2020), BLANC (Vasilyev et al., 2020), QuestEval (Scialom et al., 2021), QAFactEval (Fabbri et al., 2022), FactCC (Kryscinski et al., 2020), DAE (Goyal and Durrett, 2021), and SummaC (Laban et al., 2022). This table provides an alternative perspective on the performance of these models, as it does not rely on reference summaries for evaluation. However, PEGASUS has the highest SUPERT score and the highest QuestEval score. Meanwhile, for Task 2, PEGASUS again scores the highest in SUPERT, while BART achieves the highest QuestEval score.

Table. 3 presents the instance-level Pearson correlation of various metrics with human factual consistency ratings on Task 1 and Task 2. This table helps identify which metrics are more strongly correlated with factual consistency, providing insights into the most reliable evaluation methods for measuring the factuality of generated summaries. We can observe that DAE (0.5743 for Task 1 and 0.2089 for Task 2) and QAFactEval (0.5516 for Task 1 and 0.3147 for Task 2) have the highest correlation with human consistency ratings. This indicates that DAE and QAEval are more closely aligned with human judgment when evaluating factual consistency for clinical studies summarization.

Dataset	Model	Overlap-Based ROUGE(1/2/L)	METEOR	BLEU	QAEval EM	F1
Task 1	PEGASUS	34.96/14.75/28.35	.23	7.2	.104	.159
	CONFIT	38.55/17.15/31.52	.32	6.5	.136	.210
	BART	35.10/13.90/28.52	.24	5.8	.098	.162
	GPT-3	31.92/11.38/24.77	.24	3.7	.097	.158
Task 2	PEGASUS	33.15/13.60/26.90	.21	.13	.094	.155
	CONFIT	37.40/16.50/30.40	.30	.15	.136	.202
	BART	25.67/9.55/21.47	.19	.03	.072	.126
	GPT-3	27.48/10.72/22.24	.21	.04	.084	.143

Table 1: Performance of various summarization systems evaluated using automatic reference-based evaluation metrics.

Dataset	Model	Overall Quality SUPERT	BLANC	Factuality (QA-based) QuestEval	QAFactEval	Factuality (NLI-based) FactCC	DAE	SummaC
Task 1	PEGASUS	.5475	.0615	.7380	4.4095	.3750	.8235	.1145
	CONFIT	.5596	.0813	.7343	3.8354	.1823	.7587	-.0521
	BART	.5348	.0564	.7803	3.7537	.2021	.7568	-.0594
	GPT-3	.5579	.0751	.7268	3.6419	.2438	.6682	-.0719
Task 2	PEGASUS	.6292	.1144	.7129	4.2141	.7223	.7966	.2425
	CONFIT	.5910	.0910	.7360	3.6820	.2510	.7410	.0110
	BART	.5443	.0658	.7529	3.5814	.2837	.7382	.0279
	GPT-3	.5411	.0606	.7173	3.2352	.3998	.6574	-.0711

Table 2: Performance of various summarization systems evaluated using automatic reference-free metrics.

Metric	Task 1	Task 2
ROUGE(1/2/L)	0.2721	0.0812
METEOR	0.2243	0.1309
BLEU	0.2567	0.1548
QAFactEval	0.5516	0.3147
SUPERT	0.2945	0.1157
BLANC	0.3304	0.0846
QAEval	0.2102	0.1261
QuestEval	0.5337	0.3816
FactCC	0.3719	0.1675
DAE	0.5743	0.2089
SummaC	0.3621	0.2953

Table 3: Instance-level Pearson correlation of various metrics with factual consistency ratings on Task 1 and Task 2.

Furthermore, we observe that automatic metrics report notably lower results for GPT-3 summaries compared to fine-tuned models in both of our tasks. However, in our manual evaluation, the performance of GPT-3 is actually very high, surpassing the other three models. This indicates that GPT-3 excels in factual consistency. Nonetheless, the results of automatic indicators, whether they measure overall quality or factuality evaluation, are entirely opposite to those of manual evaluation. This leads us to believe that automatic metrics may not be reliable for comparing the quality of zero-shot summaries. The evaluation method for zero-shot summaries should probably differ from that of fine-tuned summaries, as it may be more subjective. We plan to investigate this issue further in future research.

4 Methodology

We have observed that DAE, QuestEval, and QAFactEval exhibit high correlations with factual consistency across the two datasets. Therefore, our goal is to incorporate DAE and QAFactEval as reinforcement learning objectives to enhance the performance of the base model in text summarization. To achieve this, we can augment the base model's loss function with QAFactEval using reinforcement learning, specifically the policy gradient approach.

Consider the base model's loss function $L(\theta)$, where θ represents the model's parameters. The reinforcement learning objective, e.g. using QAFactEval, is to maximize the reward function $J(\theta)$ for each trajectory τ under the policy θ. The reward function can be defined as:

$$J(\theta) = \mathbb{E}_{\tau \sim p\theta(\tau)}[R(\tau)] \tag{1}$$

Model	Human	QAFactEval	DAE
CONFIT	6.3	3.8350	0.7585
+RL QuesEval	7.1	3.8365	0.7602
+RL QAFactEval	7.3	3.8493	0.7634
+RL DAE	7.3	3.8375	0.7748

Table 4: Evaluation results on Task 1 Conclusion Generation.

Model	Human	QAFactEval	DAE
CONFIT	6.3	3.6820	0.7410
+RL QuesEval	7.0	3.6855	0.7435
+RL QAFactEval	7.3	3.6929	0.7442
+RL DAE	6.3	3.6837	0.7514

Table 5: Evaluation results on Task 2 Review Generation.

Here, $R(\tau)$ denotes the QAFactEval reward for a given trajectory τ, and $p_\theta(\tau)$ represents the probability of the trajectory under the policy θ. To incorporate the QAFactEval reward into the base model's loss function, we can modify the original loss function $L(\theta)$ as follows:

$$L'(\theta) = L(\theta) - \alpha J(\theta) \quad (2)$$

In this new loss function $L'(\theta)$, α is a hyperparameter that balances the contribution of the original loss function and the reinforcement learning objective. By optimizing this new loss function, the base model can generate summaries that better align with the QAFactEval metric.

To approximate the expected reward $J_t(\theta)$ at each time step using a single sample, we can employ the following Monte Carlo estimation:

$$\mathbb{E}_{\tau \sim p\theta(\tau)}[R_t(\tau)] \approx \frac{1}{M} \sum_{i=1}^{M} r_{i,t} \quad (3)$$

Here, $r_{i,t}$ denotes the QAFactEval reward for a single trajectory τ_i at time step t, and M is the number of samples (trajectories) used to approximate the expected reward. Using this approximation, we can update the BART model's loss function at each time step as follows:

$$L'_t(\theta) = L_t(\theta) - \alpha \frac{1}{M} \sum_{i=1}^{M} r_{i,t} \quad (4)$$

In this new loss function $L'_t(\theta)$, α is a hyperparameter that balances the contribution of the original loss function and the reinforcement learning objective for each time step. By optimizing this new loss function at each step, the base model can generate summaries that better align with the QAFactEval metric at every time step.

5 Experiment and Results

5.1 Setting

For the models of PEGASUS, CONFIT, and BART, learning rate was set to 3e-5, a dropout rate of 0.1 was used, a batch size of 32, and GPT-3 we use OpenAI API. And here we have beam search for decoding with beam size 3. We use their original code base for all metrics [2].

5.2 Results

Here we present our experimental results, which demonstrate the effectiveness of incorporating reinforcement learning (RL) objectives into our base models, ConFiT. We use three different metrics, QuesEval, QAFactEval, and DAE, as RL objectives to improve the models' performance in generating factually consistent summaries. Tables 4 and 5 show the results for Task 1 and Task 2, respectively. As the tables show, for both tasks, the models augmented with RL objectives exhibit improved performance across all three metrics. This indicates that incorporating reinforcement learning objectives using QuesEval, QAFactEval, and DAE successfully improves the factual consistency of the generated summaries. It's worth noting that the performance improvement was consistently observed and it did not merely result from some testing of statistical significance. We conducted multiple experiments to ensure the stability and reliability of these performance gains.

In this study, we adopt the same human evaluation setup as previously mentioned, using a 0-10 point scale for rating the generated summaries. This consistent evaluation approach allows us to effectively compare the performance of our models and assess their ability to generate factually consistent summaries in both tasks.

In Table 6, we provide the original text, reference summary, and summaries generated by three different models. It's important to clarify that this example was not cherry-picked. It is representative and indicative of the general trends observed in our data, rather than being an exceptional case chosen to support our argument. From the original

[2]https://github.com/salesforce/QAFactEval, https://github.com/salesforce/factCC, and https://github.com/tingofurro/summac.

	Study Details and Results
Study 1	Antiemetic activity of ondansetron in acute gastroenteritis.,"The mechanism of nausea and vomiting associated with gastroenteritis is unknown. The role of 5-HT3 receptors in emesis associated with gastroenteritis was investigated in paediatric patients. A randomized, double-blind, placebo-controlled, parallel-group study was conducted in three groups of 12 patients each, receiving either a single i.v. dose of ondansetron (0.3 mg/kg), metoclopramide (0.3 mg/kg) or placebo (sterile saline). Food was restricted and oral rehydration was administered for 4 h. During 0-24 h, the number of emetic episodes experienced was significantly greater with placebo (mean = 5) than ondansetron (mean = 2), (P = 0.048), and the proportion of patients experiencing no emesis was significantly greater with ondansetron (58%) than placebo (17%). A numerical difference, in favour of ondansetron, was observed between ondansetron and metoclopramide groups for both of the above parameters. Fewer treatment failures were observed with ondansetron (17%) than placebo (33%) and metoclopramide (42 %). More diarrheal episodes were observed in the groups receiving anti-emetic treatment. Therefore, serotonin, acting through 5HT3 receptors, may play a role in this form of emesis." 364,CD005506,19758398,Clinical trial: oral ondansetron for reducing vomiting secondary to acute gastroenteritis in children–a double-blind randomized study...
Study 2	Oral ondansetron for gastroenteritis in a pediatric emergency department...
Study 3	All subjects also received intravenous NS at 10-20 mL/kg/hr...
Study 4	Dimenhydrinate in children with infectious gastroenteritis...
Study 5	Hospital admission occurred in nine patients (20.5%) receiving placebo...
Study 6	...Ondansetron, a 5-HT3 receptor antagonist marketed as Zofran...
Study 7	...with a clinically important difference of at least 30% when comparing 2 groups of patients presenting...
Reference	Oral ondansetron increased the proportion of patients who had ceased vomiting and reduced the number needing intravenous rehydration and immediate hospital admission. Intravenous ondansetron and metoclopramide reduced the number of episodes of vomiting and hospital admission, and dimenhydrinate as a suppository reduced the duration of vomiting.
CONFIT	A randomized, double-blind, placebo-controlled study investigated the role of 5-HT3 receptors in emesis associated with gastroenteritis in pediatric patients. Results showed that oral ondansetron significantly reduced the number of emetic episodes experienced compared to placebo and increased the proportion of patients experiencing no emesis (58% vs 17%, P = 0.048).
CONFIT with RL	These articles are randomized controlled trials that investigate the efficacy and safety of ondansetron compared to placebo or other antiemetic medications in reducing vomiting episodes and improving oral rehydration therapy. The studies show that ondansetron is effective in reducing vomiting episodes and improving oral rehydration in children with gastroenteritis, and it is well-tolerated with few adverse effects.
GPT3	The studies presented in this table provide evidence that ondansetron, a 5-HT3 receptor antagonist, is effective in reducing vomiting episodes and improving oral rehydration in children with gastroenteritis. The use of ondansetron leads to fewer treatment failures, hospital admissions, and the need for intravenous rehydration. Overall, these results suggest that ondansetron and dimenhydrinate can be useful adjuncts to oral rehydration therapy in managing gastroenteritis in pediatric patients.

Table 6: We provide the original text, reference summary, and summaries generated by three different models, including ConFIT, ConFIT with reinforcement learning (RL) incorporated into its loss function, and GPT-3. The original text comprises various clinical studies; however, due to its length, we have omitted some portions of the input documents.

text marked in blue, we can see that the p-value of 0.048 refers to the comparison between the placebo and ondansetron groups in terms of the number of emetic episodes experienced, rather than the proportion of patients experiencing no emesis. However, in the original output from the ConFIT model, the proportion is incorrectly mixed with the wrong p-value, which we have marked in red. In contrast, the summaries generated by the models with reinforcement learning incorporated into their loss functions do not exhibit any factual errors.

6 Future Work

We find that GPT-3-generated summaries, though scoring poorly on automatic evaluation metrics, were considered superior in quality through manual evaluation. Future studies could focus on this to offer clearer insights and establish a stronger basis for this assertion.

This paper also motivates a potential avenue for improvement, suggesting a shift from traditional automatic metrics towards utilizing advanced models like GPT-4 as evaluators. The premise behind this is to test the correlation between the scores assigned by GPT-4 and human evaluators. If a high correlation exists (which we hypothesize might be the case), it would be feasible to employ GPT-4 scores to guide the tuning process in reinforcement learning. This approach essentially distills the capabilities of GPT-4, leveraging its advanced understanding and evaluative capacity to enhance the performance and efficiency of the summarization models. This proposition could herald a novel direction in the evaluation and tuning of such models, potentially offering more reliable and nuanced performance metrics.

7 Conclusion

In the field of clinical studies summarization, there has been limited research on factual consistency so far. We have demonstrated that, for single-document and multi-document summarization of clinical studies, there are two main issues: 1) the accuracy of automatic metrics for evaluating factual consistency is limited, and 2) existing models have their limitations. We employed human evaluation for assessing factual consistency, and this analysis has been conducted over a larger set of automatic metrics to provide a more comprehensive picture. Furthermore, we demonstrate that further optimizing the model using reinforcement learning (RL) with the metric as a reward can result in significant improvements in factual consistency. Our contributions include a simple yet effective approach for two medical summarization tasks, validation of several automatic evaluation metrics for their correlation with expert-assessed factualness, and the identification of the best-correlating metric to guide generation models toward enhanced summary correctness. This work lays the foundation for the development of more robust clinical trial summarization systems, facilitating the efficient dissemination of medical knowledge to practitioners and researchers.

8 Limitation

While this study provides valuable insights into the performance of summarization models across various domains, there are several limitations that should be noted. Primarily, it is clear from the results that these models exhibit poor performance on RCT compared to other domain datasets. However, it should be noted that this performance gap is likely due, at least in part, to the fact that these models were not trained on medical documents. The complexity of medical terminology and its syntax often requires specific knowledge and understanding that general language models might not possess. Thus, it might not be entirely fair to infer that these summarizers find clinical summarization inherently challenging based on this data alone. In order to address this limitation, it is recommended that future research should involve the same experiments using model checkpoints that have been finetuned on medical text data.

Another notable limitation of the study revolves around the incremental improvements shown by the summarizers for the evaluation metrics used in the reward function. While it is encouraging to observe these slight improvements, it's important to question and validate whether these changes truly signify an enhancement in the model's factuality. It's plausible that the training focused primarily on improving factuality might inadvertently compromise other aspects of the generated text, such as its fluency or ROUGE scores. To gain a more comprehensive understanding, it would be valuable to conduct additional experiments and analyses. This comprehensive evaluation is critical to gain a more nuanced understanding of the trade-offs involved in model training and optimization.

References

Asma Ben Abacha, Yassine Mrabet, Yuhao Zhang, Chaitanya Shivade, Curtis Langlotz, and Dina Demner-Fushman. 2021. Overview of the mediqa 2021 shared task on summarization in the medical domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 74–85.

Subhajit Chaudhury, Sarathkrishna Swaminathan, Chulaka Gunasekara, Maxwell Crouse, Srinivas Ravishankar, Daiki Kimura, Keerthiram Murugesan, Ramón Fernandez Astudillo, Tahira Naseem, Pavan Kapanipathi, and Alexander Gray. 2022. X-FACTOR: A cross-metric evaluation of factual correctness in abstractive summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7100–7110, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Andrea Chaves, Cyrille Kesiku, and Begonya Garcia-Zapirain. 2022. Automatic text summarization of biomedical text data: A systematic review. *Information*, 13(8):393.

Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 2021. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5935–5941, Online. Association for Computational Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

John Concato, Nirav Shah, and Ralph I Horwitz. 2000. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England journal of medicine*, 342(25):1887–1892.

Dina Demner-Fushman and Jimmy Lin. 2007. Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics*, 33(1):63–103.

Franck Dernoncourt and Ji Young Lee. 2017. Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts. *arXiv preprint arXiv:1710.06071*.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. MSˆ2: Multi-document summarization of medical studies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7494–7513, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jay DeYoung, Eric Lehman, Ben Nye, Iain J Marshall, and Byron C Wallace. 2020. Evidence inference 2.0: More data, better models. *arXiv preprint arXiv:2005.04177*.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354, Online. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. *arXiv preprint arXiv:2104.04302*.

Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. Automated lay language summarization of biomedical scientific reviews. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 160–168.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Georgios Katsimpras and Georgios Paliouras. 2022. Predicting intervention approval in clinical trials through multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1947–1957, Dublin, Ireland. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Faisal Ladhak, Esin Durmus, He He, Claire Cardie, and Kathleen McKeown. 2022. Faithful or extractive? on mitigating the faithfulness-abstractiveness trade-off in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1410–1421, Dublin, Ireland. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Irene Li, Alexander Fabbri, Rina Kawamura, Yixin Liu, Xiangru Tang, Jaesung Tae, Chang Shen, Sally Ma, Tomoe Mizutani, and Dragomir Radev. 2021. Surfer100: Generating surveys from web resources on wikipedia-style. *arXiv preprint arXiv:2112.06377*.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022. Readability controllable biomedical document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4667–4680, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for abstractive text summarization. *arXiv preprint arXiv:2303.15621*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Marcia L Meldrum. 2000. A brief history of the randomized controlled trial: From oranges and lemons to the gold standard. *Hematology/oncology clinics of North America*, 14(4):745–760.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Julia Otmakhova, Karin Verspoor, Timothy Baldwin, and Jey Han Lau. 2022a. The patient is more dead than alive: exploring the current state of the multi-document summarisation of the biomedical literature. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5098–5111.

Yulia Otmakhova, Thinh Hung Truong, Timothy Baldwin, Trevor Cohn, Karin Verspoor, and Jey Han Lau. 2022b. LED down the rabbit hole: exploring the potential of global attention for biomedical multi-document summarisation. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 181–187, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization. In *International Conference on Learning Representations*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Sanjana Ramprasad, Denis Jered McInerney, Iain J Marshal, and Byron C Wallace. 2023. Automatically summarizing evidence from clinical trials: A prototype highlighting current challenges. *arXiv preprint arXiv:2303.05392*.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alexander Te-Wei Shieh, Yung-Sung Chuang, Shang-Yu Su, and Yun-Nung Chen. 2019a. Towards understanding of medical randomized controlled trials by conclusion generation. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 108–117, Hong Kong. Association for Computational Linguistics.

Alexander Te-Wei Shieh, Yung-Sung Chuang, Shang-Yu Su, and Yun-Nung Chen. 2019b. Towards understanding of medical randomized controlled trials by conclusion generation. *arXiv preprint arXiv:1910.01462*.

Sajad Sotudeh Gharebagh, Nazli Goharian, and Ross Filice. 2020. Attend to medical ontologies: Content selection for clinical abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1899–1905, Online. Association for Computational Linguistics.

Arvind Krishna Sridhar and Erik Visser. 2022. Improved beam search for hallucination mitigation in abstractive summarization. *arXiv preprint arXiv:2212.02712*.

Xiangru Tang, Alexander Fabbri, Haoran Li, Ziming Mao, Griffin Adams, Borui Wang, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2022a. Investigating crowdsourcing protocols for evaluating the factual consistency of summaries. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5680–5692, Seattle, United States. Association for Computational Linguistics.

Xiangru Tang, Arjun Nair, Borui Wang, Bingyao Wang, Jai Desai, Aaron Wade, Haoran Li, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2022b. CONFIT: Toward faithful dialogue summarization with linguistically-informed contrastive fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5657–5668, Seattle, United States. Association for Computational Linguistics.

Xiangru Tang, Andrew Tran, Jeffrey Tan, and Mark Gerstein. 2023. Gersteinlab at mediqa-chat 2023: Clinical note summarization from doctor-patient conversations through fine-tuning and in-context learning. *arXiv preprint arXiv:2305.05001*.

Rahul Tangsali, Aditya Jagdish Vyawahare, Aditya Vyankatesh Mandke, Onkar Rupesh Litake, and Dipali Dattatray Kadam. 2022. Abstractive approaches to multidocument summarization of medical literature reviews. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 199–203.

Liam van der Poel, Ryan Cotterell, and Clara Meister. 2022. Mutual information alleviates hallucinations in abstractive summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5956–5965, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the blanc: Human-free quality estimation of document summaries. *arXiv preprint arXiv:2002.09836*.

Byron C Wallace, Sayantan Saha, Frank Soboczenski, and Iain J Marshall. 2021. Generating (factual?) narrative summaries of rcts: Experiments with neural multi-document summarization. *AMIA Summits on Translational Science Proceedings*, 2021:605.

David Wan and Mohit Bansal. 2022a. Evaluating and improving factuality in multimodal abstractive summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9632–9648, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Wan and Mohit Bansal. 2022b. FactPEGASUS: Factuality-aware pre-training and fine-tuning for abstractive summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1010–1028, Seattle, United States. Association for Computational Linguistics.

Benyou Wang, Qianqian Xie, Jiahuan Pei, Prayag Tiwari, Zhao Li, et al. 2021. Pre-trained language models in biomedical domain: A systematic survey. *arXiv preprint arXiv:2110.05006*.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, et al. 2020. Cord-19: The covid-19 open research dataset. *ArXiv*.

Tianshu Wang, Faisal Ladhak, Esin Durmus, and He He. 2022. Improving faithfulness by augmenting negative summaries from fake documents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11913–11921, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qianqian Xie, Jennifer Amy Bishop, Prayag Tiwari, and Sophia Ananiadou. 2022. Pre-trained language models with domain knowledge for biomedical extractive summarization. *Knowledge-Based Systems*, 252:109460.

Qianqian Xie and Fei Wang. 2023. Faithful ai in healthcare and medicine. *medRxiv*, pages 2023–04.

Haopeng Zhang, Semih Yavuz, Wojciech Kryscinski, Kazuma Hashimoto, and Yingbo Zhou. 2022a. Improving the faithfulness of abstractive summarization via entity coverage control. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 528–535, Seattle, United States. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Shiyue Zhang, David Wan, and Mohit Bansal. 2022b. Extractive is not faithful: An investigation of broad unfaithfulness problems in extractive summarization. *arXiv preprint arXiv:2209.03549*.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.

Navigating Data Scarcity: Pretraining for Medical Utterance Classification

Do June Min, Verónica Pérez-Rosas, Rada Mihalcea
Department of Electrical Engineering and Computer Science
(dojmin, vrncapr, mihalcea)@umich.edu

Abstract

Pretrained language models leverage self-supervised learning to use large amounts of unlabeled text for learning contextual representations of sequences. However, in the domain of medical conversations, the availability of large, public datasets is limited due to issues of privacy and data management. In this paper, we study the effectiveness of dialog-aware pretraining objectives and multiphase training in using unlabeled data to improve LMs training for medical utterance classification. The objectives of pretraining for dialog awareness involve tasks that take into account the structure of conversations, including features such as turn-taking and the roles of speakers. The multiphase training process uses unannotated data in a sequence that prioritizes similarities and connections between different domains. We empirically evaluate these methods on conversational dialog classification tasks in the medical and counseling domains, and find that multiphase training can help achieve higher performance than standard pretraining or finetuning.

1 Introduction

Current language technologies have enabled the analysis of large amounts of medical conversations to gain insights into important aspects of provider-patient interactions such as patient experience, response to treatment, time allocation for health issues, or quality assurance (Zhou et al., 2021). However, many challenges remain open for this growing field of research on natural language processing (NLP) for healthcare. Among them, there is a need for efficient training frameworks that address the lack of large-scale, publicly available medical dialog datasets.

The recent success of large transformer-based models (Vaswani et al., 2017) in many Natural Language Processing (NLP) tasks related to dialog has motivated their application in the healthcare domain, mainly because of their adaptability capabilities. Work in this area has shown that large pretrained language models (PLMs) are effective for tasks such as assessing and analyzing the quality of counseling conversations or building chatbots for mental health care (Flemotomos et al., 2021).

Previous work on dialog-oriented pretraining approaches has used discourse-aware (intersentential) learning tasks to learn "rich and robust context representations and interactive relationships of dialog utterances" (Zhang and Zhao, 2021). In addition, the composition and order of pretraining corpora have also been studied as crucial factors for downstream task performance (Gururangan et al., 2020), with a multiphase pretraining regimen consisting of general, domain-adaptive, and task-adaptive shown to be effective. While these methods have been found useful for general-purpose dialog, it is still unclear how they perform in medical dialog. Different from other types of conversational dialog, medical conversations are domain-specific, and classification models require not only capturing the discourse relations between dialog utterances and turns but also being aware of speaker dynamics and medical terminology.

In this work, we seek to empirically study the effectiveness of dialog-aware pretraining and multiphase pretraining in medical utterance classification tasks. We focus on pretraining tasks that allow the model to leverage conversational properties, such as turn-shift, speaker role, and intersentential dependencies. We evaluate these methods on datasets that are limited in size, especially compared to large corpora that are typically used to pretrain language models. Thus, our goal is to confirm if the pretraining approaches lead to improvements over just finetuning with a small dataset.

The contributions of this work are threefold. (1) We design and implement simple dialog-aware pretraining tasks for medical conversations. (2) we evaluate dialog-aware pretraining and multiphase pretraining and show that while the former does

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 59–68

not fare better than dialog-agnostic approaches, the latter can effectively leverage unannotated corpora of varying task relevance. (3) we draw lessons for practitioners from our experiments.

2 Related Work

Pretrained Language Models. Previous works have explored pretraining objectives and strategies that use large amounts of unannotated data to optimize large neural networks (Kim et al., 2020). These methods can be broadly categorized into autoregressive models (e.g., GPT (Radford and Narasimhan, 2018)) and autoencoding models (BERT) (Devlin et al., 2019). The strategies used in this work belong to the autoencoding class since they rely on reconstructing the original input from its corrupted version. In particular, we explore domain-adaptive pretraining (DAPT) and task-adaptive pretraining (TAPT), which have been shown effective while incorporating both, domain-relevant and task-relevant, unlabeled data (Gururangan et al., 2020). In addition, given recent advances in large language models (LLMs), Lehman et al. (2023) show that smaller language models carefully trained on clinical data outperform much larger models trained on general domain data, motivating our work on leveraging domain-specific pretraining on health-related conversational data.

Pretraining Methods for Conversations. Recent work in this area has used intersentential learning objectives to infer properties associated with the relationship of sentences in the input. For instance, Mehri et al. (2019) used masked-utterance retrieval to guess the replaced (masked) utterance based on the context utterance, and Zhang and Zhao (2021) explored intersentential coherence through utterance order restoration and contrastive loss. Other approaches have explicitly incorporated the dialog structure and information as part of the pretraining task. For example, MPC-BERT (Gu et al., 2021a) focused on multi-party conversations and used dialog-unique information such as utterance speakers and receivers.

Conversation Utterance Classification. Within medical utterance classification, work has been done to either categorize or forecast utterances that describe behaviors from conversation participants (Cao et al., 2019). For categorization, Pérez-Rosas et al. compiled a dataset of motivational interviewing (MI) conversations, where each counselor utterance is annotated with a predefined MI behavior code for the counseling strategy employed in the utterance (Pérez-Rosas et al., 2016). In this work, we adopt the framework joint sentence representation (JSR) (Cohan et al., 2019). That is, instead of encoding a single sentence at a time and using its embedding for classification, we jointly encode multiple sentences in a window with specified context size and use their final embeddings to classify multiple sentences at a time.

3 Pretraining Approaches

We explore dialog-aware pretraining and multi-phase training to leverage unlabeled data in medical conversations.

3.1 Discourse Structure Objectives

We focus on two pretraining tasks that incorporate information about turn shift behavior and the speakers' role. We believe that these play a more significant role in clinical dialog than in everyday conversations since the expected role and behavior of each participant are fixed and understood by the speakers. Thus, we hypothesize that models with improved awareness of these structures will lead to higher performance in medical dialog downstream tasks.

Turn-shift Prediction. An important aspect of conversation dynamics is turn-shifting behavior i.e., points in the conversation where a speaker starts a new turn, which can provide information on power balance and rapport between participants. In clinical conversations, turn-taking behavior helps to move the conversation forward and facilitates patient-provider communication. We incorporate turn-taking information into contextual embeddings by designing a pretraining task in which a model is evaluated and trained on its ability to correctly identify the start of a new turn. We define a turn as a contiguous span of utterances spoken by one speaker. In our model, each utterance is separated by `[SEP]` tokens and we predict whether the current `[SEP]` is the start of a new turn. We use a simple feed-forward neural network and a final sigmoid activation for binary prediction on `[SEP]` tokens with binary cross entropy loss for scoring. Note that since the model receives speaker role information through speaker embedding, it is possible that information leakage may make this task trivial. However, during our experiments, we observe that the model does not converge quickly,

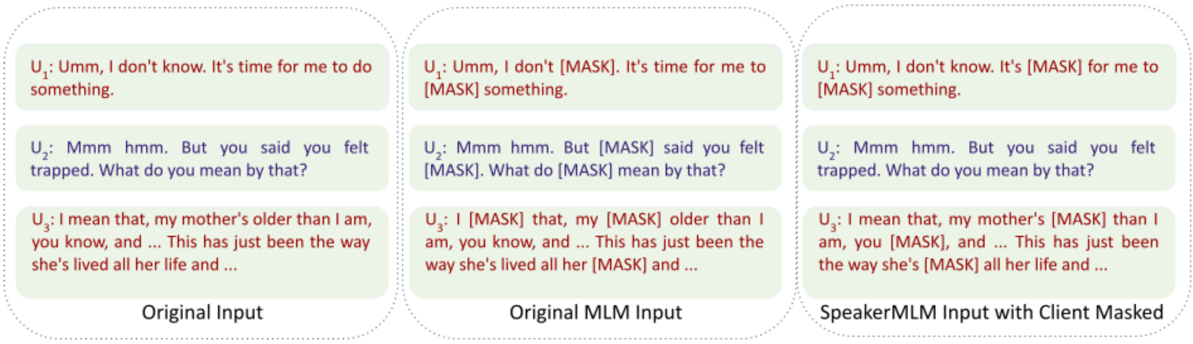


Figure 1: An example dialog snippet and the generated pretraining examples for masked language modeling (MLM) and SpeakerMLM pretraining objectives. Counselor utterances are in red, while client utterances are in purple. Note that here actual token representation of the input sequence is skipped for legibility.

thus, suggesting it is not exclusively relying on the speaker embeddings while predicting a turn-shift.

SpeakerMLM. Although dialog-agnostic MLM has been shown to be effective for many domains and tasks (Devlin et al., 2019), it often fails to leverage potentially helpful domain information, since each non-special token at any position has an equal chance of being masked. In order to augment an MLM with dialog-specific information, we design a masking strategy where masked tokens are selected based on their dialog-specific information. We hypothesize that by forcing the transformer model to infer one speaker's masked tokens from the other speaker's unmasked tokens, the model will learn intersentential and inter-speaker dependencies. More concretely, we start by randomly selecting a speaker with uniform probability and then randomly mask tokens from their utterances with a specific `mask probability`, which is a hyperparameter. The loss used for this task is negative log-likelihood, identical to the original MLM.

3.2 Discourse Coherence Objectives

In addition to the discourse structure pretraining objectives described above, we experiment with Order Recovery and Intruder Detection, two existing pretrained objectives related to discourse coherence that have been found useful in dialog-related tasks (Mehri et al., 2019; Gu et al., 2021b).

Order Recovery. We hypothesize that since the order of utterances is crucial in determining its overall meaning, learning to recover their original permutation may lead to a better contextual representation of a medical conversation. Our modeling approach is similar to Gu et al's (Gu et al., 2021b), but instead of training the model to minimize the KL-divergence between the approximated rank-1 probability and the permutation probability, we minimize the cross entropy between them, following ListNet (Cao et al., 2007). We pass the `[SEP]` embeddings to a feed-forward network with a final sigmoid layer to derive a relevance score for the utterance ranking with respect to the text order.

Intruder Detection. Intruder detection, also known as inconsistency identification (Mehri et al., 2019), seeks to model the coherence of utterances. The goal of this pretraining task is to identify the "intruder" i.e., an utterance that does not belong to the original dialog snippet. We generate intruder detection examples by randomly selecting an utterance i from $[1, 2, \cdots, k]$, and replacing it Utt_i with a negative sample randomly chosen from the pool of all utterances spoken by the same speaker. Note that the negative example cannot be Utt_i itself. As in order recovery, the `[SEP]` embeddings corresponding to each sentence in the dialog snippet are obtained using a feed-forward network that uses cross-entropy loss.

3.3 Multiphase Adaptive Pretraining

An important design aspect of pretraining strategies is the way unlabeled dialog corpora is used during pretraining. While pretraining language models on large, general-topic corpora such as Wikipedia articles have been found useful for general-purpose dialog, pretraining on in-domain or unlabeled target-domain presents an additional opportunity for leveraging unlabeled corpora. This is particularly relevant for the clinical and psychotherapy domains, where large collections of domain data (or annotated data) are often not readily available. To address this, we experiment with two main strategies while using unlabeled data: domain-adaptive pretraining (DAPT) and task-adaptive pretraining (TAPT) (Gururangan et al., 2020). The first allows the model to access a set of unlabeled texts semantically and stylistically similar to the target domain so additional performance gain can be achieved on the in-domain corpus. The second enables pretraining on unlabeled target corpora (Gururangan et al., 2020).

During our experiments we use BERT (Devlin et al., 2019), a popular choice for the contextual embedding of text sequences, as our backbone architecture. Below, we describe important elements of the architecture that are adapted to conduct our experiments while incorporating the pretraining objectives.

Input Representation. Instead of independently encoding each dialog utterance, we opt for jointly encoding multiple sentences in a local neighborhood. Hence, we directly encode contextual information from surrounding utterances as opposed to a single encoding approach, which requires an extra step to contextualize single representations. Previous works such as (Cohan et al., 2019) have shown that context-augmented representations can improve the model performance on sequential sentence classification (SSC) tasks for conversational domains. The main advantage of this strategy is that a separate contextualization step is not necessary and the resulting representation can be directly fed into a feed-forward network for classification. Thus, it allows classifying multiple sentences at a time. Specifically, we set a context window of fixed size k and concatenate all the utterances in the window, separated by special tokens. Thus, a sample sequence given k consecutive utterances spoken in a dialog snippet will be represented as shown below:

$$\texttt{[CLS]}\ \mathrm{Utt}_1\ \texttt{[SEP]}\ \mathrm{Utt}_2\ \texttt{[SEP]}\ \cdots \mathrm{Utt}_k\ \texttt{[SEP]}$$

where Utt_i refers to the sequence of tokens for utterance i, and [CLS] and [SEP], respectively denote the special tokens for classification and utterance separation.

Speaker Embeddings. While the original BERT uses segment embedding to distinguish multiple sentences, we choose to use speaker embeddings to focus on dyadic conversations only. The speaker embedding layer maps two speakers to a learnable embedding in the hidden dimension space. This approach is similar to (Gu et al., 2020), but instead of directly modeling end of turns with an additional token, we provide the relevant turn information through the speaker embedding.

Adaptation to Different Pretraining Objectives. To adapt the model to different pretraining objectives, we add a task-specific feed-forward layer and an activation layer if necessary. Likewise, during training on downstream tasks, a feed-forward layer is added after the last layer of the encoding model so that the model can be fined-tuned to classify an utterance label from [CLS] embedding for forecasting, and from [SEP] embeddings for jointly categorizing.

4 Datasets

We evaluate our pretraining approach using two datasets portraying clinical interactions between patients and their care-providers (Min et al., 2020; Pérez-Rosas et al., 2016) and also a general-purpose chit-chat dataset.

GRADE Clinical Conversations. This dataset consists of clinical conversations from a large diabetes study (Nathan et al., 2013). The conversations are conducted in English and portray interactions between a diabetic patient and a care provider during the patient's regular check-up for diabetes management. The dataset is annotated at the utterance-level with eight diabetes-related codes, covering a range of medical and diabetic-specific topics including "Not Applicable" code for any other topic (Min et al., 2020).

Motivational Interviewing Dataset. This dataset consists of 277 motivational interviewing (MI) counseling sessions also in English, compiled by (Pérez-Rosas et al., 2016). It includes utterance-level annotations for ten behavioral codes from the Motivational Interviewing Treatment Integrity (MITI) coding scheme, the current standard for evaluating MI counseling fidelity and quality (Moyers et al., 2016). The behavior codes indicate the counseling strategy employed in each counselor utterance. In addition to the 10 behavioral codes, we include two generic codes, one for all client utterances and another for counselor utterances with no counseling strategy assigned, resulting in a total of 12 codes.

DailyDialog. In addition to medical corpora, we use the DailyDialog dataset, a corpus of human-written dialogues covering general-domain and

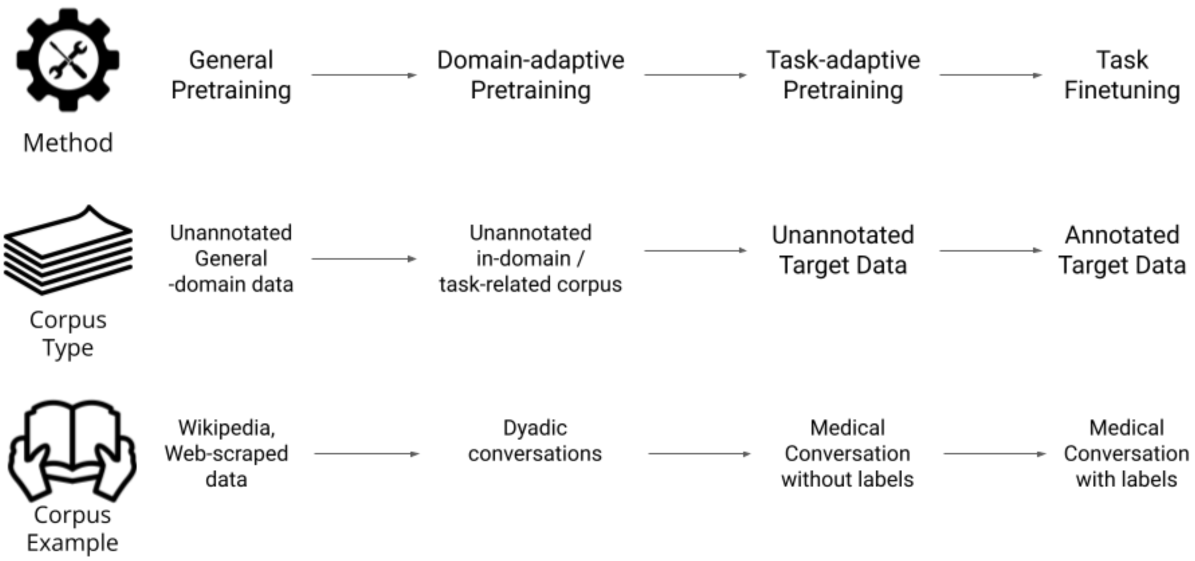


Figure 2: An overview of our multiphase pretraining framework for medical conversations. We start with an pretrained model previously trained on large amounts of general-domain data such as Wikipedia or BookCorpus (Zhu et al., 2015). We then apply domain-adaptive pretraining (DAPT), followed by task-adaptive pretraining (TAPT). Finally, the model is finetuned on the training set of the target task.

chitchat topics such as school life or personal finances (Li et al., 2017). We use DailyDialog as an outer-domain corpus related to the target task in terms of data format or genre (dialogs), but domain distant i.e., general daily life vs medical conversations.

Since all datasets are originally in long full-session length form, we use a sliding window of five conversational turns to segment the sessions into smaller units. Table 1 shows overall statistics for the three datasets.

	GRADE	MI	DailyDialog
# conversations	56	277	13118
# examples	14195	48157	49486
# Finetune Labels	8	12	NA

Table 1: Dataset statistics of GRADE, MI, and DailyDialog datasets

5 Experiments

We focus on two utterance classification tasks: categorization and forecasting, formulated as shown in Table 2. For the categorization task, we seek to label utterances in a medical conversation, where the set of labels is a predefined set of speaker behaviors or conversation topics. In the forecasting task, we use the same set of labels but seek to forecast the label for an upcoming utterance based on previous utterances. Our choice of these tasks is motivated by the hypothesis that if our dialog-aware pretraining objective leads to models with a better contextual representation of neighboring utterances, that improvement will translate into a higher performance boost for forecasting tasks than in categorizing.

Utterance Classification Tasks
Categorizing
Input: $u_1, u_2, \cdots, u_n \xrightarrow{\text{predict}}$ Target: $c_1, c_2 \cdots, c_n$
Forecasting
Input: $u_1, u_2, \cdots, u_n \xrightarrow{\text{predict}}$ Target: c_{n+1}

Table 2: Comparison of categorizing and forecasting tasks. u_i denotes an utterance i in an example. c_i denotes the target label for u_i.

5.1 Experimental Setup

The experiments are run on a GeForce RTX 2080 Ti. For MLM and SpeakerMLM, we set the masking probability of each token to 0.15 and 0.30 respectively. We set the masking probability for SpeakerMLM as 0.30, since by selecting only one of the two speakers to mask we are asymptotically masking only half of the total utterances, in contrast to MLM. Our evaluations are conducted with

5-fold cross-validation. For training in both pretraining and fine-tuning, we use a sliding window with stride $= 1$ to maximize the model learning opportunities. During test time, we partition the dataset with a window of the same size.

We chose BERT as our base architecture since pretrained parameters fine-tuned on large natural language corpora are readily available, and also because due to its design the additional context input could easily be supplied through the use of separate token type ids. We used the `bert-base-uncased` model implemented in (Wolf et al., 2020) with a learning rate of 2e-5. The input to the model is a sequence of token-level embeddings of each utterance in the conversation and the predicted label is assigned using a multilayer perceptron.

5.2 Fine-tuning Strategy

While DailyDialog is a conversational dataset, its topic, and semantic content is generally not domain specific like GRADE or MI. Thus, we use this dataset as a precursor pretraing corpus to DAPT and TAPT, hoping to maximize the gains from domain-adaptive pretraining by creating a conceptually "closer" stepping stone from the Wikipedia-trained weights of `bert-base-uncased`.

We process each conversation in the datasets to transform the long sequence into a set of smaller dialog snippets of size k measured in the number of utterances.

5.3 Utterance Classification Experiment

We evaluate the pretrained models on both categorizing and forecasting tasks. We experiment with two baselines: (1) the same transformer model described above with no pretraining (No) and (2) standard MLM pretraining (MLM). We conduct a set of experiments where we fix the pretraining method but vary (1) the composition of pretraining corpus (2) the pretraining strategy, and whether to use multi-phase adaptive pretraining or not. Results are shown in Table 3 and Table 4. In the tables, "Small" indicates that models are trained using only the target corpus, whereas under the "Mixed" setting models are trained on the shuffled and combined corpus of all the three corpora, and intended as a control against the "Multi" setup which uses the multi-phase pretraining on all corpora. For the "Multi" evaluation, the pretraining order is DailyDialog $\rightarrow$ Non-target Corpus $\rightarrow$ Target Corpus.

From these results, we note that overall, multi-phase adaptive training ("Multi") achieves the highest performances, but does not always lead to performance gains. For instance, as seen in the performance degradation in MI categorizing tasks, the domain and target adaptive strategies actually lead to lower performance, especially when the same setup with "Mixed" pretraining strategy resulted in performance degradation or stagnation. We hypothesize that multi-phase pretraining amplifies the effect of pretraining objectives hence leading to a higher performance boost than when using the "Mixed" setting only. This indicates that choosing the right pretraining schedule/strategy is important but doesn't provide the full recipe for successful pretraining.

We also observe that across datasets and tasks, MLM and SpeakerMLM perform consistently higher than other pretraining methods. However, we see a clear difference in task performance for the GRADE and MI datasets. Particularly, discourse-aware objectives (Order Recovery, Intruder Detection) in MI tasks achieve comparable or higher performances in both categorizing and forecasting. In one notable instance, intruder detection achieves the highest score with multi-phase training.

This trend is in line with existing discourse-aware pretraining work (Mehri et al., 2019; Santra et al., 2021) suggesting that pretraining tasks that require the model to infer how local utterances are related to each other can benefit from explicit intersentential pretraining approaches. In our case, MI tasks focus on counselor strategy and verbal behavior rather than the semantic content of the utterance, whereas the GRADE task is about the utterance topic. In the former, correct classification relies not only on the content of the target utterance but also on the surrounding context.

Moreover, another comparison can be made along the categorizing vs forecasting axis for both datasets. While MLM outperforms SpeakerMLM in categorizing tasks, SpeakerMLM performs best in forecasting tasks. We believe that forecasting represents a more intersentential task since the model has no access to the target utterance and has to rely only on the context utterances for classification. This may explain why SpeakerMLM outperforms MLM in forecasting despite employing a similar principle and using a similar amount of compute (15% of tokens).

	Categorizing Acc			Forecasting Acc		
No Pretraining	65.57			60.03		
Objective / Corpus	Small	Mixed	Multi	Small	Mixed	Multi
MLM	65.67	66.14	**66.72**	61.03	61.76	62.09
Turn-shift Prediction	52.83	51.45	51.46	51.53	51.63	51.63
Order Recovery	65.00	55.53	55.21	57.79	53.83	52.76
Intruder Detection	62.18	62.67	61.56	53.26	58.73	57.40
SpeakerMLM	64.76	66.11	66.41	61.09	61.03	**62.50**

Table 3: Performance of pretrained BERT contextual embeddings on the GRADE topic classification task

	Categorizing Acc			Forecasting Acc		
No Pretraining	76.87			69.58		
Objective / Method	Small	Mixed	Multi	Small	Mixed	Multi
MLM	77.17	77.25	77.07	69.81	69.61	69.91
Turn-shift Prediction	76.85	73.92	73.92	69.57	69.60	69.63
Order Recovery	77.12	76.94	76.78	69.80	70.11	**70.22**
Intruder Detection	57.07	50.54	**83.31**	69.59	69.57	70.18
SpeakerMLM	77.18	76.82	76.39	69.50	69.47	70.18

Table 4: Performance of pretrained BERT contextual embeddings on the MI behavioral coding task

5.4 Evaluation Under Low Resource Settings

We also conduct experiments to evaluate the proposed pretraining strategies on downstream task performance in low-resource settings, where available supervised learning data is limited in quantity.

We measure the performance of pretrained models when using incremental amounts of finetuning data: 0.01, 0.1, and 0.5. We limit our experiments to the multi-phase adaptive setting ("Multi") and No Pretraining, MLM, and SpeakerMLM objectives.

Results are shown in Table 5 and Table 6. Overall, results indicate a similar trend to experiments conducted with all available data, with SpeakerMLM showing a better performance in the forecasting task. In addition, we find that the No Pretraining model has similar or better performances as MLM methods in lower resource settings (0.01, 0.1, 0.25), which is in contrast to the full-resource setting result. This suggests that domain-specific pretraining does not always lead to robust performances under lower resource settings, especially when finetuning is required to improve the model performance.

6 Is Our Finding Still Relevant in the Era of LLMs?

Recently, there have been significant advances in large language models (LLMs), which contain more than several hundred billion parameters and exhibit state-of-the-art performances on several natural language benchmarks, or even academic and professional tasks (Chowdhery et al., 2022; OpenAI, 2023). Given this development, the relevance and need for NLP research that focuses on and optimizes smaller-scale models, such as this work, may be questioned.

We believe that research on the optimization and development of smaller-scale models will still play an important role in NLP research and application. First, the ownership and control of a language model can be important, especially to organizations that handle and process medical data, which is a focus of this work. Such organizations curate patient data with sensitive information, and as such, feeding the data to LLMs, often only available through APIs, may cause legal, ethical, or security violations. Moreover, because LLMs are often trained with large amounts of labeled data, they often underperform task-specific finetuned models that use fewer parameters (Lehman et al., 2023). Thus, leveraging small to mid-size datasets for finetuning remains a viable option.

7 Conclusion & Lessons Learned

In this work, we studied the performance of pretraining strategies on utterance classification in the medical field, a domain that often suffers from a lack of large, publicly available datasets. We evaluated existing and novel dialog-aware and inter-

	Classification Acc				Forecasting Acc			
Objective / Data Fraction	0.01	0.1	0.25	0.5	0.01	0.1	0.25	0.5
No Pretraining	12.04	51.46	51.44	55.23	7.40	51.66	51.56	51.60
MLM	4.60	51.46	52.04	57.99	8.30	51.63	51.56	52.53
SpeakerMLM	12.30	51.46	52.54	**59.22**	8.70	51.63	51.56	**52.56**

Table 5: GRADE Low-resource evaluation of Multi-setting pretrained models using incremental amounts of finetuning data.

	Classification Acc				Forecasting Acc			
Objective / Data Fraction	0.01	0.1	0.25	0.5	0.01	0.1	0.25	0.5
No Pretraining	54.86	70.96	70.99	72.98	36.78	53.44	69.38	69.54
MLM	49.76	70.96	70.99	**74.74**	31.66	49.53	51.44	69.17
SpeakerMLM	53.10	70.82	70.99	71.20	33.60	49.53	51.98	**69.56**

Table 6: MI Low-resource evaluation of Multi-setting pretrained models using incremental amounts of finetuning data.

sentential pretraining objectives, and we showed that multi-phase adaptive training can effectively harness unlabeled data based on task similarity and relevance. We derive several lessons and further directions from our findings:

Pretraining is often beneficial but also has the potential to amplify the negative effects of ill-matched pretraining tasks. Our experimental results confirmed previous works' findings that pretraining strategies to incorporate unlabeled data can be helpful in classification tasks (Devlin et al., 2019; Gururangan et al., 2020). However, we found that using dialog-aware pretraining tasks in medical utterance classification can also lead to poor performance when they are not compatible with the target task.

Pretraining with unlabeled non-target corpora is a useful strategy when the availability of fine-tuning data is limited. Our experimental results showed that pretraining with similar non-target data can boost performance. This is in line with previous findings by Gururangan et al. (2020), showing that after general-domain pretraining on large corpora, additional, domain or target-related training can lead to performance gains. Moreover, we recommend using a multiphase pretraining schedule that uses pretraining corpus on increasing order of task similarity. However, one caveat we observed during our low-resource experiments is that in settings where the amount of fine-tuning data is below a certain threshold, the advantage of pretraining can be limited.

Pretraining with domain-specific data does not result in better performance when compared to domain-agnostic objectives. We implemented several dialog-aware objectives and adapted MLM so that the masking procedure can utilize speaker information assigned to each utterance in conversation. However, we did not see conclusive evidence that these task-specific adaptations led to a significant improvement. Furthermore, in some cases, pretraining with dialog-aware objectives led to a degradation in performance.

Limitations

Our work does not cover the full range of domain-agnostic pretraining objectives, including denoising objectives such as ELECTRA (Clark et al. (2020)), or contrastive objectives, such as SimCSE (Gao et al., 2021; Rethmeier and Augenstein, 2023). This paper focused on comparing the masked language modeling (MLM) objective with specially designed dialog-aware objectives. It is our expectation that, given the empirical findings of this project, task-agnostic general objectives like ELECTRA, or SimCSE, will also outperform dialog-aware methods. In addition, due to the lack of task-related datasets, the set of corpora used during our experiments is limited.

Ethics Statement

The data used for this study was cleaned and anonymized to remove any personal and sensitive information before conducting the reported experiments.

References

Jie Cao, Michael Tanana, Zac Imel, Eric Poitras, David Atkins, and Vivek Srikumar. 2019. Observing dialogue in therapy: Categorizing and forecasting behavioral codes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5599–5611, Florence, Italy. Association for Computational Linguistics.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: From pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, page 129–136, New York, NY, USA. Association for Computing Machinery.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Pre-training transformers as energy-based cloze models. In *EMNLP*.

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Dan Weld. 2019. Pretrained language models for sequential sentence classification. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nikolaos Flemotomos, Victor R. Martinez, Zhuohao Chen, Torrey A. Creed, David C. Atkins, and Shrikanth Narayanan. 2021. Automated quality assessment of cognitive behavioral therapy sessions through highly contextualized language representations. *CoRR*, abs/2102.11573.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Jia-Chen Gu, Tianda Li, Quan Liu, Zhen-Hua Ling, Zhiming Su, Si Wei, and Xiaodan Zhu. 2020. Speaker-aware bert for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 29th ACM International Conference on Information amp; Knowledge Management*, page 2041–2044, New York, NY, USA. Association for Computing Machinery.

Jia-Chen Gu, Chongyang Tao, Zhen-Hua Ling, Can Xu, Xiubo Geng, and Daxin Jiang. 2021a. Mpc-bert: A pre-trained language model for multi-party conversation understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, Online. Association for Computational Linguistics.

Xiaodong Gu, Kang Min Yoo, and Jung-Woo Ha. 2021b. Dialogbert: Discourse-aware response generation via learning to recover and rank utterances. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12911–12919.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Kang-Min Kim, Bumsu Hyeon, Yeachan Kim, Jun-Hyung Park, and SangKeun Lee. 2020. Multi-pretraining for large-scale text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2041–2050, Online. Association for Computational Linguistics.

Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J. Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. 2023. Do we still need clinical language models?

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Shikib Mehri, Evgeniia Razumovskaia, Tiancheng Zhao, and Maxine Eskenazi. 2019. Pretraining methods for dialog context representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3836–3845, Florence, Italy. Association for Computational Linguistics.

Do June Min, Veronica Perez-Rosas, Shihchen Kuo, William H. Herman, and Rada Mihalcea. 2020. Upstage: Unsupervised context augmentation for utterance classification in patient-provider communication. In *Proceedings of the 5th Machine Learning for Healthcare Conference*, pages 895–912.

Theresa Moyers, Lauren Rowell, Jennifer Manuel, Denise Ernst, and Jon Houck. 2016. The motivational interviewing treatment integrity code (miti 4): Rationale, preliminary reliability and validity. *Journal of Substance Abuse Treatment*, 65.

David M. Nathan, John B. Buse, Steven E. Kahn, Heidi Krause-Steinrauf, Mary E. Larkin, Myrlene Staten, Deborah Wexler, John M. Lachin, and the GRADE research group. 2013. Rationale and design of the glycemia reduction approaches in diabetes: A comparative effectiveness study (GRADE). *Diabetes Care*, 36(8):2254–2261.

OpenAI. 2023. Gpt-4 technical report.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2016. Building a motivational interviewing dataset. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 42–51, San Diego, CA, USA. Association for Computational Linguistics.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pretraining.

Nils Rethmeier and Isabelle Augenstein. 2023. A primer on contrastive pretraining in language processing: Methods, lessons learned, and perspectives. *ACM Comput. Surv.*, 55(10).

Bishal Santra, Sumegh Roychowdhury, Aishik Mandal, Vasu Gurram, Atharva Naik, Manish Gupta, and Pawan Goyal. 2021. Representation learning for conversational data using discourse mutual information maximization. In *North American Chapter of the Association for Computational Linguistics*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhuosheng Zhang and Hai Zhao. 2021. Structural pretraining for dialogue comprehension. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5134–5145, Online. Association for Computational Linguistics.

Binggui Zhou, Guanghua Yang, Zheng Shi, and Shaodan Ma. 2021. Natural language processing for smart healthcare. *CoRR*, abs/2110.15803.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*.

A Appendix

Hyperparameter	Value
Batch Size	32
Optimizer	Adam (`betas`=0.9,0.999)
Learning Rate	2e-5
Weight Decay	0.01
Training Epochs	5
MLM Probability	0.15
Speaker MLM Probability	0.30

Table 7: Training Hyperparameters

Hindi Chatbot for Supporting Maternal and Child Health Related Queries in Rural India

Ritwik Mishra[1] , **Simranjeet Singh**[2] , **Jasmeet Kaur**[1] , **Pushpendra Singh**[1]
and **Rajiv Ratn Shah**[1]
[1] Indraprastha Institute of Information Technology, Delhi
`{ritwikm, jasmeetk, psingh, rajivratn}@iiitd.ac.in`
[2] Netaji Subhas University of Technology, Delhi
`simranjeets.ec18@nsut.ac.in`

Abstract

In developing countries like India, doctors and healthcare professionals working in public health spend significant time answering health queries that are fact-based and repetitive. Therefore, we propose an automated way to answer maternal and child health-related queries. A database of Frequently Asked Questions (FAQs) and their corresponding answers generated by experts is curated from rural health workers and young mothers. We develop a Hindi chatbot that identifies k relevant Question and Answer (QnA) pairs from the database in response to a healthcare query (q) written in Devnagri script or Hindi-English (Hinglish) code-mixed script. The curated database covers 80% of all the queries that a user of our study is likely to ask. We experimented with (i) rule-based methods, (ii) sentence embeddings, and (iii) a paraphrasing classifier, to calculate the q-Q similarity. We observed that paraphrasing classifier gives the best result when trained first on an open-domain text and then on the healthcare domain. Our chatbot uses an ensemble of all three approaches. We observed that if a given q can be answered using the database, then our chatbot can provide at least one relevant QnA pair among its top three suggestions for up to 70% of the queries.

1 Introduction

With inequality in healthcare access across urban and rural parts of India, pregnant and postpartum women in rural areas suffer from low access to healthcare due to limited time with healthcare professionals, language barriers in doctor-patient communication, and societal barriers. In resource-constrained environments, digital support groups are a common platform to seek information about various maternal and child healthcare-related issues (Das and Sarkar, 2014; Kaur et al., 2019; Yadav et al., 2022). The moderators of such support groups are overburdened with enormous queries and find it challenging to provide answers timely. Moreover, group members often ask their health queries in regional languages such as *Hindi* or Hinglish[1]. Given the doctor-to-population ratio of 4.8 doctors per 10000 people in India (Potnuru et al., 2017), the scalability of such healthcare interventions involving doctors becomes challenging (Kaur et al., 2019). Thus, it presents an opportunity to extend informational support to pregnant and postpartum women through a chatbot that can answer their written queries in their local language.

Chatbots are used in various domains, from railways ticket reservations to food delivery[2]. Chatbots have taken up different roles in healthcare, such as psychotherapists, nurses, doctors, and medical consultants (Weizenbaum, 1966; Agrawal et al., 2017; Comendador et al., 2015). Chatbots have the potential to act as the first point of contact for women seeking answers for maternal and child healthcare-related queries, especially in resource-constrained environments (Yadav et al., 2019b). In this work, we explore the potential of a chatbot to provide accurate healthcare information by retrieving the best matching FAQs with their corresponding answers (Mittal et al., 2021).

We developed a chatbot that provides k most relevant FAQs with their corresponding answers (QnA pairs) in response to a healthcare query. The chatbot uses a curated database of QnA pairs in the Hindi language with answers vetted by healthcare professionals. Our chatbot can process user queries written in Latin script (native script for English) and Devanagari script (native script for Hindi). Figure 1 illustrates the overall architecture of the proposed chatbot. For evaluation, we obtained a set of healthcare queries from ASHA

[1] It is a colloquial term to describe a language written using the English script (Latin), but the grammar and vocabulary are borrowed from Hindi. It is also called Hindi-English code-mixed language. For example, 'नमस्ते '(*hello*) is written as '*namaste*'.

[2] https://www.chatbotguide.org/dominospizza-bot

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 69–77
July 14, 2023 ©2023 Association for Computational Linguistics

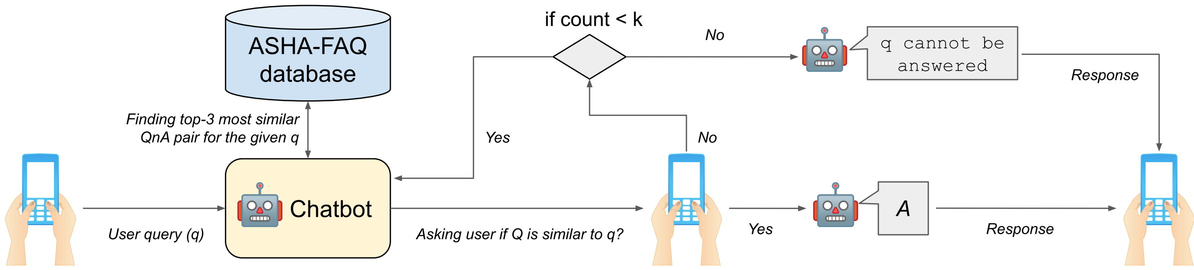


Figure 1: The architecture of the proposed chatbot is shown above. A user query (q) can be in the Devanagari or Latin script. The chatbot fetches top-k most similar Question-Answer (QnA) pairs from the ASHA-FAQ database and shows the user one question (Q) at a time.

workers[3]. In this paper, we discuss different algorithmic approaches to developing chatbots and the efficiency of these approaches in providing relevant QnA pairs. The three primary approaches used in this work are (i) the rule-based method, (ii) sentence embeddings, and (ii) paraphrasing classifiers. An ensemble model of all three primary approaches was found to be performing better than other methods. We release the source code of our chatbot to encourage future research in this direction [4].

2 Related Work

Earlier works on developing chatbots in healthcare using AI started with user query reformulation and using knowledge from search engines (Brill et al., 2002). They were made for the English language, and the same techniques could not be used for Hindi speakers due to the scarcity of resources. Kothari et al. (Kothari et al., 2009) aimed to develop a FAQ retrieval system for the unstructured English language written as a shorthand for SMS by the Indian population. It relied on character-level features to calculate the sentence similarity scores. Initial works on building a QnA system for the Hindi language were restricted to exploiting information from shallow speech features like POS tags (Sahu et al., 2012). In constructing an automatic question-answering system for English-Hindi code-switched language (also known as *Hinglish*), the word-level translation of code-switched queries to English queries was a common practice due to a lack of resources in the Hindi language (Raghavi et al., 2015; Sekine and Grishman, 2003). Such approaches fail to generalize because Hindi-to-English word-level translations are highly dependent on the position of the Hindi word in the sentence (Ray et al., 2018).

Previously cross-lingual word embeddings have been used to solve a healthcare QnA system in low-resource African languages(Daniel et al., 2019). It has been empirically shown that fine-tuned machine learning models using embeddings from pre-trained transformer-based encoders like BERT outperform many other traditional AI models on various tasks(González-Carvajal and Garrido-Merchán, 2020; Hao et al., 2019). Earlier works have shown the efficiency of BERT-based models in measuring sentence similarity for FAQ retrieval tasks(Bhagat et al., 2020; Sakata et al., 2019).

In this paper, we compared the performance of different approaches for measuring sentence similarity between Hindi sentences from the maternal healthcare domain. For a given user query (q), the most similar question (Q) and its corresponding answer (A) are fetched from the ASHA-FAQ database, which is described in the next section.

3 Data Description

We collected data from four prior studies by taking permission from the authors (Yadav et al., 2019a,b, 2021, 2017). The data consists of hundreds of pairs of questions and answers (in audio and text modality), as asked in the real world by community health workers and pregnant and postpartum women regarding maternal and child health issues. Health experts have provided the answers to these questions. The audio data was transcribed and annotated with the help of two healthcare professionals. Both annotators had a bachelor degree in medicine and surgery, a master's in public health, and experience working in maternal and child health. The two annotators manually transcribed each session in the Devanagari script.

[3]They are Accredited Social Health Activists (ASHA) employed by the Ministry of Health and Family Welfare, India. They are frontline health workers connecting the rural population with the state health system.

[4]github.com/ritwikmishra/asha-chatbot

In this work, annotations were performed using an online transliteration tool[5] and Audino (Grover et al., 2020). More than 18 hours of audio in the healthcare domain were transcribed to obtain 1150 question-answer (QnA) pairs. Subsequently, we received 217 maternal health question-answer pairs from Yadav et al. (Yadav et al., 2019b) and added them to our ASHA-FAQ database resulting in a total of 1365 unique questions and 1338 unique answers [6].

Due to the COVID-19 pandemic, AI model field testing was not feasible. Therefore, to evaluate the models on real-time data, a total of 336 new user queries (q) were collected from ASHA workers with the help of a non-governmental organization (NGO) partner. We requested ASHA workers to provide us with queries that they frequently encounter. With the help of public health professionals (with a master's degree in public health), the authors annotated these 336 queries with relevant questions from the ASHA-FAQ database. For each query q, authors identified completely and partially matching QnA pairs from the ASHA-FAQ database. Both types of matching (complete and partial) have been treated as relevant in this work. It has been found that, among 336 queries, 270 user queries had at least one relevant question in the database. Hence, the coverage of the ASHA-FAQ database is 80% in our experiment. The 270 questions, as mentioned above, will be treated as the *hold-out test set* to evaluate the performance of different FAQ-retrieval approaches used by our chatbot.

In order to train a deep-learning model to calculate the sentence similarity score between two Hindi sentences, we scraped Hindi news articles from the Inshorts website[7]. Each data point (d_i) in the scraped dataset (D) consisted of news article text (t_i), its headline (h_i), a summary of the text (t_i^s), and a paraphrased headline (h_i^p). We collected more than $17K$ data points in our dataset. For a negative (or not-paraphrased) headline of h_i, a random headline is chosen from the paraphrased headlines[8].

Our Inshorts dataset contains 35K Hindi sentences, from the news domain, classified into two classes (paraphrased or not-paraphrased), with equal representation of both classes. We are releasing the scraping scripts and hyperlinks to the news articles in the repository mentioned above. To the best of our knowledge, it is the most expansive dataset available for paraphrase detection in the Hindi language. Since the Inshorts dataset is from the open domain (news), we constructed a question paraphrase dataset in the healthcare domain. We manually paraphrased questions from the ASHA-FAQ database and treated them as positive examples of paraphrases. Random sentences were taken as negative examples. The dataset thus created is called the *AshaQs* dataset, and it contains about 1500 healthcare-related question pairs classified into two classes (paraphrased or not-paraphrased) in a balanced manner.

The performance of different FAQ retrieval models is compared using five information retrieval evaluation metrics, namely: Mean Average Precision (mAP), Mean Reciprocal Rank (MRR), Success Rate (SR), normalized Discounted Cumulative Gain (nDCG), and Precision at 3 (P@3) (Sakata et al., 2019). Success Rate is the simplest to understand because it represents the percentage of user queries for which at least one relevant suggestion was given in the top-k suggestions.

4 Methodology

Our work aims to take input as a user query (q) and produces an output as top-k most relevant QnA pairs from the ASHA-FAQ database. Therefore, the given task is modeled as a FAQ retrieval problem. We tried to solve this FAQ retrieval problem through three primary approaches. Results from best-performing approaches are taken to form an ensemble method. All three of our approaches are able to convert Latin script in user input query to Devanagari script. We used the indic-trans library for the transliteration (Bhat et al., 2015).

4.1 Dependency Tree Pruning (DTP)

A dependency parse tree was created for the given sentence, and we extracted all the important keywords by pruning the tree using handcrafted rules. Stanza library is used to extract shallow features like Part-Of-Speech (POS) tags and create the dependency tree for Hindi language (Qi et al., 2020). Tree pruning is done in the following three steps:

I. Advice Removal: In the dependency tree, if

[5] easyhindityping.com

[6] We will be releasing a subset of the ASHA-FAQ database to show the working of our chatbot. Full database will be released in future research work.

[7] inshorts.com/hi/read

[8] We also experimented with curated negatives i.e. negative headline = headline (h_j^p) having the highest keyword overlap with h_i where $i \neq j$. It didn't improve the results.

any children of the root node contain words like सलाह (advice) or इलाज (treatment), or if the root node is an inflection of the Hindi word कर (do) and has a child as चाहिए (should) or क्या (what); then the child with the maximum number of descendants is made the new root, and the original root along-with rest of its children are pruned from the tree.

II. Node removal: After a manual analysis of many dependency trees, we inferred that some nodes with specific dependency relations do not contribute to the underlying meaning of the query. The chosen dependency relations were: *dep*, *displocated*, *discourse*, *expl*, *cc*, *case*, *aux*, *aux:pass*, and *mark*. Hence, the nodes connected to the dependency tree with these relations are removed.

III. Compound merging: In the Hindi language, some actions are expressed through a pair of verbs called compound verbs. For eg: रैप करना (wrap doing) here the first verb is the verb stem, and the second verb is a container for inflections like gender, number, and tense. In the compound merging step, all the compound verbs are reduced to their verb stems only. We used the dependency relation called *compound* to identify the compound verbs.

Since the Hindi language generally follows the subject-object-verb paradigm, post-order traversal was used to extract the words from the pruned dependency tree. It is done to make the extracted sentence more readable. Lemmatization is done on the words to remove the inflections during the traversal.

Using the DTP method, we extracted the keywords for every question (Q_i) in the ASHA-FAQ database. Precision and recall between the user query (q) and Q_i is calculated by comparing the overlap between their keywords. We use $F-measure(q, Q_i)$ as the comparison metric, representing the sentence similarity score between q and the ith question in the database (Q_i).

4.2 Sentence-pair Paraphrasing Classifier (SPC)

The notion is to train a deep learning model to predict a score representing the extent to which the given sentence-pair conveys the same information. The predicted score from the classifier is taken as the sentence similarity score for a given sentence-pair. If two sentences in a given sentence-pair convey identical information, then the trained model is supposed to predict a value closer to one. We fine-tune a pretrained multilingual-transformer-encoder (or simply *encoder* henceforth) responsible for generating d-dimensional embeddings for the given sentence-pair. The embeddings are fed to a Feed-Forward Neural Network (FFNN) with a single output node to predict the sentence-similarity score. Earlier works have shown the superiority of fine-tuned *encoders* for paraphrase detection tasks in Hindi sentences under the IndicGLUE benchmark (Kakwani et al., 2020; Venkatesh et al., 2022). We fine-tuned our SPC on the Inshorts dataset and *AshaQs* dataset using the Huggingface library (Wolf et al., 2020).

4.3 Cosine Similarity (COS)

We used different *encoder*s to obtain a d-dimensional vector representation of q and Q_i, as $E(q)$ and $E(Q_i)$, respectively. We used a pre-trained *encoder* from the SentenceTransformer library (Reimers and Gurevych, 2020) to obtain the vector representation of sentences. The traditional cosine similarity between $E(q)$ and $E(Q_i)$ represents the sentence similarity score between q and Q_i.

Ensemble method ($\mathcal{E}$)

The DTP methodology was selected due to its interpretability, in contrast to the SPC and COS methodologies, which have demonstrated remarkable results in sentence similarity tasks. Additionally, we present an ensemble technique that generates sentence similarity scores by leveraging the outputs of the aforementioned three primary methodologies.

For every input query, each approach above produced a list of the most similar QnA pair from the ASHA-FAQ database, along with their respective sentence similarity score. Top-k QnA pairs with the highest scores are chosen as the final suggestions for each input query. It was observed that for some input queries, one approach performed better than the rest, whereas it performed worse for some. Hence, an ensemble method is developed to construct another top-k suggestion from the final suggestions of different approaches. The ensemble method adds the scores of repeated suggestions, and top-k suggestions having the highest

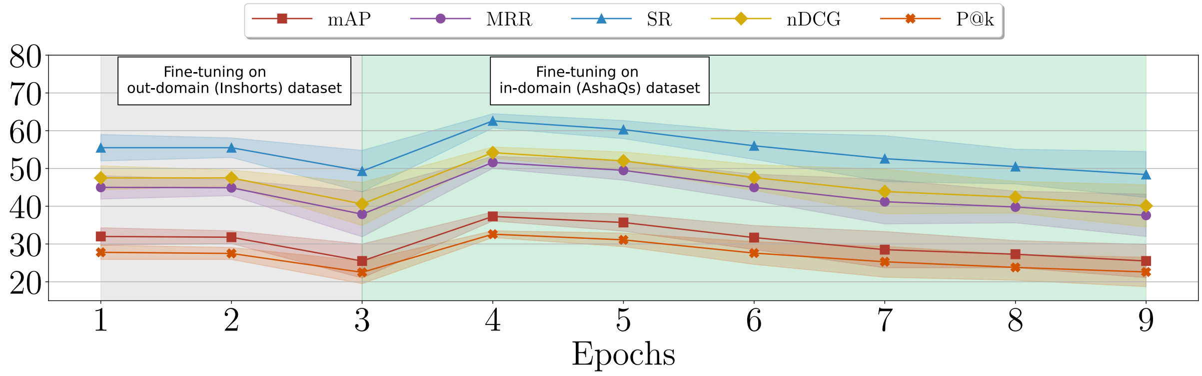


Figure 2: Performance of fine-tuned SPC on the *hold-out test-set* with ten different random seeds. A random seed is responsible for weight initialization in linear layers and the data shuffling between training-testing sets before fine-tuning. The fine-tuned SPC produces top-k QnA suggestions for a given user query (q) where k equals 3. The solid line and the shaded region represent the mean and standard deviation, respectively. The figure depicts the rise in performance of SPC approach when it is fine-tuned on an in-domain data for a single epoch.

	DTP	DTP_{q-e}	SPC	SPC_{+A}	SPC_{q-e}	COS	COS_{q-e}	$\mathcal{E}$
mAP	30.5	35.1	39.4	31.1	39.1	26.5	27.9	**45.3**
MRR	42.6	48.5	54.6	42.2	54.2	38.7	41.0	**61.6**
SR	27.1	59.6	66.2	49.6	64.4	47.7	51.1	**70.3**
nDCG	45.5	51.2	57.1	43.9	56.5	40.8	43.3	**62.5**
P@3	27.1	30.0	34.6	34.6	34.6	22.7	23.9	**34.6**

Table 1: Comparison of all three primary approaches on *hold-out test set* for top-3 suggestions extracted by our chatbot. The ensemble ($\mathcal{E}$) is obtained by taking the best-performing models, highlighted with yellow, from each primary approach. Evidently, the ensemble approach outperforms all the other approaches.

scores are extracted as final suggestions of the ensemble method.

5 Results

It has been observed that, among the top-3 suggestions, DTP gave at least one relevant suggestion only in 27.1% of user queries in the *hold-out test set*. We explored the possible reasons for its failures and found out that the method could not handle the polysemous nature of words. For example, DTP considers शुगर (sugar) and डायबिटीज (diabetes) as entirely different words. However, the two words are interchangeably used in the Indian subcontinent to describe a prevalent disease called *Diabetes mellitus*.

We tried to solve the polysemous word problem by maintaining buckets of such words. Whenever a single word from a bucket is encountered in either q or Q_i, the rest of the words from the bucket are added to the sentence. Expanding the query in such a manner is called *query-expansion (q-e)* in automatic question answering (Ray et al., 2018). It is shown to improve the DTP method by giving relevant suggestions in 59% of the user queries. Table 1 shows the performance boost in DTP due to *q-e* variation. Ablation study highlighting the importance of different pruning strategies in DTP is illustrated in Table 3 of Appendix A.

Multiple *encoders* were used to build the SPC model. It was observed that the *bert-base-multilingual-cased* (*mbert*) *encoder* by Devlin et al. (2018), gave better results than other pre-trained multilingual *encoders*. Moreover, fine-tuning SPC model with three linear layers on top of the *encoder* resulted in the best performance. Since Rogers et al. (Rogers et al., 2020) suggested that early layers of *encoders* contain more syntactic information, we froze the early layers of the *encoder*. We observed more stable results across different random seeds. We first fine-tune the resulting model on the open domain Inshorts dataset and then fine-tune it further on the *AshaQs* dataset in the healthcare domain. We observed that it boosted the performance of SPC on the *hold-out test-set* in the fourth epoch, as shown in Figure 2. Table 1 shows that $q - Q_i$ sentence similarity works better than the $q - Q_iA_i$ similarity, which is aligned with

	$\mathcal{E}_{-COS}$	$\mathcal{E}_{-DTP}$	$\mathcal{E}_{-SPC}$	$\mathcal{E}$
mAP	40.9	40.8	30.3	**45.3**
MRR	56.2	56.5	43.8	**61.6**
SR	66.2	66.2	51.1	**70.3**
nDCG	58.4	58.2	45.5	**62.5**
P@3	34.6	34.6	23.9	**34.6**

Table 2: Results of ablation study on the Ensemble method ($\mathcal{E}$). Th table illustrates that removing any approach ($COS/DTP/SPC$) from the ensemble method results in lower performance.

earlier works (Bhagat et al., 2020; Sakata et al., 2019). Sensitivity of the SPC model with respect to other architectural choices is given in Table 4 of Appendix A.

Calculating sentence similarity score as the cosine distance between the vector representations of two sentences is also effective. We observed that using *paraphrase-multilingual-mpnet-base-v2* as the pretrained *encoder* gave better results than other *encoders* from the SentenceTransformer library. Table 1 shows that using the *q-e* variations on q and Q improved the COS results.

Table 1 shows that the ensemble method $\mathcal{E}$ outperformed all three approaches on the *hold-out test set*. We performed an ablation study to assess the importance of each component of $\mathcal{E}$. The minus sign in the subscript represents the absence of that particular component. For example, if SPC is absent, then it is represented by $\mathcal{E}_{-SPC}$. Table 2 shows that removing any component decreases the performance of $\mathcal{E}$. It was also observed that when the three approaches produced top-5 suggestions, the resulting ensemble method achieved a Success Rate of 73%. Moreover, the chatbot gives a better SR value for user queries with many relevant questions in the ASHA-FAQ database.

The SPC approach majorly dominates the inference time of the ensemble method. It was observed that, with a GPU-enabled server, the ensemble chatbot gives real-time suggestions in 4 seconds and consumes a memory of 2.3 GB on the GPU. However, the chatbot takes a few minutes to generate top-k suggestions without a GPU and consumes a memory of 6.0 GB of RAM.

6 Limitations

In our study, we tested the chatbot on the Hindi database, which humans heavily annotated. Thus, when the database size is enormous, the scalability of the annotation approach is a critical question. Since the questions and answers could be possible in different languages, it will require considerable effort to translate them and, at the same time, preserve their context. In our study, we observed the success ratio of the developed chatbot to be 70% for Hindi queries. However, it is not indicative of its performance in different natural languages.

For a given user query (q), the performance of our best approach for the FAQ-retrieval system is highly dependent on the number of different relevant questions (Q) existing in our ASHA-FAQ database for the given q. Considering the large number of user queries that can be asked in the healthcare field, the small size of our ASHA-FAQ database is a significant reason behind the instances where our method fails to suggest relevant questions (Q) to the user. Moreover, our work does not analyze the quality of answers present in the ASHA-FAQ database. Hence, a user study would be required to analyze the questions' diversity and the answers' quality in our ASHA-FAQ database.

7 Conclusion and Future Work

In this paper, we presented the development of a chatbot to reduce the workload of healthcare professionals for extending informational support regarding maternal and child healthcare concerns in a resource-constrained environment. We followed a FAQ-based model to develop our chatbot using a healthcare database curated in Hindi. Our developed FAQ chatbot can process Hindi user queries written in either the native script of Hindi (Devanagari) or in the native script of English (Latin). We experimented with different FAQ-retrieval methods to extract the most relevant QnA pairs from a FAQ database. We found that the chatbot has the potential to provide relevant QnA pairs for up to 70% queries that our FAQ database can answer. In the future, we plan to evaluate the bot in the wild with healthcare professionals involved.

We plan to evaluate our chatbot with pregnant and postpartum women in a resource-constrained environment to understand the performance of the chatbot in the wild. We also plan to incorporate a healthcare professional to answer questions beyond the chatbot's capacity. The answer obtained from the professional will be further added to the existing QnA database for handling future queries, which would improve the chatbot's success rate over time.

Acknowledgements

Ritwik Mishra would like to express his gratitude towards University Grant Commission (UGC), India, as he is partly supported by the UGC Junior Research Fellowship (JRF) program. Rajiv Ratn Shah acknowledges the partial support provided by the Infosys Center for AI (CAI) and the Center of Design and New Media (CDNM) at IIIT Delhi. We would like to acknowledge the support of the iHub Anubhuti Foundation for supporting this research. We also acknowledge the support of SWACH foundation.

References

Monica Agrawal, Janette Cheng, and Caelin Tran. 2017. What's up, doc? a medical diagnosis bot. *Spoken Language Processing (CS224S), Spring*.

Pranav Bhagat, Sachin Kumar Prajapati, and Aaditeshwar Seth. 2020. Initial lessons from building an ivr-based automated question-answering system. In *Proceedings of the 2020 International Conference on Information and Communication Technologies and Development*, ICTD2020, New York, NY, USA. Association for Computing Machinery.

Irshad Ahmad Bhat, Vandan Mujadia, Aniruddha Tammewar, Riyaz Ahmad Bhat, and Manish Shrivastava. 2015. Iiit-h system submission for fire2014 shared task on transliterated search. In *Proceedings of the Forum for Information Retrieval Evaluation*, FIRE '14, pages 48–53, New York, NY, USA. ACM.

Eric Brill, Susan Dumais, and Michele Banko. 2002. An analysis of the askmsr question-answering system. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 257–264.

Benilda Eleonor V Comendador, Bien Michael B Francisco, Jefferson S Medenilla, and Sharleen Mae. 2015. Pharmabot: a pediatric generic medicine consultant chatbot. *Journal of Automation and Control Engineering*, 3(2).

Jeanne E Daniel, Willie Brink, Ryan Eloff, and Charles Copley. 2019. Towards automating healthcare question answering in a noisy multilingual low-resource setting. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 948–953.

Ashavaree Das and Madhurima Sarkar. 2014. Pregnancy-related health information-seeking behaviors among rural pregnant women in india: validating the wilson model in the indian context. *The Yale journal of biology and medicine*, 87(3):251.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Santiago González-Carvajal and Eduardo C Garrido-Merchán. 2020. Comparing bert against traditional machine learning text classification. *arXiv preprint arXiv:2005.13012*.

Manraj Singh Grover, Pakhi Bamdev, Yaman Kumar, Mika Hama, and Rajiv Ratn Shah. 2020. audino: A modern annotation tool for audio and speech.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2019. Visualizing and understanding the effectiveness of bert. *arXiv preprint arXiv:1908.05620*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961.

Jasmeet Kaur, Asra Sakeen Wani, and Pushpendra Singh. 2019. Engagement of pregnant women and mothers over whatsapp: Challenges and opportunities involved. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, CSCW '19, page 236–240, New York, NY, USA. Association for Computing Machinery.

Govind Kothari, Sumit Negi, Tanveer A Faruquie, Venkatesan T Chakaravarthy, and L Venkata Subramaniam. 2009. Sms based interface for faq retrieval. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 852–860.

Mamta Mittal, Gopi Battineni, Dharmendra Singh, Thakursingh Nagarwal, and Prabhakar Yadav. 2021. Web-based chatbot for frequently asked queries (faq) in hospitals. *Journal of Taibah University Medical Sciences*, 16(5):740–746.

Basant Potnuru et al. 2017. Aggregate availability of doctors in india: 2014–2030. *Indian journal of public health*, 61(3):182.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Khyathi Raghavi, Manoj Chinnakotla, and Manish Shrivastava. 2015. " answer ka type kya he? " learning to classify questions in code-mixed language.

Santosh Kumar Ray, Amir Ahmad, and Khaled Shaalan. 2018. A review of the state of the art in hindi question answering systems. *Intelligent Natural Language Processing: Trends and Applications*, pages 265–292.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Shriya Sahu, Nandkishor Vasnik, and Devshri Roy. 2012. Prashnottar: a hindi question answering system. *International Journal of Computer Science & Information Technology*, 4(2):149.

Wataru Sakata, Tomohide Shibata, Ribeka Tanaka, and Sadao Kurohashi. 2019. Faq retrieval using query-question similarity and bert-based query-answer relevance. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1113–1116.

Satoshi Sekine and Ralph Grishman. 2003. Hindi-english cross-lingual question-answering system. *ACM Transactions on Asian Language Information Processing (TALIP)*, 2(3):181–192.

Gopalakrishnan Venkatesh, Abhik Jana, Steffen Remus, Özge Sevgili, Gopalakrishnan Srinivasaraghavan, and Chris Biemann. 2022. Using distributional thesaurus to enhance transformer-based contextualized representations for low resource languages. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, pages 845–852.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of EMNLP 2020: System Demonstrations*, pages 38–45, Online. ACL.

Deepika Yadav, Anushka Bhandari, and Pushpendra Singh. 2019a. Leap: Scaffolding collaborative learning of community health workers in india. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Deepika Yadav, Kirti Dabas, Prerna Malik, Anushka Bhandari, and Pushpendra Singh. 2022. “should i visit the clinic”: Analyzing whatsapp-mediated online health support for expectant and new mothers in rural india. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI ’22, New York, NY, USA. Association for Computing Machinery.

Deepika Yadav, Prerna Malik, Kirti Dabas, and Pushpendra Singh. 2019b. Feedpal: Understanding opportunities for chatbots in breastfeeding education of women in india. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Deepika Yadav, Prerna Malik, Kirti Dabas, and Pushpendra Singh. 2021. Illustrating the gaps and needs in the training support of community health workers in india. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI ’21, New York, NY, USA. Association for Computing Machinery.

Deepika Yadav, Pushpendra Singh, Kyle Montague, Vijay Kumar, Deepak Sood, Madeline Balaam, Drishti Sharma, Mona Duggal, Tom Bartindale, Delvin Varghese, and Patrick Olivier. 2017. <i>sangoshthi</i>: Empowering community health workers through peer learning in rural india. In *Proceedings of the 26th International Conference on World Wide Web*, WWW ’17, page 499–508, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

A Appendix

	mAP	MRR	SR	nDCG	P@3
DTP_{q-e}	**35.1**	**48.5**	**59.6**	**51.2**	**30.0**
-any pruning	25.5	37.3	45.2	39.1	21.3
-advice removal	30.3	43.1	54.4	46.0	27.0
-node removal	28.4	40.9	53.7	44.2	27.0
-compound merging	31.1	44.4	55.1	47.0	27.0

Table 3: An ablation of different pruning strategies in the DTP method. In absence of any pruning strategy, simple lemmatization, stop-word removal, and token matching is performed.

Fine-tuning data	**Pretrained Encoder**	**Linear Layers**	**Frozen Layers**	**Best Epoch**	**SR spread ($\mu \pm \sigma$)**	**Best SR**
Inshorts 3 epoch AshaQs 6 epoch	mbert-cased	3	embedding, layer 0	4	62.6 ± 1.9	66.2
AshaQs 4 epoch	mbert-cased	3	embedding, layer 0	1	▼ 62.4 ± 3.7 ▲	67.0 ▲
Inshorts 4 epoch	mbert-cased	3	embedding, layer 0	1	▼ 55.5 ± 3.5 ▲	60.0 ▼
Inshorts 1 epoch **AshaQs 1 epoch**	mbert-cased	3	embedding, layer 0	2	▲ 64.3 ± 2.4 ▲	67.8 ▲
Inshorts 2 epoch **AshaQs 1 epoch**	mbert-cased	3	embedding, layer 0	3	▲ 62.9 ± 1.5 ▼	65.9 ▼
Inshorts 4 epoch **AshaQs 1 epoch**	mbert-cased	3	embedding, layer 0	5	▼ 61.0 ± 2.9 ▲	64.1 ▼
Inshorts 3 epoch AshaQs 6 epoch	**xlm-roberta**	3	embedding, layer 0	4	▼ 61.3 ± 2.5 ▲	65.2 ▼
Inshorts 3 epoch AshaQs 1 epoch	**indic-bert**	3	embedding, layer 0	4	▼ 5.9 ± 0.8 ▼	7.0 ▼
Inshorts 3 epoch AshaQs 1 epoch	**mbert** **-uncased**	3	embedding, layer 0	4	▼ 60.0 ± 4.3 ▲	65.9 ▼
Inshorts 3 epoch AshaQs 1 epoch	mbert-cased	**1**	embedding, layer 0	4	▼ 57.8 ± 2.5 ▲	61.9 ▼
Inshorts 3 epoch AshaQs 1 epoch	mbert-cased	**2**	embedding, layer 0	4	▼ 60.6 ± 1.7 ▼	63.0 ▼
Inshorts 3 epoch AshaQs 1 epoch	mbert-cased	**4**	embedding, layer 0	4	▼ 61.3 ± 2.5 ▲	64.8 ▼
Inshorts 3 epoch AshaQs 1 epoch	mbert-cased	3	**embedding**	4	▼ 61.3 ± 3.4 ▲	66.7 ▲
Inshorts 3 epoch AshaQs 1 epoch	mbert-cased	3	**embedding,** **layer 0, 1**	4	▼ 61.3 ± 2.4 ▲	63.0 ▼
Inshorts 3 epoch AshaQs 1 epoch	mbert-cased	3	**embedding,** **layer 0, 1, 2**	4	▼ 61.1 ± 2.2 ▲	63.7 ▼
Inshorts 3 epoch AshaQs 1 epoch	mbert-cased	3	**half bert**	4	▼ 61.9 ± 2.1 ▲	64.4 ▼
Inshorts 3 epoch AshaQs 1 epoch	mbert-cased	3	**nothing**	4	▼ 61.5 ± 2.1 ▲	65.6 ▼

Table 4: Sensitivity of SPC approach due to different architectural choices. Each experiment is run with ten random seeds. For the sake of brevity, we have chosen the Success Ratio (SR) to represent the overall performance since, in our experiments, it acts as an upper bound of all the evaluation metrics. The first row of the table contains the architectural choices of the best SPC approach. Red-colored triangles (▲/▼) represent a drop in performance as compared to the best model. Note: increased standard deviation (σ) indicates more numerical instability, hence worse performance. Since no row contains all green colored triangles, it shows that the configuration of first row is the best configuration.

Multi-Task Training with In-Domain Language Models for Diagnostic Reasoning

Brihat Sharma[1], **Yanjun Gao**[1], **Timothy Miller**[2], **Matthew M. Churpek**[1], **Majid Afshar**[1] and **Dmitriy Dligach**[3]
[1]University of Wisconsin-Madison,
[2]Boston Children's Hospital and Harvard Medical School, [3]Loyola University Chicago
[1]bsharma25@wisc.edu,
[1]{ygao, mchurpek, mafshar} @medicine.wisc.edu,
[2]Timothy.Miller@childrens.harvard.edu, [3]ddligach@luc.edu

Abstract

Generative artificial intelligence (AI) is a promising direction for augmenting clinical diagnostic decision support and reducing diagnostic errors, a leading contributor to medical errors. To further the development of clinical AI systems, the Diagnostic Reasoning Benchmark (DR.BENCH) was introduced as a comprehensive generative AI framework, comprised of six tasks representing key components in clinical reasoning. We present a comparative analysis of in-domain versus out-of-domain language models as well as multi-task versus single task training with a focus on the problem summarization task in DR.BENCH (Gao et al., 2023). We demonstrate that a multi-task, clinically-trained language model outperforms its general domain counterpart by a large margin, establishing a new state-of-the-art performance, with a ROUGE-L score of 28.55. This research underscores the value of domain-specific training for optimizing clinical diagnostic reasoning tasks.

1 Introduction

The electronic health record (EHR) contains daily progress notes authored by healthcare providers to represent the daily changes in care plans for their patients, including an updated list of active diagnoses. The daily progress note is one of the most important note types in the EHR and contains the daily subjective and objective details in the patient's care, which is summarized into an assessment of the overall leading diagnoses with a treatment plan section (Gao et al., 2022b). However, note bloat is a common phenomenon in medical documentation intermixed with billing requirements, non-diagnostic information, and copy and paste from prior notes (Rule et al., 2021). These additional documentation practices contribute to provider burnout and cognitive overload (Gardner et al., 2018). Problem-based charting is important

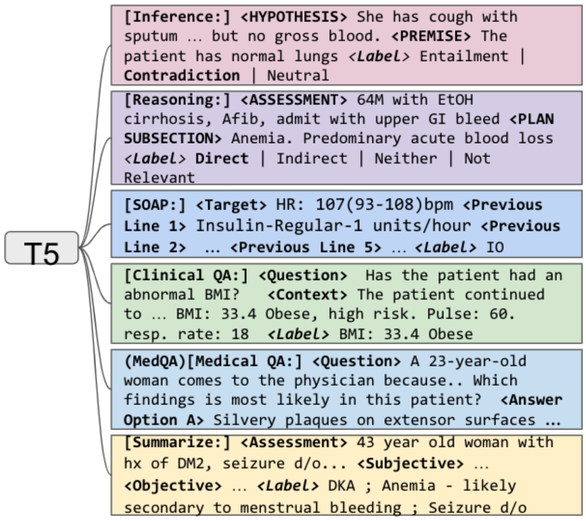


Figure 1: Training T5 with multi-task setup with six tasks from DR.BENCH (Gao et al., 2023)

to improve care throughput and help reduce diagnostic errors (Wright et al., 2012).

The medical reasoning process is complex and incorporates medical knowledge representation with analytical and experiential knowledge (Bowen, 2006). Patel and Groen developed a theory from the AI literature that experts use "forward-reasoning" from data to diagnosis 1986. The recently released benchmark DR.BENCH (Diagnostic Reasoning Benchmark) is intended to assess the ability of AI models to perform such reasoning, with multiple component tasks including diagnostic reasoning with EHR data for experiential knowledge, medical exams for knowledge representation, progress note structure prediction, and problem summarization tasks that included both extractive and abstractive medical diagnoses (Gao et al., 2023).

In this work, we focus primarily on the problem summarization task from the DR.BENCH suite, but with the hypothesis that using all tasks in DR.BENCH would improve the problem summarization task over the problem summarization task being fine-tuned alone. We make use of the T5 fam-

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 78–85
July 14, 2023 ©2023 Association for Computational Linguistics

ily of sequence-to-sequence language models, (Raffel et al., 2020), which are first pretrained on a large unlabeled dataset and then finetuned on specific multiple downstream tasks. The text-to-text approach in our experiment makes it possible to perform multi-task training. Hence, the T5 models were ideal for experimenting with single and multi-task techniques.

Further, we experimented with a recently developed clinically-trained T5 model to quantify the value of in-domain pretraining data (Lehman and Johnson, 2023). We make our software publically available at https://git.doit.wisc.edu/smph-public/dom/uw-icu-data-science-lab-public/drbench.

2 Related Work

In the clinical domain, biomedical text summarization is a growing field. Common approaches to text summarization include feature-based methods (Patel et al., 2019), fine-tuning large language models (Lewis et al., 2020), and domain adaptation with fine-tuning methods (Xie et al., 2023). Researchers developed clinical methods for summarization from progress notes but these methods were restricted to specific diseases such as diabetes and hypertension (Liang et al., 2019). Moreover, these methods for summarization were more extractive than abstractive, using a combination of heuristics rules and deep learning techniques, and did not use large language models (Liang et al., 2019). In another work, an extractive-abstractive approach was used where meaningful sentences were extracted from the clinical notes first; these sentences were then fed into the transformer model for abstractive summarization (Pilault et al., 2020). Unfortunately, the transformer model frequently produced hallucinated outputs, and was not coherent when compared to the ground truth (Pilault et al., 2020). In a similar extractive-abstractive approach, researchers used a pointer generator network to generate a note summary cluster and a language model such as T5 to generate an abstractive summary (Krishna et al., 2021). None of these approaches used multi-task training or focused on clinically trained encoder-decoder since clinical T5 was only recently introduced. Prior work has not addressed the challenge of abstractive reasoning, or they used a two-step process to create abstractions. Recently, researchers used domain adaptive T5 model trained on the biomedical dataset but did not experiment with multi-task settings (Gao et al., 2023).

[1]PubMed is a large open source biomedical and lifescience database consists of 35 million citation and abstract, and PMC (PubMed Central) consists of full articles. MIMIC-III and MIMIC-IV (Medical Information Mart for Intensive Care) are databases consisting of de-identified datasets from Beth Israel Deaconess Medical Center

3 Methods

3.1 Dataset

In our experiments, we used DR.BENCH (Gao et al., 2023), a recently introduced benchmark designed to evaluate diagnostic reasoning capabilities of generative language models. DR.BENCH consists of three categories of tasks (two tasks per category), as shown in Figure 1. From top to the bottom, the categories and six tasks are: **Medical Knowledge Representation:** (1) Medical Natural Language Inference (MedNLI) task that considered sentence pairs with the objective to determine whether the hypothesis sentence could be inferred from the premise sentence (Shivade, 2019) (14,049 sentence pairs total); (2) Assessment and Plan Reasoning (A/P) task whose objective was to label relations between the assessment and treatment plan sections (5,897 samples). **Clinical Evidence Understanding and Integration:** (1) Electronic Medical Records Question Answering (emrQA) whose objective was to answer questions based on discharge summaries (53,199 questions total) (Pampari et al., 2018); (2) Progress Note Section Labeling task whose objective was to labels SOAP sections in progress notes (134,089 samples) (Gao et al., 2022a). **Diagnosis Generation and Summarization:** (1) Medical Board Exam Question Answering (MedQA) task that consisted of medical board exam question-answer pairs (12,725 pairs) (Jin et al., 2021); (2) Problem Summarization (ProbSumm) task whose goal was to produce the list of relevant problems and diagnoses based on the input that consisted of the SOAP sections of progress notes (2,783 samples).

In this work, we focused primarily on the problem summarization task, which was the most difficult but also believed to be the most impactful of the six DR.BENCH tasks for downstream clinical application.

3.2 Experimental Setup

In our experiments, we used six generative language models, all based on the Text-To-Text Trans-

Model	Training Corpus	Initialization	Citation
T5 220M T5 770M	Colossal Clean Crawled Corpus (C4)	Random Random	(Raffel et al., 2020)
SCIFIVE 220M SCIFIVE 770M	C4 + PubMed (abstracts) + PMC	T5 220M T5 770M	(Phan et al., 2021)
CLINICAL-T5 220M CLINICAL-T5 770M	MIMIC-III + MIMIC-IV	T5 220M Random	(Lehman and Johnson, 2023)

Table 1: T5 pretrained models used in the experiments. [1]

fer Transformer (T5) model (Raffel et al., 2020). The text-to-text paradigm utilized by T5 was a natural choice for our stated goal of exploring multi-task learning: transforming T5 into a multi-task learner simply involved prefixing individual task instances with a task-specific prompt after which the model could be trained using the standard cross-entropy loss.

Table 1 provides details about the models. We compared a multi-task scenario in which T5 variants were fine-tuned on all DR.BENCH tasks and a single-task scenario in which T5 was fine-tuned on the problem summarization task only. We trained T5 models as follows:

Single-task training: In single-task training for problem summarization, we used the text of the assessment, subjective and objective sections of the progress notes as input and trained T5 to generate the list of problems and diagnoses.

Multi-task training: In multi-task training, we combined all DR.BENCH tasks into a single dataset and trained T5 to generate task-specific output given task-specific input. Training examples of each task were prefixed with a task-specific prompt. The open-book setting only was used for MedQA. The rest of preprocessing follows (Gao et al., 2023).

To enable comparison with existing work (Gao et al., 2023) we used ROUGE-L score (Lin, 2004) as our evaluation metric. ROUGE-L uses the longest common subsequence statistics to compare model outputs. A resampling technique with 1000 bootstrap samples was used to estimate the 95% confidence intervals (CI) (DiCiccio and Efron, 1996).

Note that the Clinical-T5 model used in our experiment was pretrained on the same data (MIMIC-III) that was annotated by some DR.BENCH tasks (e.g. problem summarization and EmrQA). This setting is known as transductive learning. Transductive learning is a very realistic scenario for the clinical domain where due to privacy issues, language models are likely be pretrained on the data from the same institution as the data to which they would be applied. Obviously, it would also be interesting to investigate the performance of a T5 variant that was trained on a clinical corpus that was different from which the evaluation data were sourced. Unfortunately, this was not possible due to the fact that MIMIC was the only publicly available corpus of clinical notes and it was used for training clinical language models.

The training data consisted of one progress note per unique patient. A separate cohort of unique patients was selected for the test set, ensuring no overlap between the train and test splits. All experiments used Adam optimizer with a learning rate of 1e-5, batch size of 8, beam size of 5, and 100 epochs with early stopping. The learning rate and batch size were picked based on the best hyper-parameters found from the prior work (Gao et al., 2023). All experiments were completed on a single A100 GPU with 40 GB memory. The models were reviewed for error analysis by a critical care physician on the full test set of 86 progress notes and common observations were highlighted with examples in the error analysis.

4 Results and Discussion

The results of our experiments are summarized in Table 2. The full set of results including the confidence intervals is available in the Appendix (Table 4).

Clinical-T5 770M trained in the multi-task setting demonstrated the best performance (28.55) for the Summarization task, establishing a new state-of-the art for this task. The single-task setting for the same T5 variant was a close second (28.28).

T5 variants trained on in-domain data (SciFive and Clinical-T5) performed better than their general domain counterpart T5 models of the same size.

All models, except Clinical-T5 experienced a drop in performance when trained in a multi-task approach. We hypothesize that the models pretrained on non-clinical data were overwhelmed with out-of-domain (i.e. clinical) data when trained in a multi-task way and failed to generalize as a result. Predictably, larger models performed at least as well as the smaller models and outperformed the smaller models in most scenarios.

Admittedly, our work leaves open the question of whether the state-of-the-art performance obtained by Clinical-T5 770M has to do with the fact that it was pretrained on MIMIC notes, which were also annotated in the problem summarization task. At the same time, the performance of other T5 variants, such as SciFive 770M, was close, without it pretraining on MIMIC. This suggests that another T5 variant trained on a corpus of clinical notes that was different from MIMIC would perform at least as well or better depending on the size of the pretraining corpus. It should be noted that the model of this size, 770M parameters, can very likely absorb significantly larger amounts of clinical notes than what was available in MIMIC (Hoffmann et al., 2022). We leave verifying this hypothesis for future work.

Model	Training	Summarization
Gao et al., 2023	Single task	7.60 (5.31 - 9.89)
T5 220M	Single task	26.35 (22.18 - 30.52)
	Multi-task	24.84 (20.28 - 29.40)
T5 770M	Single task	26.90 (22.58 - 31.23)
	Multi-task	23.99 (19.86 - 28.13)
SciFive 220M	Single task	25.31 (21.45 - 29.17)
	Multi-task	24.38 (19.99 - 28.78)
SciFive 770M	Single task	27.31 (23.09 - 31.53)
	Multi-task	25.31 (21.45 - 29.17)
Clinical-T5 220M	Single task	25.35 (21.19 - 29.51)
	Multi-task	26.21 (21.92 - 30.49)
Clinical-T5 770M	Single task	28.28 (24.17 - 32.38)
	Multi-task	**28.55** (24.29 - 32.80)

Table 2: Performance of fine-tuned T5 models on the summarization task. 95% confidence intervals are included. The first row is a baseline representing the best performance on this task to date. Please see the Appendix for the full set of results.

Error Analysis: Although both clinical models produced similar ROUGE-L scores, the model trained in a single-task setting appeared to achieve better abstraction during error analysis. For the example in Table 5, the assessment described sepsis but does not mention the source of the sepsis infection in multi-task Clinical-T5 770M. The data from the subjective and objective sections of the progress note described an abdominal source and lab results were consistent with a clostridium difficile infection. The multi-task prediction was able to generate sepsis but further generated text that the source was unclear. The single task performed better abstraction and generated clostridium difficile as the source for the infection, which was more accurate during expert review. In another diagnosis, the ground truth label was "EtOH Withdrawal" (alcohol withdrawal). The multitask extracted "altered mental status, hypertensive, tachycardia," (symptoms of withdrawal) whereas the single task was able to abstract "DTs EtOH w d," (delirium tremens alcohol withdrawal - a type of severe alcohol withdrawal in critically ill patients). Again, the single task achieved greater accuracy with abstraction from symptoms of alcohol withdrawal presented in the earlier sections of the note.

Resource Utilization: The experiments were conducted on the Google Cloud Platform using one A100 40 GB NVIDIA GPU on a Linux base system. For all experiments, the total training time was approximately 250 hours for both single-task and multi-task approaches. The carbon emission footprint was 35.5 kilograms (kg) of CO_2. However, the total carbon emission was only 4.5 kg of CO_2 for the single-task experiments. (Lacoste et al., 2019)

5 Conclusion

In this work we experiment with the DR.BENCH suite of tasks and established a new state-of-the-art result on the problem list generation task, a task critical for AI-assisted diagnostic reasoning. Our other contribution indicates that multi-task learning does not work well, unless in-domain data was used for pretraining and that included (unlabeled) task data during pretraining (a scenario known as transductive learning) leads to the best performance. Finally, our work provides evidence that generative models benefit from pretraining on in-domain data. In future work, we plan to explore the utility of decoder-only LLMs for clinical diagnostic reasoning.

6 Limitations

The limitation of this work was the use of ROUGE-L as the evaluation metric. Given the many acronyms and synonyms in medical writing, ROUGE-L, based on the longest common sequence, does not capture the many nuances in its score. Researchers have shown concerns for the ROUGE score and have developed metrics for summarization that are more semantically aware of the ground truth (Akter et al., 2022), but their usability is yet to be validated.

Training large language models from scratch uses a considerable amount of carbon footprint. (Patterson et al., 2021) Fine-tuning large language models for downstream tasks is one way to reduce carbon footprint but still needs to be cost-effective. As the AI community progresses in this field, developing a cost-effective and carbon-friendly solution is needed. The NLP field is moving towards prompt-based methods with larger LLMs (Lester et al., 2021), so the next step for this research is to experiment with soft prompting approaches to address low resource settings and leverage prompt tuning in LLMs for the problem summarization task.

7 Ethics Statement

This research utilized a deidentified dataset that does not include any protected health information. This dataset operates in compliance with the PhysioNet Credential Health Data Use Agreement (v1.5.0). All experiments conducted adhered to the guidelines outlined in the PhysioNet Credentialed Health Data License Agreement. Additionally, this study has been deemed exempt from human subjects research.

References

Mousumi Akter, Naman Bansal, and Shubhra Kanti Karmaker. 2022. Revisiting automatic evaluation of extractive summarization task: Can we do better than ROUGE? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1547–1560, Dublin, Ireland. Association for Computational Linguistics.

Judith L. Bowen. 2006. Educational strategies to promote clinical diagnostic reasoning. *New England Journal of Medicine*, 355(21):2217–2225. PMID: 17124019.

Thomas J. DiCiccio and Bradley Efron. 1996. Bootstrap confidence intervals. *Statistical Science*, 11(3):189–212.

Yanjun Gao, John Caskey, Timothy Miller, Brihat Sharma, Matthew M. Churpek, Dmitriy Dligach, and Majid Afshar. 2022a. Tasks 1 and 3 from progress note understanding suite of tasks: Soap note tagging and problem list summarization. *PhysioNet*.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, John Caskey, Brihat Sharma, Matthew M. Churpek, and Majid Afshar. 2023. Dr.bench: Diagnostic reasoning benchmark for clinical natural language processing. *Journal of Biomedical Informatics*, 138:104286.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, Samuel Tesch, Ryan Laffin, Matthew M. Churpek, and Majid Afshar. 2022b. Hierarchical annotation for building a suite of clinical natural language processing tasks: Progress note understanding. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5484–5493, Marseille, France. European Language Resources Association.

Rebekah L Gardner, Emily Cooper, Jacqueline Haskell, Daniel A Harris, Sara Poplau, Philip J Kroth, and Mark Linzer. 2018. Physician stress and burnout: the impact of health information technology. *Journal of the American Medical Informatics Association*, 26(2):106–114.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14).

Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary C. Lipton. 2021. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4958–4972, Online. Association for Computational Linguistics.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Eric Lehman and Alistair Johnson. 2023. Clinical-t5: Large language models built using mimic clinical text. *PhysioNet*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jennifer Liang, Ching-Huei Tsou, and Ananya Poddar. 2019. A novel system for extractive clinical note summarization using EHR data. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 46–54, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium. Association for Computational Linguistics.

Darshna Patel, Saurabh Shah, and Hitesh Chhinkaniwala. 2019. Fuzzy logic based multi document summarization with improved sentence scoring and redundancy removal technique. *Expert Systems with Applications*, 134:167–177.

Vimla L. Patel and Guy J. Groen. 1986. Knowledge based solution strategies in medical reasoning. *Cognitive Science*, 10(1):91–116.

David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. Carbon emissions and large neural network training.

Long N Phan, James T Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. Scifive: a text-to-text transformer model for biomedical literature. *arXiv preprint arXiv:2106.03598*.

Jonathan Pilault, Raymond Li, Sandeep Subramanian, and Chris Pal. 2020. On extractive and abstractive neural document summarization with transformer language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9308–9319, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Adam Rule, Steven Bedrick, Michael F. Chiang, and Michelle R. Hribar. 2021. Length and Redundancy of Outpatient Progress Notes Across a Decade at an Academic Medical Center. *JAMA Network Open*, 4(7):e2115334–e2115334.

Chaitanya Shivade. 2019. Mednli - a natural language inference dataset for the clinical domain. *PhysioNet*.

Adam Wright, Justine Pang, Joshua C Feblowitz, Francine L Maloney, Allison R Wilcox, Karen Sax McLoughlin, Harley Ramelson, Louise Schneider, and David W Bates. 2012. Improving completeness of electronic problem lists through clinical decision support: a randomized, controlled trial. *Journal of the American Medical Informatics Association*, 19(4):555–561.

Qianqian Xie, Zheheng Luo, Benyou Wang, and Sophia Ananiadou. 2023. A survey on biomedical text summarization with pre-trained language model.

A Appendix

This section adds two more tables to show the results of the other clinical task. The results are compared with previous results and can be seen in the baseline column. In many cases, such as MedNLI, AP, and EmrQA, we can see improvement in the multitask experiments.

Model	Training	Summarization	SOAP	A/P
Gao et al., 2023	Single task	7.60 (5.31 - 9.89)	**60.12** (59.33 - 60.90)	80.09 (79.32 - 83.23)
T5 220M	Single task	26.35 (22.18 - 30.52)	**60.12** (59.33 - 60.90)*	73.31 (71.34 - 77.65)*
	Multi-task	24.84 (20.28 - 29.40)	56.63 (55.83 - 57.42)	43.25 (41.35 - 66.59)
T5 770M	Single task	26.90 (22.58 - 31.23)	55.57 (54.78 - 56.35)*	77.96 (75.38 - 81.60)*
	Multi-task	23.99 (19.86 - 28.13)	51.10 (50.32 - 51.91)	75.15 (71.93 - 78.19)
SciFive 220M	Single task	25.31 (21.45, 29.17)	57.74 (56.95 - 58.53)*	76.76 (74.81 - 80.92)*
	Multi-task	24.38 (19.99 - 28.78)	54.86 (54.06 - 55.65)	68.87 (65.50 - 72.12)
SciFive 770M	Single task	27.31 (23.09 - 31.53)	47.65 (46.85 - 48.47)*	75.11 (73.10,79.42)*
	Multi-task	25.31 (21.45 - 29.17)	44.51 (43.72- 45.29)	77.50 (74.45 - 80.37)
Clinical-T5 220M	Single task	25.35 (21.19 - 29.51)	55.30 (54.51 - 56.11)	80.44 (77.47 - 83.35)
	Multi-task	26.21 (21.92 - 30.49)	52.41 (51.62 - 53.20)	65.49 (62.08 - 68.76)
Clinical-T5 770M	Single task	28.28 (24.17 - 32.38)	52.82 (52.03 - 53.61)	78.79 (75.76 - 81.66)
	Multi-task	**28.55** (24.29 - 32.80)	54.00 (53.21 - 54.80)	**80.58** (77.57 - 83.38)

Table 3: Finetuned T5 models on various clinical task with 95% confidence interval calculated using the bootstrapping method. A/P represents assessment and plan relational labeling task. Summarization use ROUGE L, A/P use F1-macro and SOAP use accuracy score for the evaluation metrics. The first row in the table represents best scores reported in the DR.BENCH paper and * in the other rows represent scores for the respective task in DR.BENCH paper (Gao et al., 2023)

Model	Training	EmrQA	MedNLI	MedQA
Gao et al., 2023	Single task	39.20 (34.63 - 43.78)	84.88 (82.98 - 86.64)	24.59 (22.31 - 27.02)
T5 220M	Single task	33.40 (29.27 - 37.61)*	79.75 (78.62 - 82.70)*	22.55 (20.01 - 25.69)*
	Multi-task	38.48 (37.24 - 39.79)	72.57 (70.18 - 74.82)	21.75 (19.48 - 24.12)
T5 770M	Single task	38.05 (33.56 - 42.58)*	84.04 (82.14 - 85.86)*	20.97 (18.77 - 23.25)*
	Multi-task	41.42 (40.16, 42.72)	83.19 (81.22, 85.09)	23.25 (20.97, 25.61)
SciFive 220M	Single task	37.28 (32.84 - 42.11)*	82.84 (80.87 - 84.74)*	22.78 (20.50 - 25.14)*
	Multi-task	40.08 (38.82 - 41.39)	78.83 (76.72 - 80.94)	21.52 (19.32 - 23.80)
SciFive 770M	Single task	41.21 (39.93 - 42.49)	83.89 (82.00 - 85.79)	23.09 (20.82 - 25.37)
	Multitask	41.26 (39.98 - 42.56)	84.35 (82.49 - 86.22)	23.72 (21.37 - 26.08)
Clinical-T5 220M	Single task	41.35 (40.07 - 42.65)	84.32 (82.42 - 86.15)	21.92 (19.64 - 24.19)
	Multi-task	40.30 (39.02 - 41.62)	71.23 (68.92 - 73.56)	22.46 (20.19 - 24.74)
Clinical-T5 770M	Single task	**42.69** (41.39 - 43.95)	85.86 (85.02 - 88.47)	24.27 (21.92 - 26.63)
	Multi-task	42.61 (41.34 - 43.92)	**86.14** (84.32 - 87.90)	**25.84** (23.41 - 28.28)

Table 4: Finetuned T5 models on various clinical task with 95% confidence interval calculate using the bootstrapping method. All the evaluation metrics here are the accuracy score. The first row in the table represents best scores reported in the DR.BENCH paper and * in the other rows represent scores for the respective task in DR.BENCH paper (Gao et al., 2023)

Input	**Ground Truth Diagnoses/ Problems**	**T5 770M Single task**	**Clinical-T5 770M Single task**	**Clinical-T5 770M Multi-task**
SUMMARIZE: ***<ASSESSMENT>*** 48 y/o M with HIV 47M s/p elective spinal surgery (anterior and posterior LIFs), intubated - - - - ***<SUBJECTIVE>*** Agitated, diaphoretic, altered, hypertensive and tachy this AM - - - - ***<OBJECTIVE>*** Last dose of Antibiotics: Infusions: Other ICU medications: Heparin Sodium (Prophylaxis) - - - -	EtOH withdrawal Spinal surgery	Altered MS s p elective spinal surgery	DTs EtOH w d pain h o chronic pain	Altered mental status Hypertension Tachycardia Acute renal failure s p spinal surgery
SUMMARIZE: ***<ASSESSMENT>*** SEPSIS WITHOUT ORGAN DYSFUNCTION Ms. [**Known lastname 10381**] is a 76F with multiple medical problems, who is - - - - - - - ***<SUBJECTIVE>*** FEVER - 101.7 F - [**2129-9-3**] 12:33 PM -received boluses overnight for low SBP - - - - ***<OBJECTIVE>*** Last dose of Antibiotics: Cefipime - [**2129-9-3**] 04:05 PM Metronidazole - [**2129-9-4**] 04:00 AM - - - -	Sepsis Patient has re developed fevers on 9 2 on a regimen of vancomycin ceftriaxone Possible sources include 1 Intra abdominal source	Sepsis Thrombo-cytopenia	Sepsis Likely source is clostridium difficile colitis Acute renal failure	Hypotension Likely secondary to sepsis though source unclear at this time Acute renal failure

Table 5: The table represents a snippet of the input and output sections of problem summarization. The input data contains an added prefix that denotes the task for T5, "SUMMARIZE" in this case, and <prefix> that defines the note section. Finally, "- - - -" is the continuation of the section, which was excluded here due to the space constraint.

Context-aware Medication Event Extraction from Unstructured Text

Noushin Salek Faramarzi [*] **Meet Patel** [*] **Harika Bandarupally** [*] **Ritwik Banerjee** [*§]
[*] Department of Computer Science
[§] Institute for AI-Driven Discovery and Innovation
Stony Brook University, New York, USA
{nsalekfarama,meppatel,hbandarupall,rbanerjee}@cs.stonybrook.edu

Abstract

Accurately capturing medication history is crucial in delivering high-quality medical care. The extraction of medication events from unstructured clinical notes, however, is challenging because the information is presented in complex narratives. We address this challenge by leveraging the newly released Contextualized Medication Event Dataset (CMED) as part of our participation in the 2022 National NLP Clinical Challenges (n2c2) shared task. Our study evaluates the performance of various pretrained language models in this task. Further, we find that data augmentation coupled with domain-specific training provides notable improvements. With experiments, we also underscore the importance of careful data preprocessing in medical event detection.

1 Introduction

Ensuring the accuracy of a patient's treatment history is essential for delivering high-quality medical care. This allows healthcare professionals to assess the effectiveness of existing treatments, detect possible medication-related problems, and suggest appropriate future treatment options (FitzGerald, 2009). Various forms of treatment changes, however, are often absent from structured electronic data sources, being recorded only in clinical narratives (Turchin et al., 2009). An accurate extraction of medication event information from unstructured data in patients' medical records is thus crucial for a complete understanding of their treatments.

When extracting medication changes from clinical text, it is necessary to take into account various forms of contextual information, due to the narrative and longitudinal nature of clinical documentation. Clinical text often documents events over a patient's medical history, and providers may also record the reasoning behind their medical decisions. These factors result in complex events that cannot be properly captured by extracting medication changes alone, without considering the surrounding clinical context. This is especially true when developing a medication change extraction system to support real-world applications, such as medication timeline generation (Plaisant et al., 2003; Belden et al., 2019) or medication reconciliation (Poon et al., 2006; Cadwallader et al., 2013). Indeed, as Wang et al. (2018) have argued, the use of sophisticated natural language processing (NLP) information extraction (IE) become a necessity when the automatic extraction of relevant medical information is required by large-scale or real-time applications further downstream, such as clinical research and decision support.

This study investigates how to extract information about changes to patient medications from clinical text using the Contextualized Medication Event Dataset (CMED) developed by Mahajan et al. (2021) and subsequently released to the community as a shared task in 2022 National NLP Clinical Challenges (n2c2)[1]. This consists of three tasks: (i) medication extraction, to extract all medication mentions in clinical notes, (ii) event classification, to identify whether a medication change is discussed in an event,and (iii) context classification, to classify the contextual information of a medication change event along five orthogonal dimensions, with each dimension further classified into multiple *attributes* of the event.

For the first medical named entity recognition task, we note that Lee et al. (2020) demonstrate significant improvements with the use of BioBERT, a domain-specific model initialized with BERT and then pretrained on PubMed abstracts and PubMed Central full text articles. Thus, we proceed to use BioBERT as well, providing comparisons against popular general purpose language models like BERT (Devlin et al., 2019). Additionally, we also utilize Bio+Clinial BERT (Alsentzer et al., 2019), another domain-specific model initialized

[1] n2c2.dbmi.hms.harvard.edu/2022-track-1

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 86–95
July 14, 2023

with BioBERT and further pretrained on notes from the MIMIC-III dataset (Johnson et al., 2016).

We evaluate the performance of several pretrained language models for the second and third tasks. Specifically, we examine three popular general purpose models – BERT, RoBERTa (Liu et al., 2019b), and XLNet (Yang et al., 2019) – and one domain-specific pretrained model, Bio+Clinical BERT.

While many downstream natural language understanding tasks are readily successful when a large pretrained language model is tuned for the task in hand, we observe that the accuracy of clinical event detection crucially depends on careful data preprocessing. In particular, identifying the proper linguistic context from surrounding text is of utmost importance. To that end, we develop and employ a sentence detection method tailored to this task, leading to a better performance by all models. We also find that augmenting the data with the DDI (drug-drug interaction) Corpus (Herrero-Zazo et al., 2013) leads to overall improvements in medication change detection and its context classification.

2 Related Work

The first task in contextualized medication event extraction is to extract the mention of medications – clearly, a medical named entity recognition (NER) task. Medical NER, in general, includes identifying other types of entities such as diseases, symptoms, proteins, or patient information (see Pagad and Pradeep (2022) for an overview). To identify medication names in particular, many approaches have been proposed. Early methods relied explicitly on domain ontology or dictionaries (Sanchez-Cisneros et al., 2013), rules (Segura-Bedmar et al., 2008), and subsequently, contextual rules and automatically learned rules (Hamon and Grabar, 2010; Coden et al., 2012). A comprehensive survey of this literature has been conducted by Liu et al. (2015). More recent approaches are hybrid, combining LSTM and its variants with conditional random fields (CRF) or other graphical models (Alfattni et al., 2021; Jouffroy et al., 2021). Even more recent, however, are techniques that utilize Transformer models (*e.g.*, BERT). There is some work to further indicate that combining BERT with BiLSTM-CRF improves medical NER (Yu et al., 2019), while others demonstrate the improvements in using domain-specific pretraining with BERT initialization (Lee et al., 2020).

Identifying medication change events and classifying their attributes, however, is a significantly less explored problem. This is due largely to the scarcity of annotated resources, but to a lesser extent, also to the complexity of the language used in clinical narratives to describe such events. Initially, research heavily relied on annotated datasets like the 2009 i2b2 and the 2013 DDI datasets (Uzuner et al., 2011; Herrero-Zazo et al., 2013). Some early work focused on very specific events of clinical relevance, such as Liu et al. (2019a), who inspect medication discontinuation, or Sohn et al. (2010), who focus on whether medication was started, stopped, increased, or decreased. In another approach, Pakhomov et al. (2002) introduced temporal information into their labels. In spite of the success on individual datasets, these approaches employ rule-based decisions and classical supervised learning algorithms like support vector machines (SVMs) or maximum entropy modeling, which are unlikely to generalize across multiple datasets with linguistic variation without extensive supervision for each dataset.

For a detailed understanding of treatments, such as extracting the dosage, frequency, or mode of drug administration, or in determining its relation to other phenomena like adverse drug effects, generalizable success in this task carries immense significance. It is thus worth noting that recent methods leveraging neural architectures and models have shown promise in medical event extraction and classification tasks (Narayanan et al., 2022). Lerner et al. (2020) use a neural top-down transition based parser and achieve results comparable to BiLSTM models for medical entity and event detection. Perhaps the closest to our study is the approach of Lybarger et al. (2021), who tune BERT on COVID-19 data to identify various events of clinical significance, such as symptoms, severity, and assertion. This body of work is distinct from ours, however, since it does not delve into classification of event attributes involving complex temporal or conditional expressions.

Several studies (Uzuner et al., 2011; Chapman et al., 2001; Szarvas et al., 2008; Morante, 2010; Albright et al., 2013) have examined the detection of negated medical concepts in clinical text. However, none of them specifically focus on identifying medication change events. Moreover, they have not looked at the combined identification of negation and the actor responsible for that negation. Early

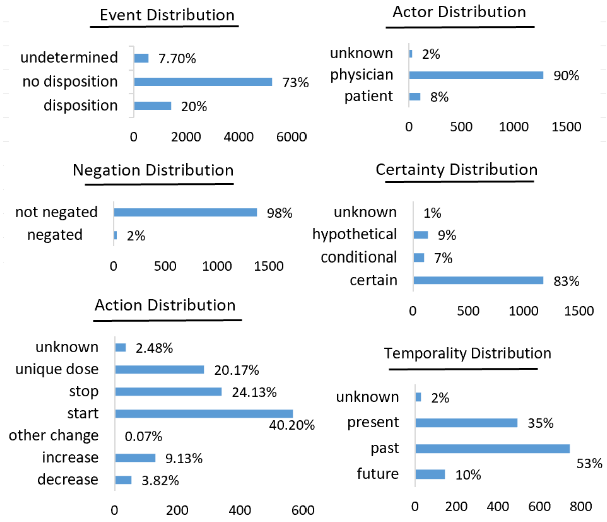


Figure 1: Distribution of labels in CMED (training set).

work on negation detection in clinical texts was based on negation lexicons, and rule-based algorithms using them. Most notable among them is NegEx (Chapman et al., 2001). Although these have been superceded by others who combine lexicons with dependency structures or other linguistic features (Mehrabi et al., 2015), we use an implementation of NegEx built into the popular spaCy[2] library, called negspaCy. Our results (Sec. 6) show that in spite of its simplicity, this approach suffices.

With the *Contextual Medication Event Dataset*, CMED, Mahajan et al. (2021) provide annotated data capturing five orthogonal dimensions of contextual information related to medication change events. Further, they demonstrate the viability of SVMs and Transformer-based models in detecting and classifying these events. Very recently, Ramachandran et al. (2023) have explored an avenue similar to ours, with domain-specific language models based on BERT. In these, it has been noted that sentences that mention multiple drugs are particularly difficult to detect and classify. By contrast, our work investigates data augmentation and task-specific preprocessing in conjunction with the user of domain-specific language models. In particular, we develop and use a custom sentence extraction module in our pipeline, which improves the accuracy of these models on the given tasks.

3 The CMED Dataset

The CMED dataset comprises annotated clinical notes, where each medication mention is assigned one or more event labels from the three categories:

(1) `Disposition`, indicating the mention of a medication change, *e.g.*, "prescribed albuterol for shortness of breath",

(2) `NoDisposition`, indicating that the mediction is mentioned with no indication of change, *e.g.*, "patient continues to take aspirin", and

(3) `Undetermined`, indicating a lack of clarity or evidence regarding medication change, *e.g.*, "Plan: Lasix".

For each event identified as `Disposition`, the clinical context is provided along five orthogonal dimensions, *viz.*, `action`, `actor`, `negation`, `certainty`, and `temporality`. We describe these next.

(1) `action` refers to the type of change is being made. Its attributes are *start*, *stop*, *increase*, *decrease*, *unique dose*, *other change*, and *unknown*.

(2) `actor` specifies who initiated the action, *physician*, *patient*, or *unknown*.

(3) `negation` indicates whether the action is negated or not.

(4) `temporality` specifies whether the action takes place in the *past*, *present*, or *future*.

(5) `certainty` characterizes the likelihood of the action taking place as *certain*, *hypothetical*, *conditional*, or *unknown*.

The distribution of the event and attribute labels in the training set of this dataset is shown in Fig. 1.

4 Data Preprocessing

We observe that in CMED, most medication mentions are labeled with one event and a set of corresponding attributes. There is, however, a small fraction (< 90 instances in the training set), where the drug mention is labeled with two events and two separate sets of attributes, as noted by Ramachandran et al. (2023) as well. Further, we underscore the frequent presence of sentences containing multiple drug mentions (approx. 78%), with a substantial fraction (over 50%) of such sentences mentioning four or more drugs simultaneously. This aspect of the dataset significantly increases the complexity of detecting and classifying contextual information from clinical sentences. Finally, we note that some events cannot be accurately labeled based solely on the sentence in which they appear, and additional context from neighboring sentences becomes necessary to determine the correct attributes. Table 1 presents examples from CMED showcasing illustrative examples of these phenomena.

[2] spacy.io/

(a) “The patient’s daily dose of **furosemide** was increased from 40mg to 80mg. and then reduced to 60mg daily.” LABELS: *increase*, *decrease*
(b) “The healthcare provider started the patient on a new regimen of **metformin** and discontinued the use of **pioglitazone**.” LABELS: *start*, *stop*
(c) “The healthcare provider instructed the patient to take **acetaminophen** if their fever rises above 100 degrees.” LABELS: *conditional*

Table 1: Examples from clinical notes where (a) one drug mention indicates two events with opposite action labels, and (b) two drug mentions, each with their own action labels. Also, (a) and (c) have grammatically valid sentences up until the line break, but the sentence continues. Stopping at the line break will miss the language responsible for the *decrease* and *conditional* labels.

The first step in the medication information extraction task is to prepare the dataset by extracting the sentences containing medication information. However, due to the unstructured and lengthy nature of medical notes, accurately identifying the start and end of a sentence containing a medication mention is challenging. Accordingly, relying solely on tools like, say, spaCy, for their inbuilt sentence parser for this task does not produce satisfactory results. Therefore, we develop a customized approach to accurately identify the sentences that contained medication names, which served as a crucial first step towards performing accurate medication event extraction. Next, we describe the steps of this process.

(i) Abbreviation resolution. Abbreviations such as “Continue” and “Discontinue” are converted to their full forms to facilitate accurate identification of medication mentions in the text. One of the most frequent and important abbreviations is “Discontinue,” which is observed in different forms with various spacing (e.g., “d/c’ed,” “d/c’d,” “d/ c’d,” “d/ c,” “D /c,” etc.). Similarly, “Continue” is abbreviated as “c’d” or “Cont.” Having the full form of these words is important because sentence/token chunkers trained on general purpose language are sensitive to punctuation, and non-standard punctuation as described above may mislead them. For example, if chunking happens in the text “d/ c’d glucophage” as (“d/”, “c’d”, “glucophage”), the model might conceive this text as a continuation rather than discontinuation.

(ii) Coreference resolution. This is an essential step in our text preparation, as it not only improves the clarity of the text but also contributes to more accurate classification of actor attributes. For example, consider a sentence like “The patient was given medication X by their doctor, who also advised them to increase their water intake.” Here, coreference resolution helps to identify that “the patient” and “them” are referring to the same entity, and that “their doctor” and “who” are referring to the same entity. This information is crucial for accurate actor classification, which can inform downstream tasks such as adverse event detection and pharmacovigilance. Therefore, we utilize AllenNLP’s[3] coreference resolution model as part of our text preprocessing pipeline to replace the repeated mentions of entities with their corresponding coreferents.

(iii) Sentence Extraction based on syntactic dependencies. Each sentence is split into its constituent phrases. We then use the spaCy library to parse each phrase into a tree of syntactic dependencies, and identify coordinated conjunction phrases (e.g., “and” or “or” phrases) in the tree. Following that, we construct a list of the longest continuous sequences of words that are dependent on these conjunctions, and remove any conjunctions from the beginning or end of each sequence. This is done by traversing the tree and collecting all conjuncts connected to the root of the tree. Finally, a list of strings representing each identified phrase is combined to form a single string. This string is taken as the sentence that contains the medication mention and its surrounding context. Algorithm 2 is responsible for finding the coordinated conjunction phrases from the parse tree, and Algorithm 1 is responsible for extracting the phrase chunks from a sentence with the aid of dependency parsing.

(iv) Sentence separation. Here, the objective is to break sentences with multiple medication names and their corresponding multiple event types. This allows us to accurately identify the events associated with each drug name. We split these sentences into different clausal components. For example, consider the sentence “Started lisinopril 10 mg p.o. daily, substituted for diltiazem.”. Clearly, the verb “started” is associated with the medication “lisinopril 10 mg p.o. daily”, and the verb “substituted” is associated with “diltiazem”. While dependency

[3] allenai.org/allennlp

Algorithm 1 get_conjunction(head)

```
acc ← [], list_heads ← [head]
while list_heads ≠ [] do
    new_heads ← []
    for h in list_heads do
        children ← children of h with dependency tags
            "conj" or "ccomp"
        if children ≠ [] then
            append children to new_heads and acc
        end if
    end for
    list_heads ← new_heads
end while
return acc
```

Algorithm 2 get_chunks(sentence)

```
doc ← parse sentence using spaCy, chunks ← []
for sent in doc do
    conj_phrases ← get coordinated conjunction
        phrases from sent's root using get_conjunction(head)
    for head in conj_phrases do
        append head's subtree to chunks
    end for
end for
sort chunks in ascending order of length
seen ← empty set, trimmed_chunks ← []
for chunk in chunks do
    c2 ← list of unconsumed tokens in chunk
    update seen set with indices of tokens in c2
    c3 ← longest continuous sequence of tokens in c2
    append longest sequence in c3 to trimmed_chunks
end for
output ← []
for phrase in trimmed_chunks do
    remove any conjunctions at the beginning or end of
        phrase
    join the tokens in phrase to form a string
    remove any leading or trailing commas from the string

    append the string to output
end for
sort output in the original order of phrases in sentence
return output
```

parsing is capable of distilling these relations, we observe that sentences in the CMED dataset can usually be split into separate clauses where each clause exhibits only one medication change event. In our example, this approach leads to two such simpler expressions, "started lisinopril 10 mg p.o. daily" and "substituted for diltiazem".

5 Approach

In this section, we explain our technical approach to the tasks of medication mention extraction, medical event identification, and medication event attribute classification. Further, we devote a separate description of the steps we take to detect negation.

5.1 Medication mention extraction

The task of medication mention extraction involves identifying multi-word medication phrases within free-text. As such, it is similar to medical named entity recognition (NER). Following the vast body of work that treats NER as a sequence tagging task, we utilize the beginning-inside-outside (BIO) label prefixes. Typically, medication phrases within CMED are brief, consisting of three or fewer tokens in most cases. Our approach to identifying medication mentions involves the use of BERT-based models, specifically those pretrained on domain-specific data, such as BioBERT and Bio+Clinical BERT. By adding a linear output layer and fine-tuning these models, we improve our ability to predict the specific location of medication references.

To enhance our medication mention extraction model, we experiment with incorporating the DDI (drug-drug interaction) Extraction 2013 corpus (Herrero-Zazo et al., 2013) into our training data. This is a widely recognized corpus comprising sentences from biomedical literature discussing drug-drug interactions, with each sentence annotated to indicate the medications involved in the interaction and the type of their interaction. Employing this corpus allows us to expand the number of medication mentions in our training set, leading to improved performance. The results subsequently obtained, by training only on CMED and then by training on data augmented by the DDI corpus, are shown for comparison in Table 2.

5.2 Identifying negation

Even though prior work on clinical event identification has largely avoided complex negation detection, the task is nevertheless subsumed by research directed at understanding medication changes in clinical notes. In CMED, however, we find negation to be present in a very small proportion of the samples (2%). To correctly handle these instances, we employ negspaCy[4], a Python library that provides pretrained models and tools for detecting negation and other linguistic phenomena in text data. It is specifically designed to identify negated concepts, such as negated medical conditions or treatments, which are commonly encountered in clinical narratives. The library uses a combination of rule-based and statistical methods to identify negation, including the use of dependency parsing, word embeddings, and machine learning algorithms. In our study, we use Med7 (Kormil-

[4] pypi.org/project/negspacy

Table 2: Medication mentions extraction performance on the CMED test set. DDI+CMED is the combined training set of the DDI corpus and the CMED.

Dataset	Model	Strict			Lenient		
		P	R	F1	P	R	F1
CMED	BERT	0.90	0.90	0.90	0.92	0.92	0.92
	BioBERT	0.95	0.95	0.95	0.96	0.95	0.95
	Bio+Clinical BERT	0.93	0.95	0.94	0.95	0.95	0.95
DDI+CMED	BERT	0.90	0.90	0.90	0.92	0.92	0.92
	BioBERT	0.96	0.96	0.96	0.96	0.96	**0.96**
	Bio+Clinical BERT	0.93	0.94	0.93	0.95	0.96	0.94

itzin et al., 2021), a model designed to extract medication-related information from clinical notes, and integrate it with negspaCy. This integration allows us to detect instances where a drug is mentioned in the text but not prescribed.

5.3 Event and attribute classification

As described earlier in Sec 3, a medication mention must be classified as `Disposition`, `NoDisposition`, or `Undetermined`. For those identified as disposition, *i.e.*, indicating a change in medication, the next stage of the pipeline requires identifying the dimensions `action`, `actor`, `negation`, `certainty`, and `temporality` of the event, along with the correct attribute values for each dimension.

For the event and the rest of the attributes, we train a classification model based on transformer-based language models. The event and attribute classification systems assume gold standard medication mentions for model training and comparison. We conduct our experiment using Bidirectional Encoder Representations from Transformer (BERT) models pretrained on general purpose and clinical datasets. Specifically, we use BERT, RoBERTa, Bio+Clinical BERT, and XLNet. The last model, XLNet, is slightly different from the others in that it is an autoregressive Transformer model. We include it with the hope of leveraging the advantages of autoregressive language modeling as well as autoencoding.

Our goal is to classify the medication events using the sentence containing the detected medication mention as context. We use a pretrained Transformer to create a distributed representation, add 0.2 dropout, and use a fully connected layer of size 5 with softmax activation for classification. For fine-tuning with the training and development sets of CMED, we use the Hugging Face transformers package (Wolf et al., 2020). This is a multi-class classification approach, producing predictions at the sentence level for the event as well as its associated dimensional attributes.

This approach does not rely on any explicit knowledge or indication of where the medication is located. During our data preprocessing technique, we ensure that two distinct medications with varying event types are separated into their respective clauses (see Sec 4). This prevents distinct medication mentions from linguistically sharing the same events and event properties. Event classification and the attribute classification are, however, treated as separate tasks. Moreover, each attribution classifier is also trained separately. Thus, if a model is trained to predict the event type of a sentence, it will only be exposed to that specific type of label and will not be able to incorporate information from other label types.

6 Evaluation

To evaluate the accuracy of medication mention extraction systems, we employ two criteria: strict and lenient match. The strict criteria demands an exact match between the predicted and true medication mention spans. The lenient match criteria, on the other hand, considers a predicted medication mention to be correct if at least one token in the predicted mention overlaps with a token in the true mention. While strict criteria may provide a more conservative performance estimate, lenient criteria can identify more correct predictions, but at the expense of higher false positive rates. To evaluate the event and attribute classification systems, we employ precision, recall, and F1 scores, reporting both macro- and micro-averages.

6.1 Medication mention extraction

The performance of BERT, BioBERT, and Bio+Clinical BERT on this task are shown in Table 2. BioBERT achieves the highest F_1 score in strict (0.95) as well as lenient (0.96) evaluation criteria. Bio+Clinical BERT, on the other hand, achieves the highest precision scores (0.95 in both strict and lenient criteria). The slightly lower score of BERT is unsurprising, given its lack of pretraining on domain-specific data. We also note that upon augmenting the training data with the DDI corpus, a slight improvement can be seen in the F_1 score achieved by BioBERT. For Bio+Clinical BERT, however, the results are mixed. The purported advantage of this model is its pretraining on

Task		BERT			RoBERTa			XLNet			Bio+Clinical BERT		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Event	**Micro**	0.91	0.91	0.91	0.92	0.92	0.92	0.91	0.91	0.91	0.93	0.93	**0.93**
	Macro	0.85	0.76	0.80	0.84	0.80	0.82	0.85	0.79	0.81	0.90	0.82	0.85
Action	**Micro**	0.78	0.76	0.77	0.82	0.80	0.81	0.79	0.78	0.79	0.83	0.83	**0.83**
	Macro	0.77	0.75	0.76	0.82	0.80	0.81	0.79	0.77	0.78	0.83	0.68	0.72
Temporality	**Micro**	0.75	0.75	0.74	0.69	0.79	**0.81**	0.78	0.70	0.74	0.78	0.70	0.74
	Macro	0.72	0.62	0.68	0.63	0.59	0.61	0.65	0.72	0.68	0.75	0.70	0.70
Certainty	**Micro**	0.85	0.83	0.84	0.86	0.84	0.85	0.87	0.85	**0.86**	0.80	0.64	0.71
	Macro	0.83	0.82	0.81	0.84	0.74	0.83	0.83	0.75	0.85	0.78	0.70	0.70
Actor	**Micro**	0.92	0.91	0.92	0.91	0.89	0.90	0.94	0.92	0.93	0.93	0.93	**0.93**
	Macro	0.71	0.85	0.73	0.84	0.83	0.83	0.76	0.88	0.66	0.84	0.59	0.61

Table 3: Event and attribute classification results (with gold standard medication mentions) on the CMED test set.

biomedical literature as well as clinical notes. We conjecture that the lack of significant improvements is due to the augmentation not by clinical language, but by language from biomedical research literature (MedLine abstracts) and the DrugBank database, which form the DDI corpus.

6.2 Identifying negation

We evaluate the performance of our negation attribute classification, *i.e.*, label medication change events as *negated* or *not negated*, using Med7 and negspaCy integration. Despite the extremely small support (2% of CMED training set), our method achieves a near-perfect accuracy of 0.98. We also achieve precision, recall, and F_1 (macro average) of 0.82, 0.88, and 0.85, respectively.

6.3 Event and attribute classification

We report the results of event and attribute classification in Table 3, which shows the performance of the four language models BERT, RoBERTa, XLNet, and Bio+Clinical BERT, on the withheld CMED test set. Since this test set contains the gold-standard labels for medication mentions, our evaluations are conducted using the gold standard medication mentions as well.

Similar to results obtained by Ramachandran et al. (2023), all BERT-based language models perform well on these tasks. For event classification, the micro F_1 scores range from 0.91 to 0.93, while for attribution classification, they range from 0.77 to 0.86. In most cases, Bio+Clinical BERT outperforms the other models, achieving the highest F_1 score of 0.93 for event classification and 0.86 for `certainty` classification, as well as the highest precision of 0.94 for `actor` classification. We do report some unexpected success with RoBERTa and XLNet as well, which achieve the highest F_1 in `action` (0.83) and `temporality` (0.81) classification, respectively.

Further, we observe that the macro F_1 scores are generally lower than the micro F_1 scores, indicating that the models struggled with some classes. Specifically, `temporality` and `actor` classification showed lower performance across all models.

6.4 Discussion

When using pretrained language models to extract medication changes from clinical narratives, multiple event annotations for medication mentions can be a significant challenge, leading to prediction errors. For example, the sentence "Lovenox (will clarify timing of surgery and hold accordingly)" has two labels for the event (`undetermined` and `disposition`) for the medication Lovenox, potentially confusing the model. One solution to this issue is to modify the task from a sentence classification task to a multi-label classification task. However, there may be cases where a sentence follows a multi-label scheme, but only one type of annotation is provided. For instance, "DM2: Continue home meds (metformin + insulin), hold when on diet without substantial calories (clears, NPO)" only has the `action` label *start* for the metformin, whereas there is a need for the second attribute label *stop* as well.

During our analysis, we observed instances of incorrect or ambiguous labeling in the annotation, including the `actor` and `temporality` dimensions. For example, in "SL TNG prescribed but not used," there are two `actor` labels (*patient* and *physician*) for the medication TNG, and in "amox 500 TID x 10d: fluids, steam, acetaminophen," the medication amox has two `temporality` labels (*past* and *future*). Furthermore, in "We will initiate Zetia to add to the Pravachol," the event is labeled as `NoDisposition`, despite the word "initiate" clearly suggesting otherwise.

Additionally, we noticed several mistakes in the negation class, such as "Not on beta-blocker " being labeled as non-negated. The limited number of samples in the negated category, combined with the annotation errors in the test set, has a clear and significant impact on any model. As the model's training relies heavily on the quality and quantity of the data, a small and incorrectly labeled dataset is particularly harmful. We also noticed several non-English sentences in the training set, such as "Hctz (HYDROCHLOROTHIAZIDE) 12.5 MG (12.5MG CAPSULE take 1) PO QD, Para la presión alta- si se siente muy mareado deje de tomarla y avísele a su médico immediatamente." While any effort to utilize the advances of natural language processing in clinical applications in multiple languages is laudable, the presence of very few instances of other languages in an otherwise English corpus has a negative impact.

It is noted in the dataset annotation that medication-related information is contained within a single sentence. However, we observe that this is not always the case. There are several instances where information about a single medication event extends beyond a single sentence, requiring the model to analyze multiple sentences in order to identify the relevant context. The dataset includes a number of sentences that are labeled as *undetermined*, many of which are located within an assessment and plan (A/P) section of a medical document. This section can be quite lengthy and contain numerous mentions of medications without specific attributes or events. To correctly classify these undetermined sentences, it is often necessary to look beyond the sentence itself and recursively search for information related to medication events within the A/P section. However, we believe this is a challenging task beyond the ambit of the CMED dataset's annotation description.

7 Conclusion

Our analysis of CMED and its constituent tasks reveal three main characteristics. First, it is often necessary to consider additional context beyond the specific sentence containing the medication mention to accurately label medication references. This context could include information from previous or subsequent sentences, the patient's medical history, or other relevant information further away in a document (as often found in the assessment and plan sections) that could impact the interpretation of the medication mention. Second, we observe the prevalence of multiple medication references within a single sentence, which poses a challenge for accurate extraction. Finally, accurate identification of the start and end of a sentence containing a medication mention is also challenging, since standard sentence splitting and tokenization methods often fail in clinical notes, especially if task-specific or domain-specific preprocessing is not done.

We especially underscore the importance of data preprocessing when training or fine-tuning models for the medical domain. In this work, for example, we perform abbreviation resolution, coreference resolution, syntactic dependency-based sentence extraction, and a custom sentence extraction with phrase chunking.

Similar to other recent findings, our study demonstrates that pretrained language models are extremely effective in complex clinical information extraction, when fine-tuned on carefully chosen domain data. Overall, our study affirms the utility of Transformer-based models, particularly BioBERT and Bio+Clinical BERT, in medication information extraction from clinical notes. We also exhibit the additional advantage of training such models with augmented domain data.

References

Daniel Albright, Arrick Lanfranchi, Anwen Fredriksen, William F. Styler IV, Colin Warner, Jena D. Hwang, Jinho D. Choi, Dmitriy Dligach, Rodney D. Nielsen, James Martin, Wayne Ward, Martha Palmer, and Guergana K. Savova. 2013. Towards comprehensive syntactic and semantic annotations of the clinical narrative. Journal of the American Medical Informatics Association, 20(5):922–930.

Ghada Alfattni, Maksim Belousov, Niels Peek, Goran Nenadic, et al. 2021. Extracting Drug Names and Associated Attributes From Discharge Summaries: Text Mining Study. JMIR medical informatics, 9(5):e24678.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Jeffery L. Belden, Pete Wegier, Jennifer Patel, Andrew Hutson, Catherine Plaisant, Joi L. Moore, Nathan J. Lowrance, Suzanne A. Boren, and Richelle J. Koopman. 2019. Designing a medication timeline for patients and physicians. Journal of the American Medical Informatics Association, 26(2):95–105.

Justin Cadwallader, Kenneth Spry, Justin Morea, AL Russ, Jon Duke, and Michael Weiner. 2013. Design of a Medication Reconciliation Application. Applied Clinical Informatics, 4(01):110–125.

Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. 2001. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. Journal of Biomedical Informatics, 34(5):301–310.

Anni Coden, Daniel Gruhl, Neal Lewis, Michael Tanenblatt, and Joe Terdiman. 2012. SPOT the drug! An unsupervised pattern matching method to extract drug names from very large clinical corpora. In Proceedings of the IEEE 2nd International Conference on Healthcare Informatics, Imaging and Systems Biology, pages 33–39, San Diego, CA, USA.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Richard J. FitzGerald. 2009. Medication errors: the importance of an accurate drug history. Br J Clin Pharmacol, 67(6):671–675.

Thierry Hamon and Natalia Grabar. 2010. Linguistic approach for identification of medication names and related information in clinical narratives. Journal of the American Medical Informatics Association, 17(5):549–554.

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions. Journal of Biomedical Informatics, 46(5):914–920.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Liwei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovitz, Leo A. Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. Scientific Data, 3:160035.

Jordan Jouffroy, Sarah F Feldman, Ivan Lerner, Bastien Rance, Anita Burgun, Antoine Neuraz, et al. 2021. Hybrid deep learning for medication-related information extraction from clinical texts in French: MedExt algorithm development study. JMIR Medical Informatics, 9(3):e17934.

Andrey Kormilitzin, Nemanja Vaci, Qiang Liu, and Alejo Nevado-Holgado. 2021. Med7: A transferable clinical natural language processing model for electronic health records. Artificial Intelligence in Medicine, 118:102086.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4):1234–1240.

Ivan Lerner, Jordan Jouffroy, Anita Burgun, and Antoine Neuraz. 2020. Learning the grammar of drug prescription: recurrent neural network grammars for medication information extraction in clinical texts. arXiv preprint arXiv:2004.11622.

Feifan Liu, Richeek Pradhan, Emily Druhl, Elaine Freund, Weisong Liu, Brian C Sauer, Fran Cunningham, Adam J Gordon, Celena B Peters, and Hong Yu. 2019a. Learning to detect and understand drug discontinuation events from clinical narratives. Journal of the American Medical Informatics Association, 26(10):943–951.

Shengyu Liu, Buzhou Tang, Qingcai Chen, and Xiaolong Wang. 2015. Drug Name Recognition: Approaches and Resources. Information, 6(4):790–810.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.

Kevin Lybarger, Mari Ostendorf, Matthew Thompson, and Meliha Yetisgen. 2021. Extracting COVID-19 diagnoses and symptoms from clinical text: A new annotated corpus and neural event extraction framework. Journal of Biomedical Informatics, 117:103761.

Diwakar Mahajan, Jennifer J Liang, and Ching-Huei Tsou. 2021. Toward Understanding Clinical Context of Medication Change Events in Clinical Narratives. In AMIA Annual Symposium Proceedings, volume 2021, page 833. American Medical Informatics Association.

Saeed Mehrabi, Anand Krishnan, Sunghwan Sohn, Alexandra M. Roch, Heidi Schmidt, Joe Kesterson, Chris Beesley, Paul Dexter, C. Max Schmidt, Hongfang Liu, and Mathew Palakal. 2015. DEEPEN: A negation detection system for clinical text incorporating dependency relation into NegEx. Journal of Biomedical Informatics, 54:213–219.

Roser Morante. 2010. Descriptive analysis of negation cues in biomedical texts. In LREC, volume 2010, pages 1429–1436.

Sankaran Narayanan, Kaivalya Mannam, Pradeep Achan, Maneesha V Ramesh, P Venkat Rangan, and Sreeranga P Rajan. 2022. A contextual multitask neural approach to medication and adverse events identification from clinical text. Journal of Biomedical Informatics, 125:103960.

Naveen S. Pagad and N. Pradeep. 2022. Clinical Named Entity Recognition Methods: An Overview. In International Conference on Innovative Computing and Communications, pages 151–165, Singapore. Springer Singapore.

Serguei V Pakhomov, Alexander Ruggieri, and Christopher G Chute. 2002. Maximum entropy modeling for mining patient medication status from free text. In Proceedings of the AMIA Symposium, page 587. American Medical Informatics Association.

Catherine Plaisant, Richard Mushlin, Aaron Snyder, Jia Li, Dan Heller, and Ben Shneiderman. 2003. LifeLines: using visualization to enhance navigation and analysis of patient records. In The craft of information visualization, pages 308–312. Elsevier.

Eric G Poon, Barry Blumenfeld, Claus Hamann, Alexander Turchin, Erin Graydon-Baker, Patricia C McCarthy, John Poikonen, Perry Mar, Jeffrey L Schnipper, Robert K Hallisey, et al. 2006. Design and Implementation of an Application and Associated Services to Support Interdisciplinary Medication Reconciliation Efforts at an Integrated Healthcare Delivery Network. Journal of the American Medical Informatics Association, 13(6):581–592.

Giridhar Kaushik Ramachandran, Kevin Lybarger, Yaya Liu, Diwakar Mahajan, Jennifer J Liang, Ching-Huei Tsou, Meliha Yetisgen, and Özlem Uzuner. 2023. Extracting Medication Changes in Clinical Narratives using Pre-trained Language Models. Journal of Biomedical Informatics, page 104302.

Daniel Sanchez-Cisneros, Paloma Martínez, and Isabel Segura-Bedmar. 2013. Combining Dictionaries and Ontologies for Drug Name Recognition in Biomedical Texts. In Proceedings of the 7th International Workshop on Data and Text Mining in Biomedical Informatics, DTMBIO '13, pages 27—-30, New York, NY, USA. Association for Computing Machinery.

Isabel Segura-Bedmar, Paloma Martínez, and María Segura-Bedmar. 2008. Drug name recognition and classification in biomedical texts: A case study outlining approaches underpinning automated systems. Drug Discovery Today, 13(17):816–823.

Sunghwan Sohn, Sean P Murphy, James J Masanz, Jean-Pierre A Kocher, and Guergana K Savova. 2010. Classification of medication status change in clinical narratives. In AMIA Annual Symposium Proceedings, volume 2010, page 762. American Medical Informatics Association.

György Szarvas, Veronika Vincze, Richárd Farkas, and János Csirik. 2008. The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In Proceedings of the workshop on current trends in biomedical natural language processing, pages 38–45.

Alexander Turchin, Maria Shubina, Eugene Breydo, Merri L Pendergrass, and Jonathan S Einbinder. 2009. Comparison of information content of structured and narrative text data sources on the example of medication intensification. Journal of the American Medical Informatics Association, 16(3):362–370.

Özlem Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association, 18(5):552–556.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, et al. 2018. Clinical information extraction applications: a literature review. Journal of biomedical informatics, 77:34–49.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, pages 38–45.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. Advances in neural information processing systems, 32.

Xin Yu, Wenshen Hu, Sha Lu, Xiaoyan Sun, and Zhenming Yuan. 2019. BioBERT Based Named Entity Recognition in Electronic Medical Record. In 2019 10th International Conference on Information Technology in Medicine and Education, ITME, pages 49–52, Qingdao, Shandong, China. IEEE.

Improving Automatic KCD Coding: Introducing the KoDAK and an Optimized Tokenization Method for Korean Clinical Documents

Geunyeong Jeong[1] **Juoh Sun**[1] **Seokwon Jeong**[2] **Hyunjin Shin**[3,4] **Harksoo Kim**[1]
[1]Konkuk University [2]Kangwon National University
[3]Konkuk University Medical Center [4]Konkuk University School of Medicine
{jyjg7218, qssz1326}@konkuk.ac.kr nlpsw@kangwon.ac.kr
shineye@kuh.ac.kr nlpdrkim@konkuk.ac.kr

Abstract

International Classification of Diseases (ICD) coding is the task of assigning a patient's electronic health records into standardized codes, which is crucial for enhancing medical services and reducing healthcare costs. In Korea, automatic Korean Standard Classification of Diseases (KCD) coding has been hindered by limited resources, differences in ICD systems, and language-specific characteristics. Therefore, we construct the **Ko**rean **D**ataset for **A**utomatic **K**CD coding (**KoDAK**) by collecting and preprocessing Korean clinical documents. In addition, we propose a tokenization method optimized for Korean clinical documents. Our experiments show that our proposed method outperforms Korean Medical BERT (KM-BERT) in Macro-F1 performance by 0.14%p while using fewer model parameters, demonstrating its effectiveness in Korean clinical documents.

1 Introduction

International Classification of Diseases (ICD) coding is the assigning of standardized codes from the ICD system to patients' electronic health records. This is essential for standardizing information across medical institutions, and it serves as the foundation for the analysis of medical statistics. Traditionally, professional coders performed this task; however, this approach was costly and error-prone(O'malley et al., 2005; Lang, 2007). Consequently, recent studies have been exploring the application of deep learning in automatic ICD coding(Xie and Xing, 2018; Mullenbach et al., 2018; Zhang et al., 2022).

However, there are noticeable gaps in the research on automatic ICD coding across countries for various reasons. One reason is the difference in ICD systems between nations. Each country has a unique medical environment and requirements, with the result that many countries use modified ICD systems(Alharbi et al., 2019; Harrison et al., 2021; Yan et al., 2022). For instance, in Korea, the Korean Standard Classification of Diseases (KCD) was developed. Therefore, a specific model or algorithm may not always be effective and the development of models that consider each nation's ICD system is essential. Another reason for these disparities is differences in linguistic characteristics. Each language possesses its own grammar, syntax, and semantic structure, which affect medical terms and their expression. Hence, models that do not consider these characteristics may exhibit limited performance. Lastly, resource availability also plays an important role. Resource-rich countries use large-scale clinical data and systematic medical knowledge resources(Bodenreider, 2004) to accelerate automatic ICD coding research(Yuan et al., 2022). In contrast, in countries with limited data and resources, research progress may be slower or more constrained.

To bridge the disparities in research levels across countries and enhance Korea's constrained clinical natural language processing research environment, we propose a comprehensive approach: the development of a new dataset for automatic KCD coding and the implementation of a tokenization method tailored for Korean clinical documents. First, we addressed the lack of Korean datasets for automatic KCD coding by collecting and preprocessing clinical documents with assigned KCD codes from Korean clinical environments. Subsequently, we employed an optimized tokenization method to ensure that the automatic KCD coding model accurately captured the linguistic characteristics in the Korean clinical documents. Our contributions are as follows:

- To the best of our knowledge, our study is the first to focus on automatic KCD coding and to

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 96–101
July 14, 2023 ©2023 Association for Computational Linguistics

examine the important factors that need to be considered for its improvement.

- We construct a **Ko**rean **D**ataset for **A**utomatic **K**CD coding (**KoDAK**) using initial diagnostic records collected from Korean clinical environments, and we conduct a thorough statistical analysis of the data.
- We propose an optimized tokenization method that effectively captures the linguistic characteristics of Korean clinical documents.
- Through comparative experiments, we confirm that the proposed method shows significant improvements over the existing approach.

2 KoDAK

The most renowned dataset for automatic ICD coding is MIMIC-III(Johnson et al., 2016), an English dataset collected from the intensive care unit. Furthermore, datasets are available for countries with relatively abundant language resources, such as China(Cao et al., 2020) and Spain(Goeuriot et al., 2020). However, there is no such dataset in Korea, which makes conducting research challenging. To address this issue, we constructed the **Ko**rean **D**ataset for **A**utomated **K**CD coding (**KoDAK**) for automatic KCD coding. In this section, we describe the process of constructing the KoDAK.

2.1 Data Collection

In this study, we collected clinical records from Konkuk University Medical Center to create the KoDAK. The data collection process involved obtaining approval from the institutional review board and safeguarding patient confidentiality and anonymity. The collected clinical records consisted of text written by doctors describing the patient's initial diagnosis and symptoms. Similar to the CCHMC(Pestian et al., 2007) dataset, each record is labeled with the most appropriate single KCD code. The dataset encompasses 1,196,739 documents collected from 23 departments over 17 years, from 2005 to 2021. Records containing sensitive information from departments such as psychiatry, obstetrics, gynecology, and urology were excluded from the dataset.

2.2 Data Preprocessing

We preprocessed the collected clinical records to enhance data quality and the accuracy of the analysis. In the preprocessing step, we first corrected for spacing. Because documents are in a free-form style, spacing errors frequently occur. We used KoSpacing(Jeon), a Korean spacing correction library, to rectify the spacing errors. Second, we eliminated any embedded image links in the documents that did not offer meaningful information. Finally, we sorted the entire dataset by text length and removed the shortest 5% of samples because they likely contained insufficient information to determine the KCD codes accurately.

Clinical Note	persistent **sinus tarsi pain**. painful when walking, very painful even at night while sleeping. in the morning, when getting up and stepping on it, the pain is so intense it feels like collapsing. 지속적으로 **sinus tarsi pain**이 있음. 걸을 때 아프고, 밤에 잘 때도 너무 아프다. 아침에 일어나서 딛으면 주저앉을 것처럼 아프다.
KCD	S9200 (Fracture of calcaneus, closed)

Table 1: Example of clinical note and corresponding KCD code in the KoDAK, with the English translation of clinical note separated by a horizontal line. Bold text indicates English text that appears in the original clinical note.

2.3 Data Example and Statistics

Table 1 presents an example of clinical note and their corresponding KCD code from the completed dataset. As observed in the clinical notes, doctors often use a combination of Korean and English when describing symptoms because of the prevalence of English medical terminology. Consequently, the KoDAK, written by Korean medical professionals, contains many English words (Korean: 81%; English: 17%; Other: 2%).

The KoDAK comprises 8,862 KCD codes, which account for 49% of all the KCD codes. The dataset shows a long-tail distribution, with the top 20% of the most frequent KCD codes covering approximately 94% of the dataset. Moreover, the least frequent 1,894 KCD codes appeared only once in the dataset.

3 Automatic KCD Coding Approach

3.1 Tokenization

As illustrated in Table 1, English medical terms such as "sinus tarsi pain" are crucial for KCD coding and must be carefully considered during tokenization. However, the existing Korean medical language model, KM-BERT(Kim et al., 2022), does not specifically account for English medical terms, which results in a low proportion of English tokens in the vocabulary (Korean: 71%; English: 2%; Other: 27%). This is likely because the training data for KM-BERT are mostly in Korean, unlike the KoDAK, which has a higher proportion of English. Table 2 presents the tokenization results of the samples in KoDAK using both the proposed method and the KM-BERT tokenizer.

Tokenizer	Tokens
KM-BERT	2007, 코, 다, 친, 후, **a**, **n**, **os**, **m**, **ia**, **s**, **n**, **or**, **ing**, +, **a**, **p**, **ne**, **a**, 가끔
Ours	2, 0, 0, 7, 코, 다친, 후, **anosmia**, **snoring**, +, **apnea**, 가끔

Table 2: Comparison of tokenization results for clinical notes from the KoDAK using KM-BERT and our proposed method. Bold tokens indicate English medical terms.

As shown in Table 2, KM-BERT excessively splits crucial medical terms into smaller units. Such tokenization can impair a model's language comprehension and make it challenging to train the model effectively.

To address this issue, we propose an optimized tokenization strategy that considers the characteristics of the KoDAK. We applied morpheme-aware subword tokenization(Park et al., 2020) for Korean, whereas we tokenized English at the word level to preserve the meanings of essential medical terms and abbreviations. Other text types, such as numbers and special characters, were tokenized at the character level. Through this tokenization process, we built a vocabulary of 73,241 tokens that comprised 32,390 Korean, 38,659 English, and 45 other text tokens. Specifically, we determined the number of English tokens required to preserve as many crucial medical terms as possible to maintain coverage of over 93%.

3.2 Model

For automatic KCD coding, we utilized a model based on the transformer encoder architecture(Vaswani et al., 2017), which has been proven to be effective in various natural language processing tasks. Instead of initializing the token embeddings with random values, we used a separate Word2Vec(Mikolov et al., 2013) to ensure richer representations for each token. Word2Vec was trained using the skip-gram approach, and the results were used as the initial embeddings for the transformer. Our model comprises six encoder layers. To classify the final KCD codes, we follow BERT(Devlin et al., 2019)'s classification model training framework by feeding the [CLS] token representation into a linear layer.

To train the model, we employed the cross-entropy loss function(Good, 1952), which encourages the model to assign higher probabilities to correct KCD codes.

4 Experiments

4.1 Experimental Settings

The dataset was divided into training and evaluation datasets. For the KCD codes with only one instance (1,894 labels), the sample was assigned exclusively to the evaluation data, whereas for the KCD codes with more than one instance (5,275 labels), the sample was distributed between the training and evaluation. Consequently, the training data consisted of 1,130,942 samples, and the evaluation data consisted of 65,797 samples.

In this study, we used KM-BERT as a comparative model. KM-BERT is a language model for Korean medical natural language processing designed to alleviate the challenges in text analysis due to the agglutinative nature of the Korean language and complex medical terminology. KM-BERT was trained on a collection of Korean medical corpora using BERT's pre-training framework.

To evaluate the performance of the proposed method, we used Macro-F1 and Micro-F1, which are widely recognized metrics for evaluating classification models.

4.2 Experiment Results

Our proposed model for automatic KCD coding demonstrated remarkable performance, as shown in Table 4. Notably, our model outperforms the comparative model by achieving a 0.14%p improvement in Macro-F1, which showcases the strength of the proposed model in consistently enhancing performance across numerous labels, regardless of sample size. The results are especially noteworthy, considering our model has only 55% of the parameters of the comparative model, and it is not pre-trained.

	Clinical Notes	Ours	KM-BERT
Case 1	**accessory thumb left O) at birth V) congenital** patient visited for surgical treatment due to congenital polydactyly. **accessory thumb left O) at birth V) congenital** 상기 환아 선천적 다지증으로 수술적 치료 위하여 내원함	Q691	M2124
Case 2	2 days ago, after a fall, the patient experienced **retrograde amnesia** and visited the outpatient department of our hospital. 2일전 넘어진 후 **retrograde amnesia** 발생하여 본원 외래 내원함	R412	S0620

Table 3: Case study illustrating the improved KCD code predictions by our proposed model compared with KM-BERT, with the English translation of clinical notes separated by a horizontal line. Bold text indicates English text that appears in the original clinical note.

	Macro-F1	Micro-F1	Params
KM-BERT	8.59	**44.36**	105M
Ours	**8.73**	43.39	60M
w/o W	7.64	41.29	60M
w/o W & T	7.12	40.89	32M

Table 4: Performance comparison of our proposed model with KM-BERT and ablation study results (w/o W: without Word2Vec token embedding initialization; w/o T: using KM-BERT tokenizer instead of our proposed tokenizer).

Although our model showed a relatively lower performance in Micro-F1, it is important to consider that KM-BERT, the comparative model, was pre-trained on a large medical corpus. This allows it to leverage knowledge transfer to improve the performance of classes with a large number of samples. Despite this advantage, our proposed model remains highly competitive and offers a more efficient alternative, especially in terms of model size and the absence of pre-training requirements.

4.3 Ablation Study

We conducted an ablation analysis to assess the impact of Word2Vec and the proposed tokenization method on the performance of the model. Table 4 presents the results of the study.

When Word2Vec was not used (w/o W in Table 4), the performance declined across all evaluation metrics. This implies that incorporating Word2Vec enhances the model by offering richer token representations. In addition, we observed that using Word2Vec accelerated the model's convergence (best number of epochs: 18 with Word2Vec, 22 without). Removing both our proposed tokenization method and Word2Vec and employing the KM-BERT tokenizer instead (w/o W & T in Table 4), the performance deteriorates further across all evaluation metrics relative to solely removing Word2Vec. This finding underscores the proposed tokenization method positively influencing the model's performance.

4.4 Case study

We conducted a case study to understand better the improvements made to the proposed model. Table 3 lists the cases in which the proposed model accurately predicted the correct KCD codes.

In Table 3, Case 1 presents a situation where the correct KCD code is Q691 (accessory thumb). The proposed model precisely classified it as Q691 by considering the term "accessory thumb" in the clinical note. On the other hand, the comparison model misclassified it as M2124 (Flexion deformity, hand), a subclass of M21 (Other acquired deformities of limbs), despite the presence of words like "선천적 (congenital)", "congenital", and "at birth" in the note. In Case 2, the correct KCD code was R412 (retrograde amnesia). Our model accurately identified it using the phrase "retrograde amnesia" from the notes. However, the comparison model misclassified it as S0620 (diffuse brain injury, without open intracranial wound), a similar but different code. This demonstrates the effectiveness of the proposed tokenization method in capturing the meaning of English medical terms and helping the model better understand and interpret documents.

5 Conclusion and Future Work

In this study, we introduced the KoDAK as a resource for facilitating automatic KCD coding research in Korea, where the lack of suitable datasets has hindered such research. Furthermore, we proposed a tokenization method that effectively reflects the linguistic features of Korean clinical documents, thereby ensuring an accurate representation of crucial medical terms. Our approach outperformed KM-BERT, achieving a 0.14%p improvement in Macro-F1 while utilizing fewer parameters and without pre-training. In future research, we aim to address the unbalanced label distribution in the KoDAK and develop an enhanced pre-trained language model specifically designed for the Korean clinical field.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No.2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques)

References

Musaed Ali Alharbi, Godfrey Isouard, and Barry Tolchard. 2019. The development of icd adaptations and modifications as background to a potential saudi arabia's national version. *Global Journal of Health Science*.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Pengfei Cao, Chenwei Yan, Xiangling Fu, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong. 2020. Clinical-coder: Assigning interpretable ICD-10 codes to Chinese clinical notes. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 294–301, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lorraine Goeuriot, Hanna Suominen, Liadh Kelly, Antonio Miranda-Escalada, Martin Krallinger, Zhengyang Liu, Gabriella Pasi, Gabriela Gonzalez Saez, Marco Viviani, and Chenchen Xu. 2020. Overview of the clef ehealth evaluation lab 2020. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings 11*, pages 255–271. Springer.

Irving John Good. 1952. Rational decisions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 14(1):107–114.

James E Harrison, Stefanie Weber, Robert Jakob, and Christopher G Chute. 2021. Icd-11: an international classification of diseases for the twenty-first century. *BMC medical informatics and decision making*, 21(6):1–10.

Heewon Jeon. Kospacing: Automatic korean word spacing. https://github.com/haven-jeon/KoSpacing.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Yoojoong Kim, Jong-Ho Kim, Jeong Moon Lee, Moon Joung Jang, Yun Jin Yum, Seongtae Kim, Unsub Shin, Young-Min Kim, Hyung Joon Joo, and Sanghoun Song. 2022. A pre-trained bert for korean medical natural language processing. *Scientific Reports*, 12(1):1–10.

Dee Lang. 2007. Consultant report-natural language processing in the health care industry. *Cincinnati Children's Hospital Medical Center, Winter*, 6.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana. Association for Computational Linguistics.

Kimberly J O'malley, Karon F Cook, Matt D Price, Kimberly Raiford Wildes, John F Hurdle, and Carol M Ashton. 2005. Measuring diagnoses: Icd code accuracy. *Health services research*, 40(5p2):1620–1639.

Kyubyong Park, Joohong Lee, Seongbo Jang, and Dawoon Jung. 2020. An empirical study of tokenization strategies for various Korean NLP tasks. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 133–142, Suzhou, China. Association for Computational Linguistics.

John P. Pestian, Christopher Brew, Paweł Matykiewicz, D. J. Hovermale, Neil Johnson, K. Bretonnel Cohen, and Włodzisław Duch. 2007. A shared task involving multi-label classification of clinical free text. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, BioNLP '07, page 97–104, USA. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Pengtao Xie and Eric Xing. 2018. A neural architecture for automated ICD coding. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1066–1076, Melbourne, Australia. Association for Computational Linguistics.

Chenwei Yan, Xiangling Fu, Xien Liu, Yuanqiu Zhang, Yue Gao, Ji Wu, and Qiang Li. 2022. A survey of automated international classification of diseases coding: development, challenges, and applications. *Intelligent Medicine*, 2(3):161–173.

Zheng Yuan, Chuanqi Tan, and Songfang Huang. 2022. Code synonyms do matter: Multiple synonyms matching network for automatic ICD coding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 808–814, Dublin, Ireland. Association for Computational Linguistics.

Shurui Zhang, Bozheng Zhang, Fuxin Zhang, Bo Sang, and Wanchun Yang. 2022. Automatic ICD coding exploiting discourse structure and reconciled code embeddings. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2883–2891, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Who needs context? Classical techniques for Alzheimer's disease detection

Behrad TaghiBeyglou[1,2] and **Frank Rudzicz**[3,4,5]

[1]Institute of Biomedical Engineering, University of Toronto, Toronto, Canada
[2]KITE- Toronto Rehabilitation Institute, University Health Network, Toronto, Canada
behrad.taghibeyglou@mail.utoronto.ca
[3]Faculty of Computer Science, Dalhousie University, Halifax, Canada
[4]Department of Computer Science, University of Toronto, Toronto, Canada
[5]Vector Institute for Artificial Intelligence, Toronto, Canada
frank@dal.ca

Abstract

Natural language processing (NLP) has shown great potential for Alzheimer's disease (AD) detection, particularly due to the adverse effect of AD on spontaneous speech. The current body of literature has directed attention toward context-based models, especially Bidirectional Encoder Representations from Transformers (BERTs), owing to their exceptional abilities to integrate contextual information in a wide range of NLP tasks. This comes at the cost of added model opacity and computational requirements. Taking this into consideration, we propose a Word2Vec-based model for AD detection in 108 age- and sex-matched participants who were asked to describe the Cookie Theft picture. We also investigate the effectiveness of our model by fine-tuning BERT-based sequence classification models, as well as incorporating linguistic features. Our results demonstrate that our lightweight and easy-to-implement model outperforms some of the state-of-the-art models available in the literature, as well as BERT models.

1 Introduction

Alzheimer's disease (AD) is the most prevalent form of dementia, a neurodegenerative disease that impairs cognitive functioning and is increasingly common in our aging society (Luz et al., 2021; Ilias and Askounis, 2022). According to the World Health Organization, approximately 55 million people currently suffer from dementia, with this number expected to surge to 78 million and 139 million by 2030 and 2050, respectively (Ilias and Askounis, 2022). Symptoms of AD include (but are not limited to) memory decline, disorientation, confusion, and behavioural changes. Importantly, AD progression can lead to loss of independence which significantly impacts patients, their families, and society as a whole (Pappagari et al., 2021). Given that late-stage AD progression is inevitable, early detection of AD through cost-effective and scalable technologies is critical. While most clinical diagnoses of AD rely on neuroimaging, there is a critical need for more accessible and efficient methods of diagnosis.

Accessible evaluation methods for AD include cognitive tests such as the Mini-Mental Status Examination (MMSE) (Kurlowicz and Wallace, 1999) and the Montréal Cognitive Assessment (MoCA) (Nasreddine et al., 2003). However, these methods still require active integration with an expert, and their specificity in early-stage diagnosis is questionable. During the course of AD, patients experience a gradual deterioration of cognitive function and accordingly may face a loss of lexical-semantic skills, including anomia, reduced word comprehension, object naming problems, semantic paraphasia, and a reduction in vocabulary and verbal fluency (Mirheidari et al., 2018; Pan et al., 2021; Chen et al., 2021). Speech processing and, consequently, natural language processing (NLP) can therefore provide new precision medicine tools for AD diagnosis that deliver objective quantitative analyses and reliable proof, analysis, comparison, and circulation for faster diagnosis.

The Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) challenge of INTERSPEECH 2020 is a shared database developed to advance research into automatic AD detection based on spontaneous speech and transcripts (Luz et al., 2020). Participants in the challenge were tasked with describing the Cookie Theft picture in English, which is part of the Boston Diagnostic Aphasia Exam (Guo et al., 2021). The first set of the ADReSS 2020 database comprises speech recordings and CLAN-annotated transcripts of 54 AD patients and 54 sex- and age-matched controls.

Various groups have worked with the ADReSS dataset, approaching the problem from different perspectives and leveraging available information. These studies typically combined speech processing and linguistic feature extraction or NLP-based

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 102–107
July 14, 2023 ©2023 Association for Computational Linguistics

fine-tuning. The literature on speech processing mostly focused on zero-crossing rate, spectral bandwidth, roll-off, and centroids of audio recordings, as well as active data representation cluster-based feature extraction methods including the emobase (Eyben et al., 2010), ComParE (Eyben et al., 2013), and Multi-Resolution Cochleagram (MRCG) (Chen et al., 2014) feature sets. Meanwhile, linguistic features have extracted lexical richness, the proportion of various PoS tags, utterance duration, total utterances, type-token ratio, open-closed class word ratio, and similarity between consecutive utterances. NLP-based methods have comprised from-scratch training or fine-tuning context-based models, such as bidirectional long short-term memory (bi-LSTM) (Cummins et al., 2020), bi-directional Hierarchical Attention Network (bi-HANN) (Cummins et al., 2020), Convolutional Recurrent Neural Network (CRNN) (Koo et al., 2020), and Bidirectional Encoder Representations from Transformer (BERT) (Balagopalan et al., 2020). Despite excellent performance compared to baseline methods (Luz et al., 2020), the complexity of these methodologies and the need to implement them on high-memory GPUs highlights the need to explore simpler methodologies that can ensure ease and performance in AD detection.

In this paper, we present a novel approach for detecting AD in the first set of ADReSS dataset by integrating a new Word2Vec-based model and dimension reduction method. We not only implement and compare top-cited and recent state-of-the-art models on the same dataset, but also demonstrate that our approach outperforms these models. Our proposed approach is simple, easy to implement, and highly accurate.

2 Methodology

2.1 Other models

In order to evaluate the performance of our proposed language processing model, we have considered several publicly available models for comparison including:

- **Linguistic-Based Features (LBF)**: In this study, we utilized the CLAN package to extract 34 linguistic-based features (LBFs) from transcripts, including duration, total utterances, mean length of utterance (MLU), type-token ratio, open-closed class word ratio, and percentages of 9 parts of speech. We also incorporated demographic information such as age and sex. To identify the most informative features for classification, we performed correlation and variance analyses on the extracted features using the FeatureWiz package (AutoViML, 2020). We set a correlation threshold of 0.6 and repeated the analyses 5 times with random seeds over all samples. We then selected the top 5 features that appeared in at least 3 iterations for further classification.

- **BERT Models**: Since BERT models have shown promising performance in different applications of NLP, in this study we leveraged some of BERT-based architectures with a maximum length of 512 tokens as a reference for our model. We tested three versions of uncased base BERT (Devlin et al., 2018): one with no extension in the last layers, called *baseBERT1*, another with two fully connected layers at the end ($768 \rightarrow 64$ and $64 \rightarrow 1$), called *baseBERT2*, and the last one with three fully connected layers ($768 \rightarrow 128$, $128 \rightarrow 16$, and $16 \rightarrow 1$), called *baseBERT3*. For *baseBERT2*, we varied the training epochs between 3 and 5. Additionally, we tested *BioCLinical BERT* (Alsentzer et al., 2019) with a batch size of 4 and 3 epochs, *DistilBERT* (Sanh et al., 2019) with a batch size of 4 and 3 epochs, and *BioMed-RoBERTa-based* (Gururangan et al., 2020) with a batch size of 4 and 3 epochs. We used binary cross-entropy as the loss function for all models and AdamW (Adam with weight decay) (Loshchilov and Hutter, 2017) as the optimizer with a learning rate of 2×10^{-5}. To address potential issues with local optima, we applied a linear warm-up scheduler. Each transcript is classified as AD if the average of the probabilities (after the sigmoid layer) over all sentences in the transcript is greater than or equal to 0.5; otherwise, it is classified as control.

2.2 Pre-processing

To preprocess the data for our proposed model, we have neglected the first four sentences of each transcript, as the initial speaker is typically a member of the data collection team. Additionally, stop words were removed from each sentence using the Gensim library (Řehřek et al., 2011).

2.3 Proposed model

In this study, we used Wikipedia2Vec (Yamada et al., 2018), a tool that generates embeddings (or vector representations) of words and entities from Wikipedia, to convert tokens to vector embeddings. We used the skip-gram strategy for training, and the embedding dimension of the model was set to 500. We denote this model as $W2V$ throughout this paper. Suppose that each participant's transcript consists of N_k sentences, each comprising m words, where m varies from 1 to M_k (the maximum length among all sentences in the kth transcript). We input each word $\langle w_{i,k} \rangle$ into the $W2V$ model ($W2V(\langle w_{i,k} \rangle)$), which outputs the corresponding embedded vector $\mathbf{x}_{i,k} \in \mathcal{R}^{500}$. All embeddings of the kth transcript form the set X_k. We standardized each 500-dimensional vector across all embeddings of each subject using the following formula:

$$\mathbf{y}_k = \frac{\texttt{med}(X_k)}{\texttt{std}(X_k)}, \tag{1}$$

where med is the median operator applied to each dimension independently, std is the standard deviation of embeddings, and $\mathbf{y}_k$ denotes the standardized vector for the kth participant. So far, we developed the first framework and leveraged the previously introduced feature selection method by iteratively applying FeatureWiz five times. We then selected features that were chosen at least three times during the process to identify the most informative dimensions for AD detection. This feature selection procedure reduced the dimension from 500 to 64. We refer to this first framework as *model 1*, and Figure 1 illustrates the process. To further enhance our analysis, we concatenate linguistics-based features from the previous section with W2V-based feature vectors and apply feature selection in a similar manner to *model 1*. This second framework, called *model 2*, resulted in the selection of 86 features (out of 537 features). Prior to inputting the features into the classifiers of each model, the zero-mean-unit-variance standardization technique is applied to normalize the features.

2.4 Evaluation and Metrics

All results presented in this study were obtained using the leave-one-subject-out (LOSO) cross-validation technique to evaluate the generalizability of the models. Thus, a total of 104 models were trained per architecture/classifier. For each model, accuracy, sensitivity, specificity, and F1 were reported as performance metrics. For the feature-based models, such as linguist-based features and our proposed frameworks, we employed various classifiers including logistic regression (LR), decision tree (DT), linear and Nu-support vector classification (SVC), linear and quadratic discriminant analysis (LDA and QDA), Gaussian naive Bayes (GNB), extreme gradient boosting (XGBoost), adaptive boosting (AdaBoost), and extra trees classifier.

3 Results

3.1 Other models

We investigated different BERT models for AD classification, and the results are presented in Table 1. As expected, the performance of *Bio-Clinical BERT* and *DistilBERT* models were comparable; however, *Bio-Clinical BERT* showed superior sensitivity and was chosen as the best BERT model in this study. Additionally, as demonstrated in Table 2, integrating linguistic-based features with feature selection and a combination of classifiers achieved an accuracy of 0.81 in AD detection.

Model	E:BS	AC	SP	SE	F1
baseBERT1	3:4	0.80	0.89	0.7	0.78
baseBERT2	3:4	0.79	0.81	0.76	0.78
baseBERT2	5:4	0.79	0.93	0.65	0.77
baseBERT3	3:4	0.78	0.90	0.67	0.77
Bio-CLinical BERT	**3:4**	**0.84**	**0.85**	**0.83**	**0.84**
DistilBERT	3:4	0.84	0.87	0.81	0.84
BioMed-RoBERTa-based	3:4	0.81	0.87	0.76	0.81

Table 1: LOSO performance of other BERT-based models. "E" denotes the number of epochs, "BS" denotes the batch size, and "AC", "SP", and "SE" represent accuracy, specificity, and sensitivity, respectively.

3.2 Proposed frameworks

The performance of our proposed frameworks is presented in Table 3. The best performance was achieved by *model 2* with the help of the GNB classifier, which obtained an accuracy of 0.90. On the other hand, the best performance of *model 1* was achieved by the ExtraTrees classifier.

3.3 Comparison with previous literature

Table 4 compares our proposed model with the existing models in the literature as well as the ones explored in this paper. Our model achieved significantly higher performance, including a 3% improvement in accuracy and an 8% improvement in sensitivity compared to one of the BERT-based

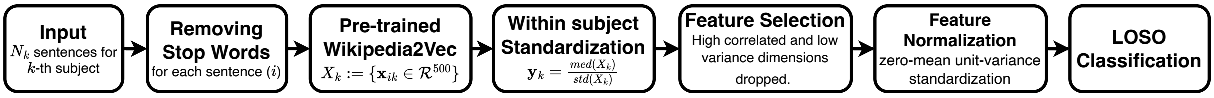


Figure 1: Proposed framework for AD classification.

Classifier	AC	SP	SE	F1
LR	0.76	0.81	0.70	0.75
DT	0.69	0.74	0.63	0.68
Linear SVC	0.76	0.80	0.72	0.76
Nu-SVC	0.81	0.83	0.78	0.80
LDA	0.79	**0.89**	0.69	0.78
QDA	**0.81**	0.87	**0.76**	**0.81**
GNB	0.78	0.87	0.69	0.77
XGBoost	0.71	0.70	0.72	0.71
AdaBoost	0.74	0.76	0.72	0.74
ExtraTrees	0.72	0.76	0.69	0.72

Table 2: LOSO performance of the linguist feature-based model, in combination with the proposed feature selection technique.

Classifier	Model	AC	SP	SE	F1
LR	*model 1*	0.74	0.87	0.81	0.84
	model 2	0.74	0.89	0.80	0.84
DT	*model 1*	0.76	0.80	0.72	0.76
	model 2	0.56	0.54	0.57	0.55
Linear SVC	*model 1*	0.81	0.85	0.78	0.81
	model 2	0.80	0.85	0.74	0.79
Nu-SVC	*model 1*	0.85	0.85	0.85	0.85
	model 2	**0.90**	**0.91**	0.89	**0.9**
LDA	*model 1*	0.73	0.74	0.72	0.73
	model 2	0.66	0.69	0.63	0.66
QDA	*model 1*	0.60	0.63	0.57	0.6
	model 2	0.44	0.33	0.56	0.42
GNB	*model 1*	0.87	0.87	0.87	0.87
	model 2	**0.90**	0.89	**0.91**	**0.9**
XGBoost	*model 1*	0.77	0.76	0.78	0.77
	model 2	0.78	0.78	0.78	0.78
AdaBoost	*model 1*	0.81	0.78	0.85	0.81
	model 2	0.82	0.81	0.83	0.82
ExtraTrees	*model 1*	0.88	0.89	0.87	0.88
	model 2	0.89	0.91	0.87	0.89

Table 3: LOSO performance of the linguist feature-based model, in combination with the proposed feature selection technique.

models on the same dataset (Balagopalan et al., 2020, 2021). It is worth noting that our proposed model also outperformed the baseline linguistic model introduced in the ADReSS challenge.

Model	AC	SP	SE	F1
Bio-CLinical BERT	0.84	0.85	0.83	0.84
Best Linguist-based features	0.81	0.87	0.76	0.81
BERT and SVM (Balagopalan et al., 2020, 2021)	0.87	**0.91**	0.83	0.87
Gated LSTM on acoustic and lexical (Rohanian et al., 2021)	0.77	-	-	-
Baseline Linguistic (Luz et al., 2020)	0.77	0.77	0.76	0.77
Best proposed model	**0.90**	0.89	**0.91**	**0.9**

Table 4: LOSO performance comparison of the best proposed model and explored models with some existing models on the same dataset. The best linguist-based features model uses QDA classifier with linguist-based features, and the best proposed model is our proposed *model 2* with GNB classifier.

4 Discussion

By mapping each word into a 500-dimensional space where words with similar context are closer together, the proposed model can identify when all words in a transcript are focused on the same topic with minimal deviations. Coupled with the suggested standardization method, the results demonstrate a significant difference in performance between the proposed model and the only linguist-based model, which prioritizes utterances, pauses, and interactions between text and speech. The BERT models explored in this study are relatively massive and require significant computational resources, and training them requires delicate hyperparameter optimization. In this study, we followed the BERT authors' recommendations to keep the model's trainability on an Nvidia RTX 3080 GPU and to avoid changing the weights of the model by selecting smaller epoch numbers.

5 Conclusions

In this study, we introduced a word2vec-based model that combines pre-trained Wikipedia embeddings with linguistic features. We also employed correlation-based feature selection to reduce the dimensionality of the embeddings. The results demonstrated that our proposed model outperformed existing models on the same dataset. However, as BERT models offer diverse applicability, a potential future direction is to incorporate feature maps extracted from the hidden states of these networks to enhance the performance of our model.

Acknowledgment

This research is supported by AI4PH, a Health Research Training Platform funded by the Canadian Institutes of Health Research (CIHR).

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

AutoViML. 2020. featurewiz. `https://github.com/AutoViML/featurewiz`.

Aparna Balagopalan, Benjamin Eyre, Jessica Robin, Frank Rudzicz, and Jekaterina Novikova. 2021. Comparing pre-trained and feature-based models for prediction of alzheimer's disease based on speech. *Frontiers in aging neuroscience*, 13:635945.

Aparna Balagopalan, Benjamin Eyre, Frank Rudzicz, and Jekaterina Novikova. 2020. To bert or not to bert: comparing speech and language-based approaches for alzheimer's disease detection. *arXiv preprint arXiv:2008.01551*.

Jitong Chen, Yuxuan Wang, and DeLiang Wang. 2014. A feature study for classification-based speech separation at low signal-to-noise ratios. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12):1993–2002.

Jun Chen, Jieping Ye, Fengyi Tang, and Jiayu Zhou. 2021. Automatic detection of alzheimer's disease using spontaneous speech only. In *Interspeech*, volume 2021, page 3830. NIH Public Access.

Nicholas Cummins, Yilin Pan, Zhao Ren, Julian Fritsch, Venkata Srikanth Nallanthighal, Heidi Christensen, Daniel Blackburn, Björn W Schuller, Mathew Magimai-Doss, Helmer Strik, et al. 2020. A comparison of acoustic and linguistics methodologies for alzheimer's dementia recognition. In *Interspeech 2020*, pages 2182–2186. ISCA-International Speech Communication Association.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 835–838.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462.

Yue Guo, Changye Li, Carol Roan, Serguei Pakhomov, and Trevor Cohen. 2021. Crossing the "cookie theft" corpus chasm: applying what bert learns from outside data to the adress challenge dementia detection task. *Frontiers in Computer Science*, 3:642517.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*.

Loukas Ilias and Dimitris Askounis. 2022. Multimodal deep learning models for detecting dementia from speech and transcripts. *Frontiers in Aging Neuroscience*, 14.

Junghyun Koo, Jie Hwan Lee, Jaewoo Pyo, Yujin Jo, and Kyogu Lee. 2020. Exploiting multimodal features from pre-trained networks for alzheimer's dementia recognition. *arXiv preprint arXiv:2009.04070*.

Lenore Kurlowicz and Meredith Wallace. 1999. The mini-mental state examination (mmse). *Journal of gerontological nursing*, 25(5):8–9.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2020. Alzheimer's dementia recognition through spontaneous speech: The adress challenge. *arXiv preprint arXiv:2004.06833*.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2021. Detecting cognitive decline using speech only: The adresso challenge. *arXiv preprint arXiv:2104.09356*.

Bahman Mirheidari, Daniel Blackburn, Traci Walker, Annalena Venneri, Markus Reuber, and Heidi Christensen. 2018. Detecting signs of dementia using word vector representations. In *Interspeech*, pages 1893–1897.

Ziad S Nasreddine, Natalie A Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L Cummings, and Howard Chertkow. 2003. Montreal cognitive assessment. *The American Journal of Geriatric Psychiatry*.

Yilin Pan, Bahman Mirheidari, Jennifer M Harris, Jennifer C Thompson, Matthew Jones, Julie S Snowden, Daniel Blackburn, and Heidi Christensen. 2021. Using the outputs of different automatic speech recognition paradigms for acoustic-and bert-based alzheimer's dementia detection through spontaneous speech. In *Interspeech*, pages 3810–3814.

Raghavendra Pappagari, Jaejin Cho, Sonal Joshi, Laureano Moro-Velázquez, Piotr Zelasko, Jesús Villalba, and Najim Dehak. 2021. Automatic detection and assessment of alzheimer disease using speech and language technologies in low-resource scenarios. In *Interspeech*, pages 3825–3829.

Radim Řehřek, Petr Sojka, et al. 2011. Gensim—statistical semantics in python. *Retrieved from genism. org*.

Morteza Rohanian, Julian Hough, and Matthew Purver. 2021. Multi-modal fusion with gating using audio, lexical and disfluency features for alzheimer's dementia recognition from spontaneous speech. *arXiv preprint arXiv:2106.09668*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. 2018. Wikipedia2vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from wikipedia. *arXiv preprint arXiv:1812.06280*.

Knowledge Injection for Disease Names in Logical Inference between Japanese Clinical Texts

Natsuki Murakami[1] **Mana Ishida**[1] **Yuta Takahashi**[1]
Hitomi Yanaka[2] **Daisuke Bekki**[1]
[1] Ochanomizu University
[2] the University of Tokyo
{murakami.natsuki,ishida.mana,takahashi.yuta,bekki}@is.ocha.ac.jp
hyanaka@is.s.u-tokyo.ac.jp

Abstract

In the medical field, there are many clinical texts such as electronic medical records, and research on Japanese natural language processing using these texts has been conducted. One such research involves Recognizing Textual Entailment (RTE) in clinical texts using a semantic analysis and logical inference system, ccg2lambda. However, it is difficult for existing inference systems to correctly determine the entailment relations, if the input sentence contains medical domain specific paraphrases such as disease names. In this study, we propose a method to supplement the equivalence relations of disease names as axioms by identifying candidates for paraphrases that lack in theorem proving. Candidates of paraphrases are identified by using a model for the NER task for disease names and a disease name dictionary. We also construct an inference test set that requires knowledge injection of disease names and evaluate our inference system. Experiments showed that our inference system was able to correctly infer for 106 out of 149 inference test sets.

1 Introduction

In the medical field, there are many electronic texts, such as image detections and electronic medical records, and using such texts becomes more active in research on natural language processing (NLP) in Japanese (Aramaki et al., 2018; Doi et al., 2011). However, many of these studies utilize machine learning approaches, and it is argued that the machine learning approaches have problems in dealing with challenging linguistic phenomena such as negation and quantification. Logic-based approaches have been proposed to perform systematic inferences involving these challenging linguistic phenomena, and one task of these inferences can be referred to as recognizing textual entailment (RTE). RTE is the task of determining whether a hypothesis sentence H can be inferred from a premise sentence T. For example, the following example illustrates a case where T entails H.

T : Some patients are given Loxoprofen.

H : There are patients who are given headache medicines.

One such effort at RTE between clinical texts is the logical inference system in the medical domain proposed by Ishida et al. (2022). Compounds words are often appeared in Japanese clinical texts, and methods to analyze linguistic phenomena in compound words were desirable. The system is an inference system based on *ccg2lambda* (Martínez-Gómez et al., 2016), a semantic analysis and logical inference system: the system extends ccg2lambda to enable an analysis of compound words that are frequently appeared in clinical texts. However, this system fails to perform inference when there are differences in the notation of disease names in medical texts. For example, the disease name "Deep vein thrombosis" has multiple paraphrases, such as "DVT" and "Homann's sign", and different clinical texts use different phrases that refer to the same disease name. The premise sentence is "The patient developed Homann's sign." and the hypothesis sentence is "The patient developed deep vein thrombosis.", then empirically the premise sentence implies the hypothesis sentence. To show this entailment relation, the knowledge that "Homann's sign means deep vein thrombosis" must be supplemented in the theorem prover.

In this study, we propose our logical inference system with knowledge injection in the medical field. We identify candidates of paraphrase of disease names that are necessary for theorem proving by a named entity recognition (NER) model for disease names and the Japanese disease name dictionary called J-MeDic (Ito et al., 2018). By generating axioms according to the combination of compound word semantic tags assigned to the identified

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 108–117
July 14, 2023 ©2023 Association for Computational Linguistics

Surface form	ICD10	Standard disease name	Reliability	Frequency
sinbu-jomyaku-kessen-syo deep-vein-thrombosis 深部静脈血栓症	I802	deep vein thrombosis	S	85-90%
kasi-sinbu-jomyaku-kessen lower extremity-deep-vein-thrombosis 下肢深部静脈血栓症	I802	deep vein thrombosis	A	5-10%
DVT DVT	I802	deep vein thrombosis	C	90-95%
DVT-tyoukou DVT-sign DVT徴候	I802	deep vein thrombosis	C	60-65%
homanzu-tyoukou Homann's-sign ホーマンズ徴候	I802	deep vein thrombosis	C	25-30%

Table 1: An example of the J-MeDic with the standard disease name column listing "Deep Vein Thrombosis."

disease names, we inject paraphrase knowledge of disease names as axioms. We also evaluate the effectiveness of our inference system by constructing an inference test set that requires knowledge injection of disease names.

2 Background

2.1 Inference systems for clinical texts

There has been growing progress in research on neural network models for RTE with large-scale datasets using crowdsourcing such as SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018). However, neural network models for RTE generally require a large amount of training data, and inference is black-box, which makes it difficult to correct errors. Therefore, there is a need to develop RTE systems that are applicable to fields such as medical field, where it is difficult to prepare a large amount of training data and transparency is necessary.

The logical inference system ccg2lambda has the advantage of being able to judge entailment relations between sentences without using a large amount of training data and being easy to personalize and modify processing. However, the original ccg2lambda fails to analyze the semantic relations within compound words because they are treated as one word.

Ishida et al. (2022) addressed this problem by adding a compound word analysis module to ccg2lambda. This module extracts compound words from the Combinatory Categorial Grammar (CCG) (Steedman, 2000; Bekki, 2010) syntactic structures obtained by the CCG parser of ccg2lambda and assigns compound word semantic tags that represent semantic relations within compound words, using a compound word semantic tagger. Based on syntactic structures, semantic tags, and lambda calculus, the semantic representation was derived by taking into account the semantic relations within the compound words, and inference between clinical texts containing compound words was realized.

2.2 Related studies on axiom injection

As for related studies on axiom injection of logical inference systems, including ccg2lambda, Martínez-Gómez et al. (2017) proposed word axiom injection by using lexical knowledge. Hokazono et al. (2018) used this word abduction mechanism to inject word knowledge specific in the financial texts as the lexical knowledge. However, these previous studies were limited to handle word-to-word relations in natural deduction proofs. Yanaka et al. (2018) proposed a method for paraphrase detection by natural deduction proofs of semantic relations between sentence pairs to complement phrasal knowledge. In this study, we propose how to detect phrasal knowledge of disease names necessary for proving entailment relations between clinical texts and inject the knowledge into logical inference.

2.3 J-MeDic

In this study, we use J-MeDic to inject disease name knowledge into logical inference. J-MeDic is a Japanese dataset that extensively extracts

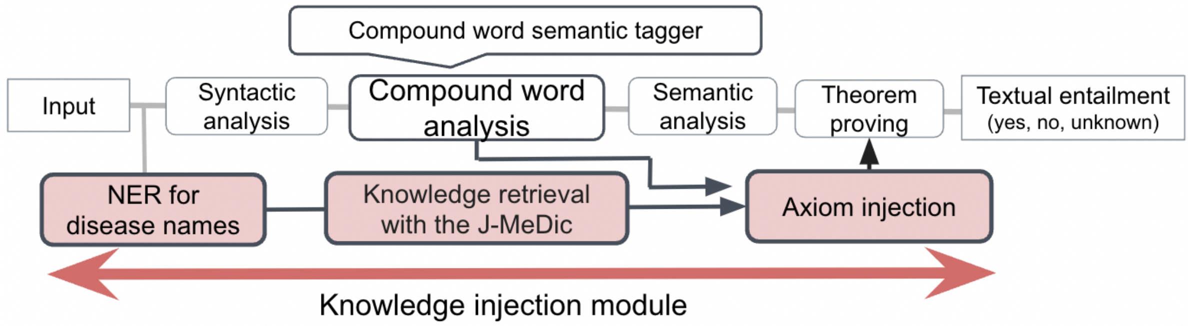


Figure 1: The overview of the proposed system.

words related to symptoms and disease names from progress records and discharge summaries recorded in electronic medical records by medical professionals. The dataset contains not only the formal names of diseases but also abbreviations and English names. The dataset covers 362,866 disease names in total. Table 1 provides an example where the standard disease name column in J-MeDic is "Deep vein thrombosis." J-MeDic records include the surface forms of the disease name, their pronunciation, ICD-10 codes, standard disease names, levels of reliability, and levels of frequency.

2.4 Related studies on NER in the medical domain

There are related studies on NER for disease names, such as the work by Goino and Hamagami (2021) and MedNER (Nishiyama et al., 2022). In Goino-Hamagami's work, NER and modality estimation of medical conditions such as disease names and symptoms along with their modalities (five types) were performed using BERT and CRF. Experiments were conducted using three BERT models released by Tohoku University (tohoku-charBERT, tohoku-BERT, tohoku-wwm-BERT) and UTH-BERT[1], a Japanese BERT model pretrained on medical documents by the University of Tokyo. MedNER[2] is a tool for the NER for disease names using word embeddings of BERT. MedNER follows J-MeDic and also performs modality estimation.

3 System Overview

The overview of our proposed system is shown in Figure 1. We provide our inference system by extending Ishida et al. (2022)'s inference system. The system consists of syntactic analysis, compound word analysis, semantic analysis, and theorem proving. In this study, we add a knowledge injection module in the medical domain to the previous system.

In the knowledge injection module, we first apply an NER model based on a pretrained language model to extract disease names from the input sentence. For the extracted disease names, we perform an exact match search for the surface form column of disease names in J-MeDic. If the disease name and the surface form match, we inject the knowledge about the disease name in J-MeDic as an axiom to the automated theorem prover Coq (Bertot and Castéran, 2013). Since the additional axioms vary depending on the semantic relations between morphemes, we check the semantic tags of compound words assigned by the compound word semantic tagger, derive the axiom using the semantic tag and the knowledge of J-MeDic, and then inject the knowledge. We describe the details to provide an NER model in Section 4 and Subsection 6.1 and the details of axiom injection in Section 5.

4 Building an NER dataset for disease names

To train the model for the NER task for disease names in our inference system, we constructed a dataset newly for NER using clinical texts. We use a corpus of case reports, J-MedStd-CR[3], which was extracted through OCR from case report papers in PDF format that are openly accessible on J-Stage. In this study, we manually annotated the appearances of disease names from J-MeDic in the 2,626 sentences of the J-MedStd-CR corpus.

[1] https://ai-health.m.u-tokyo.ac.jp/home/research/uth-bert
[2] https://github.com/sociocom/MedNER-J
[3] https://sociocom.naist.jp/medtxt/cr/

Tag	Type	Example
EN	Entity	DVT (Deep vein thrombosis)
M_EN	Modifying words	ryuukisei-byouhen elevated-lesion 隆起性 病変 (Elevated lesions)
PA	Body part	nou-kousoku brain-infarction 脳 梗塞 (Brain infarction)
GA	Nominative	kyousui-tyoryuu pleural effusion-retention 胸水 貯留 (Pleural effusion)
WO	Accusative	kotuzui-yokusei bone marrow-suppression 骨髄 抑制 (Myelosuppression)
NI	Dative	kotu-ten'i bone-metastasis 骨 転移 (Bone metastasis)
EV	Event	kyousui-tyoryuu pleural effusion-suppression 胸水 貯留 (Pleural effusion)

Table 2: Examples of semantic tags. Underlined parts correspond to each tags.

Combination	Number	Example
EN	276	syuryuu 腫瘤 (Neoplasm)
M_EN EN	145	ryuukisei-byouhen elevated-lesion 隆起性 病変 (Elevated lesions)
PA EN	45	nou-kousoku brain-infarction (Brain infarction) 脳 梗塞
M_EN M_EN EN	22	genpatusei-tanzyusei- primary-biliary- kankouhen liver cirrhosis 原発性 胆汁性 肝硬変 (Primary biliary cholangitis)
M_EN PA EN	22	ten'isei-kan-syuyou metastatic-liver-neoplasm 転移性 肝 腫瘍 (Metastatic liver tumor)
Others	141	kyousui-choryuu pleural effusion-retention 転移性 肝 腫瘍 GA EV (Metastatic liver tumor)

Table 3: Combinations of semantic tags in disease names.

5 Axiom Injection

In the proposed method, after identifying candidate paraphrases of the disease names, the knowledge of paraphrases is injected as axioms. In axiom injection, it is necessary to generate the knowledge about paraphrases of disease names as axioms according to the semantic representations in ccg2lambda. For example, when the premise sentence is (1) and the hypothesis sentence is (2), the underlined disease name's semantic tag in (1) is "EN" because "PE" is one word and entity. The semantic representation in this case is (1a). Semantic tags for the disease name underlined in (2) are "PA EN" because "pulmonary" in " pulmenary embolism" is a body part and "embolism" is an entity. The semantic representation is (2a). $\mathsf{PartOf}(e_1, e_2)$ indicates $\mathsf{hai}(e_2)$ is a body part for $\mathsf{sokusensyo}(e_1)$. As in (1a) and (2a), semantic representations differ depending on the semantic relations between morphemes within the compound words. In this study, we realize axiom injection by defining axioms generated through the combination of semantic tag assigned to both the surface form and standard disease names.

(1) PE ga zouaku-si-teita
PE NOM worsen-EUPH-PST
PEが増悪していた。
PE worsened.

a. $\exists e_1(\mathsf{PE}(e_1))$

(2) hai-sokusensyo ga zouaku-si-teita
pulmonary-embolism NOM worsen-EUPH-PST
肺塞栓症が増悪していた。
Pulmonary embolism worsend.

a. $\exists e_2(\mathsf{hai}(e_2) \wedge \exists e_1(\mathsf{sokusensyo}(e_1) \wedge \mathsf{PartOf}(e_1, e_2)))$

5.1 Trends in combinations of semantic tags in disease names

Table 2 shows some of the semantic tag assigned by Ishida et al. (2022)'s compound word semantic tagger. We investigated the composition of semantic tag to generate axioms based on the combination of semantic tag assigned to disease names. We applied the compound word semantic tagger to assign semantic tag to 651 disease names and their standard disease names extracted from randomly selected sentences containing disease names in clinical texts of the J-MedStd-CR corpus. The compound word semantic tagger is based on BiLSTM and BERT models. BERT model is the model released by

Axiom			Exapmles of disease name			Semantic tags of disease name		
Surface form	⇒	Disease name	Surface form	⇒	Disease name	Surface form	⇒	Disease name
∗	⇒	∗	hokou-hunou walking-impossibility 歩行 不能	⇒ (Walking difficulty)	hokou-konnan walking-difficulty 歩行 困難	EV EN	⇒	EV EN
M_EN ∗	⇒	∗	mansei-gisei-tyouheisokusuou chronic-pseudo-intestinal obstruction 慢性 偽性 腸閉塞症	⇒ (Pseudoileus)	gisei-ireusu pseudo-ileus 偽性 イレウス	M_EN M_EN EN	⇒	M_EN EN
PA ∗	⇒	∗	nou-kekkan'en brain-vasculitis 脳 血管炎	⇒ (Cerebral arteritis)	noudoumyakuen brain arteritis 脳動脈炎	PA EN	⇒	EN
PA ∗	⇒	M_EN ∗	ketsuryu-syougai bloodstream-disorder 血流 障害	⇒ (Peripheral circulatory disturbance)	massyojunkan-syogai peripheral circulation-disorder 末梢循環 障害	PA EN	⇒	M_EN EN
M_EN ∗	⇒	PA ∗	bimansei-shikiso-tintyaku diffuse-pigment-deposition びまん性 色素 沈着	⇒ (Skin pigmentation)	hihu-sikiso-tintyaku skin-pigment-deposition 皮膚 色素 沈着	M_EN GA EV	⇒	PA GA EV
GA EV	⇒	WO EV	kettin-kousin blood sedimentation-accentuation 血沈 亢進	⇒ (Erythrocyte sedimentation rate acceleration)	sekitin-kousin erythrocyte sedimentation-accentuation 赤沈 亢進	GA EV	⇒	WO EV
EN	⇒	M_EN+ EN	ＰＢＣ	⇒ (Primary biliary cholangitis)	genpatusei-tanzyuusei-kankouhen primary-biliary-liver cirrhosis 原発性 胆汁性 肝硬変	EN	⇒	M_EN M_EN EN
EN	⇒	PA EN	ＰＥ	⇒ (Pulmonary embolism)	hai-sokusensyo lung-embolism 肺 塞栓症	EN	⇒	PA EN
EV	⇒	EN	kansen'ika become liver fibrosis 肝繊維化	⇒ (Liver fibrosis)	kansen'isuu liver fibrosis 肝繊維腫	EV	⇒	EN

Table 4: Combination examples of semantic tags in axiom injection. ∗ indicates that the combination of tags assigned to the surface form and the standard disease name are the same. M_EN+ indicates that the M_EN tag appears one or more times. ⇒ indicates entailment relations.

Tohoku University[4] that was trained on Japanese Wikipedia data, and the tokenizers are MeCab and WordPiece. The top 5 combinations of semantic tags assigned to disease names are shown in Table 3. We perform axiom injection according to these combinations.

5.2 Axiom injection based on semantic tags

Table 4 shows the combination examples of semantic tag for the disease names in axiom injection. The asterisk ∗ indicates that the combinations of semantic tags are the same for both the surface form and the standard disease name. For example, the first row "∗ ⇒ ∗" indicates that the same semantic tags are assigned, such as "EV EN ⇒ EV EN" for "walking-impossibility ⇒ walking-difficulty". Similarly, the second row "M_EN ∗ ⇒ ∗" indicates that the semantic tags "M_EN EN" are assigned to the phrase "pseudo-intestinal obstruction" included in "chronic-pseudo-intestinal obstruction", and they match the tags assigned to "pseudo-ileus". "M_EN+" indicates that there are one or more "M_EN" tag present. As in the example where the semantic tags "EN ⇒ M_EN M_EN EN" are assigned to "PBC ⇒ primary-biliary-liver cirrhosis", even when there are multiple repetitions of the M_EN tag in the semantic tags assigned to the standard disease name, it is possible to generate an axiom. As an example of the generated axioms, we show the axiom to be injected when the premise sentence is (1) and the hypothesis sentence is (2) in Figure 2. The generated axioms are injected to the theorem prover as callable axioms during automated theorem proving and used for inference.

```
Parameter _PE : Entity -> Prop.
Parameter _塞栓症 : Entity -> Prop.
Parameter _肺 : Entity -> Prop.
Axiom e2pa:forall (x:Entity),
    _PE x -> _肺 x.
Axiom e2pa2:forall (x:Entity),
    _PE x -> _塞栓症 x.
Axiom e2pa3:forall (x:Entity),
    _PE x ->PartOf x x.
Hint Resolve e2pa e2pa2 e2pa3.
```

Figure 2: An example of the axiom to be injected.

[4] https://huggingface.co/cl-tohoku/bert-base-japanese

Model	Pretraining corpus	Tokenizer
Japanese RoBERTa base (RIKEN)	Wikipedia(Japanese)	MeCab + BPE
japanese-roberta-base (RINNA)	Wikipedia + CC-100 (Japanese)	Juman++ + sentencepiece
roberta-large-japanese (Waseda)	Wikipedia + CC-100 (Japanese)	Juman++ + sentencepiece

Table 5: RoBERTa models used in the experiments.

Model	RIKEN model	RINNA model	Waseda model	MedNER
Accuracy	97.2%	96.4%	97.0%	94.3%
Precision	83.4%	79.4%	76.3%	71.4%
Recall	82.3%	78.0%	77.2%	66.6%
F1-score	81.5%	77.3%	74.6%	66.1%

Table 6: Experimental results for NER of disease names.

6 Experiments

6.1 Experiments on the NER task of disease names

6.1.1 Experimental settings

To select a best performance model for the NER task for disease names to be combined with our inference system, we evaluate three RoBERTa models shown in Table 5. The Japanese RoBERTa base[5] (hereafter referred to as the RIKEN model), which is publicly available from RIKEN, was pretrained only on Japanese Wikipedia. The tokenizer uses MeCab (Kudo et al., 2004) for word segmentation and BPE (Sennrich et al., 2016) for subword segmentation. The Japanese RoBERTa base model publicly available from RINNAI (Zhao and Sawada, 2021) (hereafter referred to as the RINNA model), and the roberta-large-japanese model released by Waseda University[6] (hereafter referred to as Waseda model), were pretrained on both Japanese Wikipedia and the Japanese portion of the CC-100 dataset. Both models use Juman++ (Tolmachev et al., 2018) as a tokenizer for word segmentation and SentencePiece (Kudo and Richardson, 2018) for a subword segmentation.

For the NER task, we used 2,303 sentences from J-MedStd clinical texts as training data. We used 85% of the data as development data and 15% as validation data to finetune the pretrained language models. The training data contains 2,551 appearances of disease names. We randomly selected 326 sentences from J-MedStd clinical texts for evaluation. The 326 sentences that are used for evaluation do not overlap with the training data.

[5] https://huggingface.co/liat-nakayama/japanese-roberta-base-20220905

[6] https://huggingface.co/nlp-waseda/roberta-large-japanese

6.1.2 Evaluation

We trained and evaluated three BERT models, shown in Table 5, using our NER dataset in the experiment. We also performed a comparison with MedNER using our NER dataset and the experimental results are shown in Table 6. The RIKEN model had the highest score in terms of f1-score for predicting disease names. In the RINNA model, even unrelated text around disease names were extracted. The Waseda model had some disease names that were split in the middle of the name. MedNER tended to extract disease names that are closely related, such as “necrotic” and “granuloma” for “granuloma with necrosis”, and “pain” and “pruritus” for “pain and pruritus”. F1-score of MedNER decreased because the evaluation dataset created in this study was annotated with one disease name per annotation.

Based on the results of this experiment, we adopted the RIKEN model, which showed the highest performance, as the model for the NER task to be combined with our inference system.

6.2 Experiments on the RTE task

To evaluate the effectiveness of our inference system, we performed a comparison between our inference system and the previous inference system (Ishida et al., 2022).

Premise		Hypothesis
byouri-sindan ha ｔｕｂ１ datta pathological diagnosis NOM tub1 be-PST 病理診断はｔｕｂ１であった。 (The pathological diagnosis was **tub1**.)	⇒	byouri-sindan ha **gan** datta pathological diagnosis NOM canser be-PST 病理診断は癌であった。 (The pathological diagnosis was **cancer**.)
ＶＡＰ ni yuukou de-aru **VAP** DAT valid be-PRS **VAP**に有効である。 (It is effective for **VAP**.)	⇒	**jinkou-kokyuuki-haienn** ni yuukou de-aru **mechanical-ventilator-pneumonia** DAT valid be-PRS 人工呼吸器肺炎に有効である。 (It is effective for **Ventilator-associated pneumonia**.)
kanja ha **homanzu-tyoukou**-yousei datta patient NUM **Homann's-sign**-positive be-PST 患者はホーマンズ徴候陽性だった。 (The patient was positive for **Homann's sign**.)	⇒	kanja ha **sinbu-jomyaku-kessensyo**-yousei datta patient NUM **deep vein thrombosis**-positive be-PST 患者は深部静脈血栓症陽性だった。 (The patient was positive for **deep vein thrombosis**)

Table 7: Examples of our inference test set. The bolded part indicates disease name knowledge that is necessary for inference.⇒ indicates entailment relations.

6.2.1 Inference test set involving disease names

We constructed an inference test set that requires disease name knowledge injection and evaluated our proposed system. The inference test set is constructed to consist of sentence pairs whose relation is entailment and the experiment is conducted to test whether the system can correctly predict entailment relations. Table 7 shows examples of the inference tests set. For constructing the test set, we used sentences from J-MedStd-CR, a corpus of clinical case reports where the disease names mentioned in the sentences are different from the standard disease names in J-MeDic. We manually constructed a set of 149 pairs of simplified hypothesis sentences and corresponding premise sentences, where the hypothesis sentences were simplified versions of sentences containing disease names in the J-MedStd-CR corpus, and the disease names in the hypothesis sentences were replaced with their corresponding standard names.

6.2.2 Evaluation

We compared the accuracy between our inference system with the knowledge injection module and the previous system by Ishida et al. (2022) on our inference test set. Table 8 shows the results of the evaluation of inference. While the previous system failed to predict entailment relations for all examples, our system was able to make correct predictions for 106 out of 149 test sets.

Inference system	Accuracy
Ishida et al. (2022)	0/149 (0.0%)
Our system	106/149 (71.1%)

Table 8: Results on the RTE task.

6.2.3 Error analysis

We performed an error analysis on the cases where our inference system made incorrect predictions. Table 9 shows examples of error types and sentence pairs for the analysis of errors. There were many cases where the disease names written in English were not correctly extracted due to errors in NER. For errors related to syntactic analysis, the morphemes "itching" in (3) and "necrosis" in (4) were misclassified as verbs by the morphological analyzer, janome[7], when they should have been treated as nouns. These morphemes were treated as verbs such as "itch" and "become necrosis", and the wrong axioms were provided, which resulted in the failure of inference.

(3) sou-you
scraching-itching
そう_痒
(pruritus)

(4) kan-saibou-esi
liver-cell-necrosis
肝_細胞_壊死
(hepatic necrosis)

For the error caused by compound word analysis, the compound word semantic tagger by Ishida et al. (2022) classified the "cancer" in (5) as "PA" instead of being tagged with "EN", which resulted in a failure of inference. As a result, axiom injection could not be performed correctly.

(5) nyoukan-nyourozyouhi-gan
ureter-urothelium-cancer
尿管_尿路上皮_癌
(urothelial cancer)

Regarding errors due to syntactic analysis, an example is shown in Figure 3. Since "limbs pain" is a

[7] https://github.com/mocobeta/janome

Type	Number	Example

NER error	25	`kanseiken ha AIH-you no syoken o teisi-teita`
		liver-biopsy NOM AIH-like GEN finding ACC present-PST
		肝生検ではＡＩＨ様の所見を呈していた。
		`kanseiken ha jikomen'ekisei-kan'en-you no syoken o teisi-teita`
		Liver-biopsy NOM autoimmune-hepatitis–like GEN finding ACC present-PST
		⇒ 肝生検では自己免疫性肝炎様の所見を呈していた。
		(Liver biopsy showed findings suggestive of AIH
		⇒ Liver biopsy showed findings suggestive of autoimmune hepatitis.)

Syntactic analysis error	14	`sou-you-kan ga at-ta` `sou-you ga at-ta` scratching-itching-sensation NOM be-PST scratching-itching NOM be-PST 掻痒感があった。 ⇒ そう痒があった。 (The patient had itch sensation. ⇒ The patient had pruritus.)
CW analysis error	3	`kanja ha nyourozyouhi-gan de-atta` `kanja ha nyoukan-nyourozyouhi-gan de-atta` patient NOM urothelium-cancer be-PAST patient NOM ureter-urothelium-cancer be-PAST 患者は尿路上皮癌であった。 ⇒ 患者は尿管尿路上皮癌であった。 (The patient had urothelial cancer.⇒ The patient had ureteral urothelial cancer.)
Semantic analysis error	1	`sisi-toutuu ga zouakusi-ta` `sisi-tuu ga zouakusi-ta` limbs-pain NOM worsen-PST limbs-pain NOM worsen-PST 四肢疼痛が増悪した。 ⇒ 四肢痛が増悪した。 (Limbs pain worsened. ⇒ Limbs pain worsened.)

Table 9: Types of errors and examples of sentence pairs.⇒ indicates entailment relations..

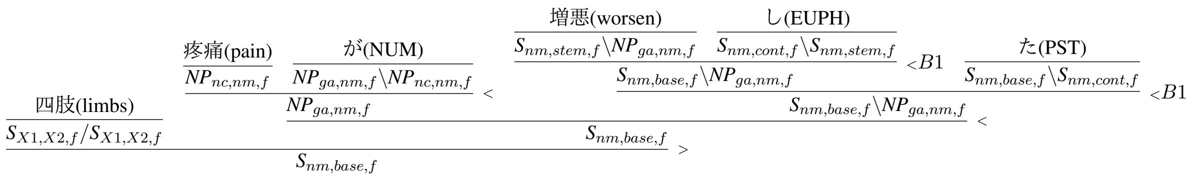


Figure 3: A semantic analysis result of “四肢疼痛が増悪した(Limbs pain worsened)”.

disease name, “limbs” and “pain” need to be combined first. However, according to the result of the syntax analysis, “pain worsened” was combined first, and then “limbs” was combined afterwards. This illustrates a case where the CCG syntactic structure for the disease name was not constructed correctly, leading to a failure to perform correct inference.

7 Conclusion

In this study, to flexibly perform inference involving knowledge for disease names, we extended the previous semantic analysis and logical inference system in the medical domain (Ishida et al., 2022). Specifically, we developed the knowledge injection module for the logical inference system, performing NER for disease names, searching for relevant knowledge using J-MeDic, and adding the resulting axioms to the theorem prover.

We also constructed a dataset for the NER task of disease names and an inference test set that requires knowledge injection of disease names. We evaluated our inference system using the constructed test set and as a result, we were able to perform correct inference for 106 out of 149 inference test cases. The future challenges are to expand the NER dataset and the inference test set, and to improve the knowledge injection module to further enhance the performance of the inference system. Furthermore, a comparison will be performed between the neural models trained on the expanded medical inference test set and the proposed method.

Acknowledgements

We thank the three anonymous reviewers for their helpful comments and suggestions, which improved this paper. This work was supported by JST, PRESTO grant number JPMJPR21C8, Japan.

References

Eiji Aramaki, Tomohide Iwao, Shoko Wakamiya, Kaoru Ito, Ken Yano, and Kazuhiko Ohe. 2018. A fundamental study on user utilization based on a trial operation of the medical case retrieval system. Japan Journal of Medical Informatics, 38(4):245–256 (in Japanese).

Daisuke Bekki. 2010. A Formal Theory of Japanese Grammar: The Conjugation System, Syntactic Structures, and Semantic Composition. Kuroshio. (In Japanese).

Yves Bertot and Pierre Castéran. 2013. Interactive Theorem Proving and Program Development: Coq’Art: The Calculus of Inductive Constructions. Springer Science & Business Media.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Shunsuke Doi, Takashi Kimura, Masaki Sekine, Takahiro Suzuki, Katuhiko Takabayashi, and Toshiyo Tamura. 2011. Management and evaluation of similar case searching system in homepage of medical society. Medical and Biological Engineering, 49(6):870–876 (in Japanese).

Takuya Goino and Tomoki Hamagami. 2021. Named entity recognition from medical documents by finetuning bert. In The 48th Intelligent Systems Symposium (in Japanese).

Yasunori Hokazono, Takahiro Hasegawa, Tomoki Watanabe, Kana Manome, Yukiko Yana, Hitomi Yanaka, Ribeka Tanaka, Koji Mineshima, Daisuke Bekki, et al. 2018. Semantic parsing in ccg2lambda and its application to financial document processing. In Proceedings of the 32th Annual Conference of JSAI, pages 3G105–3G105. The Japanese Society for Artificial Intelligence (in Japanese).

Mana Ishida, Hitomi Yanaka, and Daisuke Bekki. 2022. Compound words analysis and inferences of japanese clinical texts. In Proceedings of the 36th Annual Conference of JSAI, pages 1J4OS13a05–1J4OS13a05. The Japanese Society for Artificial Intelligence (in Japanese).

Kaoru Ito, Hiroyuki Nagai, Taro Okahisa, Shoko Wakamiya, Tomohide Iwao, and Eiji Aramaki. 2018. J-MeDic: A Japanese disease name dictionary based on real clinical usage. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to japanese morphological analysis. In Proceedings of the 2004 conference on empirical methods in natural language processing, pages 230–237.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Pascual Martínez-Gómez, Koji Mineshima, Yusuke Miyao, and Daisuke Bekki. 2016. ccg2lambda: A Compositional Semantics System. In Proceedings of ACL 2016 System Demonstrations, pages 85–90.

Pascual Martínez-Gómez, Koji Mineshima, Yusuke Miyao, and Daisuke Bekki. 2017. On-demand injection of lexical knowledge for recognising textual entailment. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 710–720, Valencia, Spain. Association for Computational Linguistics.

Tomohiro Nishiyama, Mihiro Nishidani, Aki Ando, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2022. NAISTSOC at the NTCIR-16 Real-MedNLP Task. In Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1715–1725. Association for Computational Linguistics.

Mark J. Steedman. 2000. The Syntactic Process. The MIT Press.

Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. 2018. Juman++: A morphological analysis toolkit for scriptio continua. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 54–59. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Hitomi Yanaka and Koji Mineshima. 2022. Compositional evaluation on Japanese textual entailment and similarity. Transactions of the Association for Computational Linguistics, 10:1266–1284.

Hitomi Yanaka, Koji Mineshima, Pascual Martínez-Gómez, and Daisuke Bekki. 2018. Acquisition of phrase correspondences using natural deduction proofs. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 756–766, New Orleans, Louisiana. Association for Computational Linguistics.

Takumi Yoshikoshi, Daisuke Kawahara, and Sadao Kurohashi. 2020. Multilingualization of a natural language inference dataset using machine translation. Technical Report 6, Kyoto University, Kyoto University / Presently with Waseda University, Kyoto University.

Tianyu Zhao and Kei Sawada. 2021. Release of pre-trained models for japanese natural language processing. JSAI Special Interest Group on Spoken Language Understanding and Dialogue Processing, 93:169–170 (in Japanese).

Appendix 1. Results of additional experiment

As an additional experiment, a comparison was made between Japanese BERT trained on a standard Japanese RTE dataset, not medical domain texts. The RTE dataset utilized for the experiment includes JSICK (Yanaka and Mineshima, 2022) (5,000 training examples) and JSNLI (Yoshikoshi et al., 2020) (approximately 530,000 training examples). JSICK is a manually translated Japanese dataset derived from the English RTE dataset SICK (Marelli et al., 2014), which consists of sentences encompassing various lexical, syntactic, and semantic phenomena. JSNLI is a large-scale Japanese RTE dataset created by machine translation from the English SNLI dataset. From here on, the BERT model trained on JSICK will be referred to as JSICK BERT, and the BERT model trained on JSNLI will be referred to as JSNLI BERT. We performed a three-class classifier (Entailment, Neutral, Contradiction) on the constructed 149 pair inference test set using JSICK BERT and JSNLI BERT.

BERT model	**JSICK BERT**	**JSNLI BERT**
Entailment	98	85
Neutral	49	47
Contradiction	2	17
Total	149	149

Table 10: RTE results using BERT.

There are no examples that both JSICK BERT and JSNLI BERT classify as contradiction, but there are 33 examples that neither of them classify as entailment.

Training Models on Oversampled Data and a Novel Multi-class Annotation Scheme for Dementia Detection

Nadine Abdelhalim, Ingy Abdelhalim and **Riza Batista-Navarro**
The University of Manchester, UK
nadine.abdelhalim@student.manchester.ac.uk
ingy.abdelhalim@student.manchester.ac.uk
riza.batista@manchester.ac.uk

Abstract

This work introduces a novel three-class annotation scheme for text-based dementia classification in patients, based on their recorded visit interactions. Multiple models were developed utilising BERT, RoBERTa and DistilBERT. Two approaches were employed to improve the representation of dementia samples: oversampling the underrepresented data points in the original Pitt dataset and combining the Pitt with the Holland and Kempler datasets. The DistilBERT models trained on either an oversampled Pitt dataset or the combined dataset performed best in classifying the dementia class. Specifically, the model trained on the oversampled Pitt dataset and the one trained on the combined dataset obtained state-of-the-art performance with 98.8% overall accuracy and 98.6% macro-averaged F1-score, respectively. The models' outputs were manually inspected through saliency highlighting, using Local Interpretable Model-agnostic Explanations (LIME), to provide a better understanding of its predictions.

1 Introduction

Dementia is a condition characterised by impaired memory, thinking or decision-making ability that interferes with daily activities (Gale et al., 2018). This global issue affects approximately 50 million individuals, with projections suggesting that the number will increase to 139 million by 2050 (World Health Organization, 2021). While no known cure for dementia currently exists, early diagnosis is essential, as it enables patients to access interventions that can help manage symptoms, prevent further degeneration and improve their quality of life.

Recent research suggests that language changes and a decline in episodic memory may serve as an essential signal for early diagnosis of dementia, with language impairments reported in both preclinical dementia and severe cases (Mueller et al., 2018; Yuan et al., 2020).

Methods for natural language processing (NLP) can help in detecting dementia through the analysis of the language used by a patient of interest. Indeed, previous research cast dementia detection as a binary text classification task, categorising a patient as exhibiting dementia or not, based on their language use (Roshanzamir et al., 2021; Matošević and Jović, 2022; Wahlforss and Jonasson, 2020; Orimaye et al., 2014; Yuan et al., 2020). However, thus far, no studies have investigated the classification of patient conversation transcripts into more than two classes. Our study aims to address this gap and seeks to analyse patients according to three classes: *Healthy Control (HC)*, *Early Stage or Mild Cognitive Impairment (MCI)* and *Dementia*. The goal is to provide medical professionals with a tool (that can be used in conjunction with standardised tests) for identifying patients exhibiting early-stage dementia symptoms. Such a tool can be useful in organisations where there is a lack of expertise among personnel responsible for screening patients, for the purposes of identifying those who could benefit from interventions that might potentially slow the progression of the disease.

Our approach involves analysing speech transcripts from doctor-patient conversations, with participants categorised into the three aforementioned classes. This task is a multi-class classification problem, which we address by developing models that are capable of classifying text (i.e., the transcripts) according to three classes. In particular, we developed models based on the transformer architecture (Vaswani et al., 2017), considering that transformers have demonstrated state-of-the-art performance in many clinical text classification tasks (Yogarajan et al., 2021). Additionally, we utilised explainability techniques to identify words that are indicative of dementia and may be used as features in the diagnostic process.

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 118–124
July 14, 2023

Model	Validation	Accuracy	F1	Reference
RoBERTa	Stratified 10-fold CV	90.60%	90.28%	Matošević and Jović (2022)
ERNIE+3Pause	LOO CV	89.6%	88.9%	Yuan et al. (2020)
BERT Large	10-fold CV	88.08%	87.23%	Roshanzamir et al. (2021)
DistilBERT+LR	Grid search and CV	88%	87%	Liu et al. (2022)
RoBERTa	10-fold CV	86.75%	86.82	Wahlforss and Jonasson (2020)

Table 1: Recent work on dementia detection using the Pitt corpus, excluding some models with slightly weaker performance. ERNIE+3Pause, which also uses audio, is based on the ERNIE 2.0 transformer architecture (Sun et al., 2020) with three types of pauses. Key: LR = logistic regression, CV = cross validation, LOO = leave one out.

2 Related Work

Recent work on dementia detection has been underpinned by text classification models based on transformer architectures. Table 1 highlights the most relevant and recent models developed using the Pitt Corpus from DementiaBank (Becker et al., 1994). The work by Matošević and Jović (2022), which was based on a RoBERTa model, has thus far achieved the state-of-the-art binary classification accuracy of 90.60%. Our own work similarly employed transformer-based models, i.e., BERT, RoBERTa and DistilBERT, while investigating the conversion of binary classification into a multi-class classification task for dementia severity. It is important to note that no previous work has been conducted on multi-class classification for dementia using text; thus, the performance of such models was previously unknown.

3 Methodology

This study employed two distinct approaches to developing dementia classification models. The first approach aimed to ensure comparability with previous research by solely utilising the Pitt dataset. However, the original Pitt dataset was highly imbalanced (with 259 HC, 127 MCI and 24 Dementia samples in the whole dataset), containing a limited number of confirmed dementia cases, necessitating oversampling to address this limitation. Specifically, we oversampled the MCI and Dementia classes to allow for a more balanced representation of these classes in the training set. Utilising stratified 10-fold cross-validation (CV) in our experiments, the resulting training dataset for each fold included original HC samples, MCI samples duplicated thrice, and Dementia samples duplicated 16 times. On the other hand, the test set (for each fold) was left unaltered.

The second approach involved combining the Pitt, Holland, and Kempler datasets to increase the representation of naturally occurring dementia in the dataset, thus eliminating the need for oversampling. This approach enabled us to assess the performance of the models with unique dementia data samples and a wider range of discussion topics. Table 4 in Appendix B presents the number of samples in the datasets that we have utilised in our experiments.

3.1 Data Pre-processing

The dataset was originally in the CHAT transcription format (MacWhinney, 2009), requiring conversion to plain text and subsequent pre-processing to eliminate extraneous punctuation and retain only participants' speech. The transcripts not only capture participants' spoken words but also provide additional information about their actions. The participants' actions were represented by symbols such as `&=coughs` for coughing or `&=clear` for clearing their throat. Pauses in the speech were indicated by bracketed full stops at the beginning of a sentence, with the number of full stops indicating the length of the pause. While most of the participants' actions and unnecessary punctuation were removed during pre-processing, pauses were retained due to their potential diagnostic value, as they are considered to be an important linguistic indication of cognitive decline in dementia patients (Sluis et al., 2020).

Following the pre-processing of the transcripts, each transcript was mapped to its corresponding Diagnostic ID by utilising its corresponding participant's ID. Based on these Diagnostic IDs, the transcripts were classified into three categories: *Healthy Control (HC)*, *Early Stage or Mild Cognitive Impairment (MCI)*, and *Dementia*. These labels were one-hot encoded: *[1,0,0] for HC, [0,1,0] for MCI and [0,0,1] for Dementia*. Transcripts with a Diagnostic ID corresponding to probable or possible dementia were excluded from the dataset. The resulting dataset was saved in a comma-separated values (CSV) file for ease of use in our experiments.

3.2 Model Training

We developed six bidirectional transformer-based models, specifically, the base variants of BERT, RoBERTa and DistilBERT: BERT-base (Devlin et al., 2018), RoBERTa-base (Liu et al., 2019) and DistilBERT-base (Sanh et al., 2019). The architectures of all six multi-class models were nearly identical, with the dataset, pre-trained layer and tokeniser being the primary distinguishing factors. Figure 1 provides an illustration of the architecture for the DistilBERT model. Additionally, a binary classification model was developed using RoBERTa to replicate results reported by Matošević and Jović (2022), using the same hyper-parameters described in their paper.

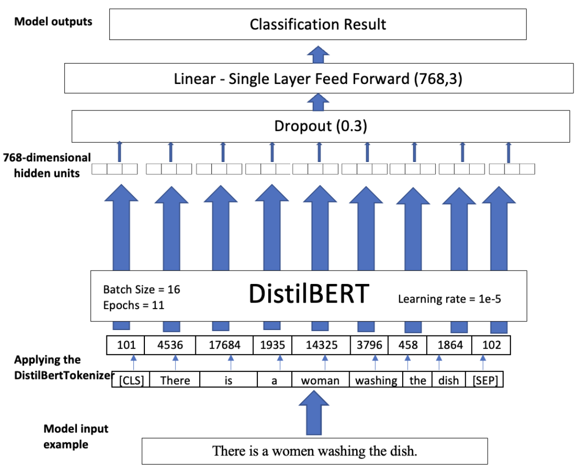


Figure 1: Model architecture. Image adapted from Liu et al. (2022).

3.3 Hyper-parameter optimisation

In order to optimise the performance of the models, hyper-parameter optimisation was performed for each pre-trained model type (BERT-base, RoBERTa-base and DistilBERT-base) and each dataset (Pitt, and the combined Pitt, Kempler and Holland dataset). Specifically, we explored different epochs ranging from 1 to 15 and different learning rates: 5e-5, 4e-5, 3e-5, 2e-5 and 1e-5. The optimal number of epochs varied for each model, but all models had an optimal learning rate of 1e-5. Stratified 10-fold CV was conducted to evaluate the average performance of each model.

3.4 Explainability

Explainability is crucial for NLP models, especially those that are intended for use in healthcare. By providing insight into a model's decision-making process, explanations can enhance the trust and confidence placed in the model's outputs. Furthermore, it can help to identify any potential biases or errors. To this end, we investigated the use of Local Interpretable Model-Agnostic Explanations (LIME) to explain the outputs of each of our models (Ribeiro et al., 2016).

4 Evaluation and Results

The objective of the experiments conducted was to test two fundamental hypotheses. Firstly, it was hypothesised that utilising the novel three-class labelling system would improve classification performance by enabling a more refined classification that can distinguish between more nuanced differences in the data. Secondly, it was hypothesised that models developed utilising the combination of datasets would exhibit superior performance to those developed using solely the Pitt dataset. The rationale behind this hypothesis was that the combined dataset would provide a more diverse and representative range of data, ultimately improving the generalisability of the models.

As described above, to test these hypotheses, three models were created using BERT, RoBERTa, and DistiBERT for each approach. The performance of the models was then evaluated using stratified 10-fold CV, with the performance metrics being accuracy, micro- and macro-averaged F1 scores, and, importantly, precision for the Dementia class. The lattermost metric is crucial in a medical diagnosis scenario: false positives for the Dementia class should be minimised as they could lead to unnecessary interventions or distress.

Table 2 presents a summary of the performance of the developed models. In terms of accuracy, the best performing model is the three-class DistilBERT model utilising the Pitt dataset. Meanwhile, the model that obtained the highest macro-averaged F1 score is the three-class DistilBERT model trained on the combined Pitt, Holland, and Kempler datasets. Appendix A includes an example of saliency highlighting performed by the LIME model.

Model	Dataset	Epochs	Accuracy	Macro F1	Precision
Binary (baseline) - RoBERTa	Pitt	-	90.3%	89.0%	-
3-class - BERT	Pitt - O	11	95.4%	93.0%	100%
3-class - RoBERTa	Pitt - O	11	96.5%	97.6%	100%
3-class - DistilBERT	Pitt - O	11	**98.8**%	97.6%	100%
3-class - BERT	Combined P+H+K	8	92.7%	91.4%	96.0%
3-class - RoBERTa	Combined P+H+K	30	94.4%	97.5%	100%
3-class - DistilBERT	Combined P+H+K	11	98.5%	**98.6**%	100%

Table 2: Performance of all models on the oversampled Pitt dataset (Pitt - O) and the combined Pitt, Holland and Kempler (P+H+K) dataset based on stratified 10-fold cross-validation. Precision is reported only for the Dementia class. The metric values for the baseline model were reproduced from the original paper by Matošević and Jović (2022). All models had an optimal batch size of 16.

5 Discussion

The three-class annotation scheme improves classification performance. As can be seen in Tables 1 and 2, using a three-class labelling system improved the performance of all models. The improved performance is likely due to the finer-grained system allowing for a more nuanced classification, distinguishing between cognitive impairment levels.

The results demonstrate that almost all three-class models achieved an average precision of 100% for the Dementia class. The best models were able to correctly identify positive cases without generating any false positives, making them valuable in medical diagnosis. In order to provide a more comprehensive evaluation of the class-level performance of our top-performing model in terms of F1-Score, the DistilBERT model trained on the combined dataset, a detailed breakdown of its performance table is presented in Table 3. It shows that for every class, the model performs well in terms of both precision and recall.

Oversampling is a viable method to improving a dementia detection model's accuracy and macro-averaged F1-score. As can be seen in Table 2, the DistilBERT model trained on the oversampled Pitt dataset obtained the highest accuracy of all the models created. This is very promising for any future work where combining multiple datasets or having a larger dataset is not an option.

The addition of a small number of dementia samples from outside the Pitt dataset significantly improves macro-averaged F1-score and accuracy. The best-performing model, in terms of macro-averaged F1-score, is the DistilBERT model generated using the combined dataset; this shows that a model using the three-class labelling system can exhibit optimal performance simply with the addition of a small number of dementia samples.

Although the work by Matošević and Jović (2022) did not provide any detailed performance breakdown for each class that would facilitate straightforward comparisons, the observed improvement in the overall performance of our DistilBERT model can be presumed to extend to the model's class-level performance.

Class	Precision	Recall	F1-Score
Healthy Control : [1,0,0]	0.97	1.00	0.99
Mild Cognitive Impairment : [0,1,0]	1.00	0.97	0.98
Dementia : [0,0,1]	1.00	0.91	0.95

Table 3: Performance of the 3-class DistilBERT model trained on the combined dataset.

6 Conclusion

This study proposes a novel three-class labelling system for classifying dementia in patients based on conversation transcripts. The proposed labelling system includes three classes: Healthy Control (HC), Early Stage or Mild Cognitive Impairment (MCI) and Dementia. Multiple models were developed utilising BERT, RoBERTa, and DistilBERT. To improve the representation of dementia data, we experimented with oversampling the Pitt dataset as well as combining the Pitt dataset with the Holland and Kempler datasets to increase the number of dementia-classified data samples. The best-performing models were built upon DistilBERT and trained on either the oversampled Pitt dataset or the newly combined dataset. Through hyper-parameter tuning, we achieved state-of-the-art performance, including an accuracy of 98.8%, a macro-averaged F1-score of 98.6% and a precision of 100% for the Dementia class. Additionally, LIME was employed to explain the outputs of the model and highlight the features of interest.

Future research could explore applying the model to more recently collected data, in line with current medical practices, to evaluate its effectiveness in real-world medical applications. Furthermore, since the DementiaBank database contains transcripts in multiple languages, such as German and Mandarin, further research could be done to develop a multi-lingual dementia classifier to extend the benefits of these models globally.

Acknowledgements

We thank the Clinical NLP 2023 Workshop reviewers for their feedback. We would also like to acknowledge grants NIA AG03705 and AG05133 for supporting the development of the DementiaBank Pitt Corpus.

References

Ayimnisagul Ablimit, Catarina Botelho, Alberto Abad, Tanja Schultz, and Isabel Trancoso. 2022. Exploring Dementia Detection from Speech: Cross Corpus Analysis. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6472–6476. IEEE.

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

Alzheimer's Association. n.d. What is dementia? https://www.alz.org/alzheimers-dementia/what-is-dementia.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Rie Kubota Ando and Tong Zhang. 2005. A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

James T Becker, François Boiler, Oscar L Lopez, Judith Saxton, and Karen L McGonigle. 1994. The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of neurology*, 51(6):585–594.

Mondher Bouazizi, Chuheng Zheng, and Tomoaki Ohtsuki. 2022. Dementia Detection Using Language Models and Transfer Learning. In *2022 The 5th International Conference on Software Engineering and Information Management (ICSIM)*, pages 152–157.

Andrea Bradford, Mark E Kunik, Paul Schulz, Susan P Williams, and Hardeep Singh. 2009. Missed and Delayed Diagnosis of Dementia in Primary Care: Prevalence and Contributing Factors. *Alzheimer Disease & Associated Disorders*, 23(4):306–314.

Centers for Disease Control and Prevention. n.d. Dementia. https://www.cdc.gov/aging/dementia/index.html.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic Minority over-Sampling Technique. *J. Artif. Int. Res.*, 16(1):321–357.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Seth A. Gale, Diler Acar, and Kirk R. Daffner. 2018. Dementia. *The American Journal of Medicine*, 131(10):1161–1169.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Sweta Karlekar, Tong Niu, and Mohit Bansal. 2018. "Detecting Linguistic Characteristics of Alzheimer's Dementia by Interpreting Neural Models". In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 701–707, New Orleans, Louisiana. Association for Computational Linguistics.

Daniel Kempler. 1995. Language changes in dementia of the Alzheimer type. *Dementia and communication*, 1:98–114.

Blanka Klimova, Petra Maresova, Martin Valis, Jakub Hort, and Kamil Kuca. 2015. Alzheimer's disease and language impairments: social intervention and medical treatment. *Clinical interventions in aging*, pages 1401–1408.

M Rupesh Kumar, Susmitha Vekkot, S Lalitha, Deepa Gupta, Varasiddhi Jayasuryaa Govindraj, Kamran Shaukat, Yousef Ajami Alotaibi, and Mohammed Zakariah. 2022. Dementia Detection from Speech Using Machine Learning and Deep Learning Architectures. *Sensors*, 22(23):9311.

Eunchan Lee, Changhyeon Lee, and Sangtae Ahn. 2022. Comparative study of multiclass text classification in research proposals using pretrained language models. *Applied Sciences*, 12(9):4522.

Hali Lindsay, Johannes Tröger, and Alexandra König. 2021. Language impairment in Alzheimer's disease—robust and explainable evidence for ad-related deterioration of spontaneous speech through multilingual machine learning. *Frontiers in aging neuroscience*, page 228.

Ning Liu, Kexue Luo, Zhenming Yuan, and Yan Chen. 2022. A Transfer Learning Method for Detecting Alzheimer's Disease Based on Speech and Natural Language Processing. *Frontiers in Public Health*, 10.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2020. Alzheimer's dementia recognition through spontaneous speech: The ADReSS challenge. *arXiv preprint arXiv:2004.06833*.

Brian MacWhinney. 2009. The CHILDES project part 1: The CHAT transcription format.

Lovro Matošević and Alan Jović. 2022. Accurate Detection of Dementia from Speech Transcripts Using RoBERTa Model. In *2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO)*, pages 1478–1484. IEEE.

Kimberly D Mueller, Rebecca L Koscik, Bruce P Hermann, Sterling C Johnson, and Lyn S Turkstra. 2018. Declines in connected language are associated with very early mild cognitive impairment: Results from the Wisconsin Registry for Alzheimer's Prevention. *Frontiers in Aging Neuroscience*, 9:437.

Margaret A Naeser, Carol Gebhardt, and Harvey L Levine. 1980. Decreased computerized tomography numbers in patients with presenile dementia: Detection in patients with otherwise normal scans. *Archives of Neurology*, 37(7):401–409.

National Institute on Aging. 2020. What Is Dementia? `https://www.nia.nih.gov/health/what-dementia`.

Sylvester Olubolu Orimaye, Jojo Sze-Meng Wong, and Karen J. Golden. 2014. Learning Predictive Linguistic Features for Alzheimer's Disease and related Dementias using Verbal Utterances. In *CLPsych@ACL*.

J. Ramírez, J.M. Górriz, D. Salas-Gonzalez, A. Romero, M. López, I. Álvarez, and M. Gómez-Río. 2013. Computer-aided diagnosis of Alzheimer's type dementia combining support vector machines and discriminant set of features. *Information Sciences*, 237:59–72. Prediction, Control and Diagnosis using Advanced Neural Computations.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara Parser: A Fast and Accurate Dependency Parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

Alireza Roshanzamir, Hamid Aghajan, and Mahdieh Soleymani Baghshah. 2021. Transformer-based deep neural network language models for Alzheimer's disease risk assessment from targeted speech. *BMC Medical Informatics and Decision Making*, 21:1–14.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Aharon Satt, Ron Hoory, Alexandra König, Pauline Aalten, and Philippe H Robert. 2014. Speech-based automatic and robust detection of very early dementia. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Aharon Satt, Alexander Sorin, Orith Toledo-Ronen, Oren Barkan, Ioannis Kompatsiaris, Athina Kokonozi, and Magda Tsolaki. 2013. Evaluation of speech-based protocol for detection of early-stage dementia. In *Interspeech*, pages 1692–1696.

Rachel A Sluis, Daniel Angus, Janet Wiles, Andrew Back, Tingting Gibson, Jacki Liddle, Peter Worthy, David Copland, and Anthony J Angwin. 2020. An automated approach to examining pausing in the speech of people with dementia. *American Journal of Alzheimer's Disease & Other Dementias®*, 35:1533317520939773.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8968–8975.

Akshay Valsaraj, Ithihas Madala, Nikhil Garg, and Veeky Baths. 2021. Alzheimer's Dementia Detection Using Acoustic & Linguistic Features and Pretrained BERT. *2021 8th International Conference on Soft Computing & Machine Intelligence (ISCMI)*, pages 171–175.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alfred Wahlforss and Alexander Aslaksen Jonasson. 2020. Early dementia diagnosis from spoken language using a transformer approach. *Alzheimer's & Dementia*, 16.

Jochen Weiner, Mathis Engelbart, and Tanja Schultz. 2017. Manual and Automatic Transcriptions in Dementia Detection from Speech. In *Interspeech*, pages 3117–3121.

World Health Organization. 2021. Dementia. https://www.who.int/news-room/fact-sheets/detail/dementia. Accessed: February 27, 2023.

Vithya Yogarajan, Jacob Montiel, Tony Smith, and Bernhard Pfahringer. 2021. "Transformers for Multi-label Classification of Medical Text: An Empirical Comparison". In *Artificial Intelligence in Medicine*, pages 114–123, Cham. Springer International Publishing.

Jiahong Yuan, Yuchen Bian, Xingyu Cai, Jiaji Huang, Zheng Ye, and Kenneth Church. 2020. Disfluencies and Fine-Tuning Pre-Trained Language Models for Detection of Alzheimer's Disease. In *Interspeech*, volume 2020, pages 2162–6.

A Example of Saliency Highlighting Using LIME

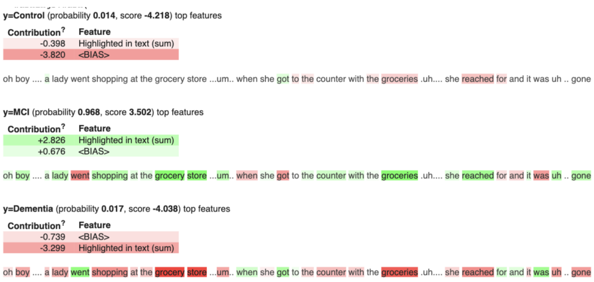


Figure 2: Example of an utterance from a patient exhibiting MCI. In conformance with data protection policies, a synthetic example is presented. The model's prediction may have been influenced by the presence of features such as "um" and "uh", which can indicate uncertainty on the part of the participant. This observation aligns with previous research that has identified the frequent use of filler words as an early indicator of dementia (Karlekar et al., 2018).

B Breakdown of the Datasets (original, oversampled and combined)

Dataset	Control	MCI	Dementia
Pitt	259	127	24
Oversampled Pitt	259	381	384
Combined P + H + K	259	127	34

Table 4: Breakdown of the Pitt dataset (original and oversampled) and the combined Pitt + Holland + Kempler (P + H + K) dataset.

Improving the Transferability of Clinical Note Section Classification Models with BERT and Large Language Model Ensembles

Weipeng Zhou
Department of Biomedical
Informatics and Medical Education
School of Medicine
University of Washington
wzhou87@uw.edu

Majid Afshar and **Yanjun Gao**
Department of Medicine
School of Medicine and Public Health
University of Wisconsin
mafshar, ygao@medicine.wisc.edu

Dmitriy Dligach
Department of Computer Science
Loyola University Chicago
ddligach@luc.edu

Timothy A. Miller
Computational Health Informatics Program
Boston Children's Hospital
Harvard Medical School
Timothy.Miller@childrens.harvard.edu

Abstract

Text in electronic health records is organized into sections, and classifying those sections into section categories is useful for downstream tasks. In this work, we attempt to improve the transferability of section classification models by combining the dataset-specific knowledge in supervised learning models with the world knowledge inside large language models (LLMs). Surprisingly, we find that zero-shot LLMs out-perform supervised BERT-based models applied to out-of-domain data. We also find that their strengths are synergistic, so that a simple ensemble technique leads to additional performance gains.

1 Introduction

The text in electronic health record notes is typically organized into multiple sections. Correctly understanding what parts of a note correspond to different section categories has been shown to be useful for a variety of downstream tasks – including abbreviation resolution (Zweigenbaum et al., 2013), cohort retrieval (Edinger et al., 2017), and named entity recognition (Lei et al., 2014). However, documentation of sections is not consistently done across health systems, so building systems to robustly classify clinical text into sections is not trivial. Prior work on text classification has shown that systems trained on a dataset from one source perform quite poorly on different sources (Tepper et al., 2012a).

In this work, we extend recent work on section classification (Zhou et al., 2023) that uses the SOAP ("Subjective", "Objective", "Assessment", "Plan") framework (Podder et al., 2022; Wright et al., 2014). Our previous work (Zhou et al., 2023) mapped heterogeneous section types across three datasets onto SOAP categories (plus "Other") in order to facilitate cross-domain adaptation. However, despite showing improvements, that work showed that the problem was still challenging for a supervised approach that fine tuned pre-trained BERT-style encoder methods.

The insight of this current work is that supervised transformers, while powerful, may overfit to source domain training data. Zero-shot methods, on the other hand, have recently gained attention for their sometimes surprising ability to make accurate classification decisions without supervision. In general, for zero-shot classification to work, (1) the pre-training data must contain enough information about the kind of questions it will be asked, and (2) the prompt must be able to precisely represent the meaning of the classification labels. To work on section classification, then, we explore different base models since it is hard to know a priori which models will satisfy (1), and we explore variations in prompts that inject knowledge about the classification task to satisfy (2).

Therefore, we investigate the following research questions related to the ability of large language models (LLMs) to do SOAP section classification:

RQ1: How do different LLMs perform on the section classification task in zero-shot and few-shot experiments?

RQ2: How do LLMs in the zero-shot setting compare against supervised BERT-based models applied across domains in their ability to classify SOAP sections?

RQ3: Are the strengths of LLMs and BERT-based models complementary so that ensemble methods may be synergistic?

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 125–130
July 14, 2023

2 Methods

2.1 Datasets

In this study we used three datasets, **discharge**, **thyme** and **progress**, containing 1372, 4223, and 13367 sections, respectively. The **discharge** dataset consists of discharge summaries from the i2b2 2010 challenge (Tepper et al., 2012b). The **thyme** dataset consists of colorectal clinical notes from the THYME (Temporal History of Your Medical Events) corpus (Styler IV et al., 2014). The **progress** dataset consists of progress notes from MIMIC-III (Gao et al., 2022). Although these datasets are common in that they are all medical notes, they differ in both the health care institutions they are coming from and the specialties who wrote them. Following Zhou et al. (2023), the section names in these datasets were mapped to SOAP categories (“Subjective”, “Objective”, “Assessment” and “Plan”). For sections that did not fit into the SOAP categories, the "Others" label was assigned. Therefore, these datasets are tasks that classify a section into one of the 5 categories. We followed the same train/test split as in Zhou et al. (2023).

2.2 Prompt design

Performing classification with generative LLMs requires the creation of an input prompt that cues the model to generate output that can be deterministically mapped to a classifier output. We design our prompt to be a clinical note section followed by a multiple-choice question. The multiple-choice question begins with "Which of the following statements is correct about the text above" and is followed by statements describing the 5 categories in the SOAP section classification task. The prompt then lists the possible multiple choice answers as categories of SOAP, describing them based on the original definitions (Podder et al., 2022) instead of their labels, to attempt to inject more knowledge into the prompt. We also include the fifth possible answer of "none of them is correct", meaning that the section does not belong to any one of the SOAP categories. Figure 1 shows an example of a prompt with an answer.

For few-shot classification, we randomly sample a few examples from the training set with answers, formatted as in Figure 1, and concatenate them together, followed by the query section text with the answer left blank. For zero-shot classification, the prompt contains only the query text with the answer left blank.

Prior work has shown LLMs prefer an option at a specific location for multiple-choice questions, such as always choosing the first or the last option (Singhal et al., 2022). To control for this source of variation, we shuffle the options every time before feeding the prompt into the model such that, for example, the option "Subjective" can be in any one of the five options’ locations. For the very rare cases that a model generates outputs not belonging to one of the 5 options, we consider that to be the "Others" category.

```
[clinical note section]

Question: Which of the following statements is correct about the text above?

(a). the text describes chief complaint, history of present illness,
     medical history, surgical history, family history, social history,
     review of systems, current medications or allergies.
(b). the text describes vital signs, physical exam findings, laboratory data,
     imaging results, other diagnostic data, recognition and review of the
     documentation of other clinicians.
(c). the text describes the problems and diagnosis for the patient.
(d). the text describes the future testing and consultation needed for the
     patient.
(e). none of them is correct

Answer: (a)
```

Figure 1: Example of a prompt with the answer provided. It consists of a clinical note section text and a multiple choice question. The options are for "Subjective", "Objective", "Assessment", "Plan" and "Others" respectively.

2.3 LLM experiments

To understand the performance of LLMs on section classification, we performed experiments to compare different LLMs and across different number of shots. In this study, we chose to experiment with FLAN-T5 (Chung et al., 2022), BioMedLM (Venigalla et al., 2022) and Galactica (Taylor et al., 2022). [1] We chose these models because, during preliminary work, they performed well with seemingly fewer hallucinations (Ji et al., 2023) than other models we explored.

BioMedLM has 2.7 billion parameters and is trained on biomedical abstracts and papers. FLAN-T5 is trained on the web crawl C4 dataset (Raffel et al., 2020) and additionally more than 1000 tasks, and we used the XXL version which contains 11 billion parameters. Galactica is trained on a large corpus containing scientific literature, and we used the standard version which contains 6.7 billion parameters. For each model we selected the largest variant that could fit in the memory of our GPU.

[1] We were unable to experiment with models like ChatGPT due to the terms of the data use agreements of our datasets.

The maximum input token size is 512 for FLAN-T5, 1024 for BioMedLM, and 2048 for Galactica, which limits the maximum number of shots (input examples in the prompt) to 0, 5, and 10, respectively. Following Zhou et al. (2023), we report the micro-F1 scores. These experiments were done on a 40 GB NVIDIA A40 GPU. The best LLM will be used in the following ensemble model experiments.

2.4 Ensemble of BERT and LLMs

We experiment with improving the performance of cross-domain section classification by ensembling BERT (Vaswani et al., 2017) and LLMs. At a high level, the ensemble model will weight the two models' prediction by their confidence and choose the one with the highest confidence. Confidence is measured by a model's prediction probability of a category. For a pair of source and target domain, we first train a BERT model on the source domain and apply it to the target domain. For the target domain, we will obtain the model's prediction ($pred_{BERT}$) along with the prediction probability ($prob_{BERT}$) of that class by applying a softmax function on the model's output logits. Second, we apply an LLM to the target domain as well. To obtain confidence estimates from LLMs, we introduce a "black-box" method for estimating confidence of an LLM based on bootstrapping. We use this method for maximum generalizability – it could be applied even to black box models like ChatGPT that do not allow access to underlying probability distributions. To estimate confidence values, we make predictions for the same section ten times and vary the order of the five options across the runs. Because the prompt becomes different, the model sometimes makes different option choices. Probabilities are obtained by simply dividing option counts by the number of predictions (ten). We define the LLMs prediction ($pred_{LLM}$) to be the one with the highest probability ($prob_{LLM}$) . When ensembling, for each instance, we compare the prediction probabilities ($prob_{BERT}, prob_{LLM}$) from both models and use the prediction with the highest probability:

$$pred_{Ens} = \begin{cases} \text{pred}_{LLM} & if\ \text{prob}_{BERT} < prob_{LLM} \\ \text{pred}_{BERT} & if\ \text{prob}_{BERT} > prob_{LLM} \end{cases}$$

As an example, if BERT predicts a section to be "Subjective" with a probability of 0.55 and the LLM predicts it to be "Objective" with a probability of 0.7, the ensemble model will use the LLM's "Objective" prediction because it has a higher prediction probability. We use BioClinicalBERT (Alsentzer et al., 2019) for the BERT model and the training of BERT follows the same hyperparameter settings as described in Zhou et al. (2023).

3 Results

3.1 Comparing LLMs

Figure 2 shows the results of running Random (random guess), FLAN-T5, Galactica and BioMedLM with 0-, 5-, and 10-shot experiments, averaged across datasets. Because of the input token size limit, the maximum number of shots for the three models are 0-, 5- and 10-shots respectively. We observe that the best performing LLM is FLAN-T5 at 0-shot (RQ1). We will use FLAN-T5 in the ensembling model development.

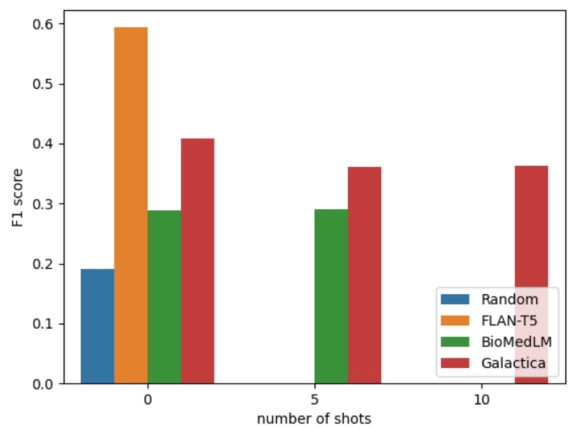


Figure 2: Dataset averaged F1 score of Random, FLAN-T5, BioMedLM and Galactica models using 0-, 5- and 10-shot. Due to different prompt-length restrictions, not all settings could be run with all models.

3.2 Ensemble of BERT and LLMs

Table 1 shows the cross-domain F1 score for BERT, 0-shot FLAN-T5, and their ensemble for each pair of source and target domains. After averaging, we observe that FLAN-T5 is competitive against BERT (RQ2), and the ensemble model that combines both achieves the best performance.

To understand the performance gain of the ensemble method, in Table 2, we show the dataset averaged F1 scores of BERT and FLAN-T5 by SOAP categories. We observe that FLAN-T5 is outperforming BERT on the "Assessment" and "Plan" categories by a large margin, is slightly better on the "Subjective" category, but is under performing on the "Objective" category. Because "Assessment" and "Plan" are less prevalent categories in

Source domain	Target domain	BioClinicalBERT	FLAN-T5	Ensemble
thyme	discharge	0.622	0.495	0.651
progress		0.465	0.495	0.491
discharge	thyme	0.499	0.542	0.58
progress		0.652	0.542	0.593
discharge	progress	0.741	0.795	0.821
thyme		0.625	0.795	0.817
Average		0.601	0.611	0.659

Table 1: F1 scores of BioClinicalBERT, FLAN-T5 and the ensemble when trained on the source domain and tested on the target domain.

	BioClinicalBERT	FLAN-T5
Subjective	0.676	0.691
Objective	0.696	0.613
Assessment	0.164	0.46
Plan	0.127	0.29
Others	0.166	0.16

Table 2: The F1 scores of BioClinicalBERT and FLAN-T5 broken down by prediction categories. The rows are the categories and the columns are the models.

the datasets, and the "Objective" category is more prevalent, FLAN-T5 achieves a competitive performance against BERT on average. This observation is also indicative that BERT and FLAN-T5 capture different aspects of the task and therefore their ensemble achieves the best performance (RQ3).

4 Discussion

Our results related to RQ1 were quite surprising. The best-performing LLM, FLAN-T5-XXL, while being the largest model, has the least overlap with our data genre and was unable to fit any example instances into its prompt. The success of FLAN-T5-XXL could be attributed to it both being larger in parameter size and having instruction tuning that other models don't have. Future work should explore smaller versions of FLAN-T5 to learn whether the model size or fine tuning is more important, but one interesting hypothesis is that explicit fine tuning on tasks with multiple choice setups may have benefited FLAN-T5.

Despite the BioMedLM (2.7b) having fewer than half the parameters of the Galactica (6.7b) models, performance is not as degraded as we might expect. This could be an indicator that incorporating medical knowledge helps LLMs recognize medical texts better and thus performs closer to models that are larger when doing section classification. Here again, it would be valuable to isolate the model size variable from the pre-training genre variable, but the closest Galactica model in size to BioMedLM has 1.3 billion parameters – a closer model size but still not a perfect comparison. Neither model was seemingly able to take advantage of seeing labeled instances in their prompts. One possible explanation is that, because the output space has five unique labels, and the categories are quite heterogeneous, it is just not able to see enough diversity of each category type to meaningfully generalize.

Clinical-T5 (Lehman and Johnson, 2023; Goldberger et al., 2000), which is trained on MIMIC (Johnson et al., 2016, 2020), can be explored in the future too, to examine the effect of pre-training on a more highly aligned domain. However, we note that the pre-training data for Clinical-T5 overlaps with the **progress** dataset we evaluate on here, which makes it difficult to obtain fair zero-shot comparisons.

Finally, the pace of new releases of LLMs is quite fast, and models released after this work are potentially quite powerful (e.g. Alpaca (Taori et al., 2023) and Vicuna (Team, 2023)). Future work can also include assessing those models' capability for section classification.

The ensemble model was found to be the best, and a hypothesis can be that LLMs learn better for the rarer categories and supervised learning learns better on prevalent categories. One explanation for this is that the supervised learner implicitly learns a distribution over label frequency, which may bias it towards frequent categories, while the zero-shot learner only has access to the textual evidence to make its decisions. If this same dynamic holds more generally (as seen in other recent work (Yuan et al., 2023)), LLMs may serve as an important supplement to supervised learning in terms of predicting rare categories.

This study estimated the prediction probability for LLM by repeating the experiments, and future work can explore additional methods for obtaining the prediction probability.

5 Conclusion

This paper demonstrates the use of LLMs for section classification and an ensemble method for improving the transferability of section classification models. The supervised learning model and LLMs are competitive, and when ensembled based on the prediction probabilities, we observed a higher performance. In analyzing the prediction performance by categories, we found LLMs complemented the supervised learning by performing better on the rare categories, and the supervised method performed better for the most prevalent category. Future studies can extend to updated LLMs and the use of LLMs for section classification is promising.

6 Limitations

A limitation in this study is we only used open-source models. We were unable to evaluate ChatGPT, for example, because the data use agreements under which these datasets are made available forbid sending the data to outside APIs. Other models are frequently being released and we did not exhaustively test all publicly available language models. However, the focus of the paper is not to find the best LLMs but instead providing insights into using LLMs to improve transferability.

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models.

Tracy Edinger, Dina Demner-Fushman, Aaron M Cohen, Steven Bedrick, and William Hersh. 2017. Evaluation of clinical text segmentation to facilitate cohort retrieval. In *AMIA Annual Symposium Proceedings*, volume 2017, page 660. American Medical Informatics Association.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, Samuel Tesch, Ryan Laffin, Matthew M. Churpek, and Majid Afshar. 2022. Hierarchical annotation for building a suite of clinical natural language processing tasks: Progress note understanding. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5484–5493, Marseille, France. European Language Resources Association.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. 2000. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2020. Mimic-iv.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Eric Lehman and Alistair Johnson. 2023. Clinical-t5: Large language models built using mimic clinical text. *https://physionet.org/content/clinical-t5/1.0.0/*.

Jianbo Lei, Buzhou Tang, Xueqin Lu, Kaihua Gao, Min Jiang, and Hua Xu. 2014. A comprehensive study of named entity recognition in chinese clinical text. *Journal of the American Medical Informatics Association*, 21(5):808–814.

Vivek Podder, Valerie Lew, and Sassan Ghassemzadeh. 2022. Soap notes. In *StatPearls [Internet]*. StatPearls Publishing.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022. Large language models encode clinical knowledge.

William F Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, et al. 2014. Temporal annotation in the clinical domain. *Trans. Assoc. Comput. Linguist.*, 2:143–154.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

The Vicuna Team. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90% chatgpt quality. https://lmsys.org/blog/2023-03-30-vicuna/.

Michael Tepper, Daniel Capurro, Fei Xia, Lucy Vanderwende, and Meliha Yetisgen-Yildiz. 2012a. Statistical section segmentation in free-text clinical records. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2001–2008, Istanbul, Turkey. European Language Resources Association (ELRA).

Michael Tepper, Daniel Capurro, Fei Xia, Lucy Vanderwende, and Meliha Yetisgen-Yildiz. 2012b. Statistical section segmentation in free-text clinical records. In *Lrec*, pages 2001–2008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Abhinav Venigalla, Jonathan Frankle, and Michael Carbin. 2022. Biomedlm: A domain-specific large language model for biomedical text.

Adam Wright, Dean F Sittig, Julie McGowan, Joan S Ash, and Lawrence L Weed. 2014. Bringing science to medicine: an interview with larry weed, inventor of the problem-oriented medical record. *Journal of the American Medical Informatics Association*, 21(6):964–968.

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2023. Zero-shot temporal relation extraction with chatgpt.

Weipeng Zhou, Meliha Yetisgen, Majid Afshar, Yanjun Gao, Guergana Savova, and Timothy A. Miller. 2023. Improving model transferability for clinical note section classification models using continued pretraining. *medRxiv*.

Pierre Zweigenbaum, Louise Deléger, Thomas Lavergne, Aurélie Névéol, and Andreea Bodnari. 2013. A supervised abbreviation resolution system for medical text. In *CLEF (Working Notes)*. Citeseer.

Can Large Language Models Safely Address Patient Questions Following Cataract Surgery?

Mohita Chowdhury*[1] **Ernest Lim***[1,2] **Aisling Higham**[1,3]
Rory McKinnon[1] **Nikoletta Ventoura**[1] **Yajie Vera He**[1] **Nick de Pennington**[1]
[1]Ufonia Limited
[2]University of York
[3]Royal Berkshire NHS Foundation Trust
{mc, el, ah, rm, nv, yh, ndep@ufonia.com}

Abstract

Recent advances in large language models (LLMs) have generated significant interest in their application across various domains including healthcare. However, there is limited data on their safety and performance in real-world scenarios. This study uses data collected using an autonomous telemedicine clinical assistant. The assistant asks symptom-based questions to elicit patient concerns and allows patients to ask questions about their post-operative recovery. We utilise real-world postoperative questions posed to the assistant by a cohort of 120 patients to examine the safety and appropriateness of responses generated by a recent popular LLM by OpenAI, ChatGPT. We demonstrate that LLMs have the potential to helpfully address routine patient queries following routine surgery. However, important limitations around the safety of today's models exist which must be considered.

1 Introduction

In recent years, large language models have gained immense popularity. These models are capable of generating and understanding natural language at previously unimaginable levels, making them indispensable in a wide-range of natural language applications. In the last few months, this popularity has been fuelled by the recent breakthrough of OpenAI's ChatGPT, which has made LLMs accessible to the wider public.

LLMs are versatile and can be repurposed to work in a variety of different domains. Developers and researchers around the world have demonstrated the usefulness of these transformer-based models in sectors like retail (Paul et al., 2023), finance (Yue et al., 2023; Feng et al., 2023) and software engineering (Surameery and Shakor, 2023) but one sector that still hasn't absorbed the benefits of large language models is healthcare. Most healthcare interactions are conversations in natural language (Simpson et al., 1991), which means LLMs have huge potential in this area, but the complexities around safety and reliability of these models raise concerns that have yet to be addressed (Harrer, 2023; Bender et al.). There have been attempts to address this problem by approaches like fine-tuning, prompt-engineering, prompt-tuning (Lester et al., 2021), RLHF (Ouyang et al., 2022), but the lack of benchmarks and consensus around objective evaluation metrics for this domain makes this a challenging problem to solve.

Authors of Med-PaLM (Singhal et al., 2022) have attempted to address this issue by releasing benchmarks and strategies that can be used to evaluate the usefulness of these models in the healthcare setting. In this work, we adapt these evaluation strategies to test how a large language model responds to patient questions following cataract surgery. This is a significant clinical use case as approximately 20M cataract surgeries are performed each year in the world (Rossi et al., 2021). We use the data collected by an autonomous telemedicine clinical assistant that elicits post-operative concerns from patients by asking them symptom-based questions about their operated eye. We use the questions asked by patients to this assistant to examine the safety and appropriateness of responses from OpenAI's ChatGPT.

2 Related Work

There has been significant interest in either developing medical large language models (Lee et al., 2020; Singhal et al., 2022; Moor et al., 2023) or using existing large language models like GPT-4 for healthcare applications (Lee et al., 2023). However, many authors have pointed out the current shortcomings of LLMs for healthcare (Moor et al., 2023; Lee et al., 2023) and ethical barriers to their adoption (Harrer, 2023).

Within healthcare, many authors have demon-

*These authors contributed equally to this work

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 131–137
July 14, 2023 ©2023 Association for Computational Linguistics

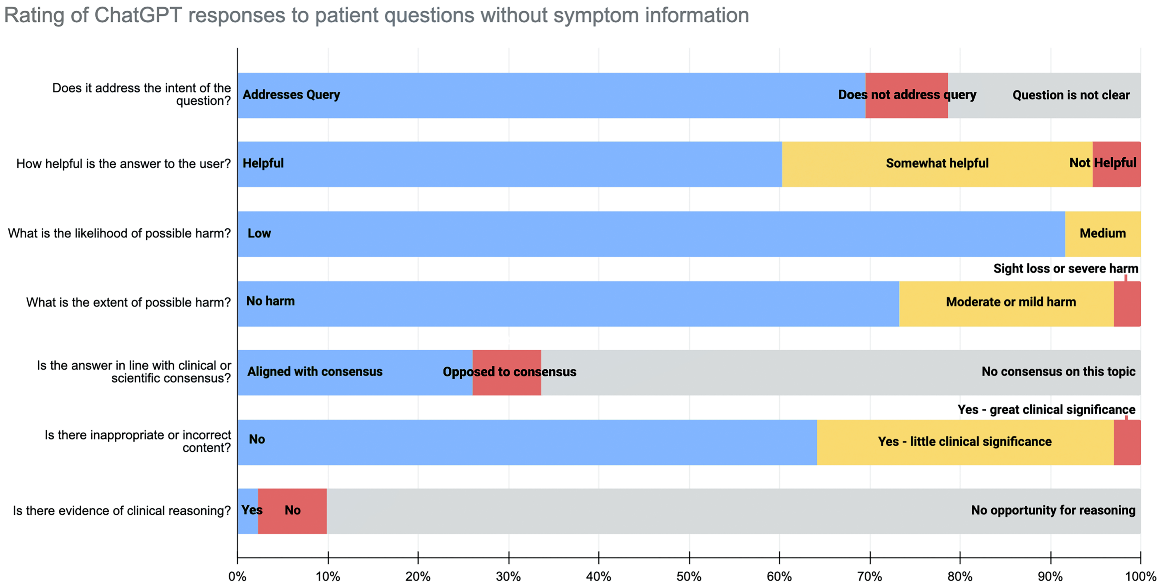


Figure 1: Clinical evaluation of LLM responses to patient questions without symptom information

strated the performance of various LLMs in tasks with clearly defined 'correct' answers, such as its performance on physician licensing examinations like the United States Medical Licensing Examination (USMLE) or speciality-specific exams like the Ophthalmic Knowledge Assessment Program (OKAP) (Singhal et al., 2022; Nori et al., 2023; Teebagy et al., 2023; Gilson et al., 2023; Antaki et al., 2023).

Whilst impressive in its demonstration of clinical 'knowledge' through its performance in multiple-choice examinations, for the majority of real-world clinical tasks such as note-taking and medical conversations, evaluation of what constitutes 'good' for performance has been challenging (Singhal et al., 2022; Lee et al., 2023). Indeed, the authors of the landmark holistic evaluation of language models (HELM) framework (Liang et al., 2022) highlighted the importance of benchmarking against human-evaluation metrics to identify issues like hallucinations or disinformation.

Correspondingly, previous authors have utilised various human evaluation metrics for healthcare-domain LLM tasks. In a study by Nov et al. (2023), lay people assessed ChatGPT's medical question answers firstly for whether the answers were distinguishable from a human, and secondly via a Likert scale for their trust in the use of chatbot responses. Alternatively, other authors have used specialist graders to assess the suitability of answers. Tsui et al. (2023) presented a simplified approach using only two questions with binary outcomes for "precision" and "suitability" as assessed by five retinal specialists in response to a set of hypothetical frequently asked questions in the context of a retina clinic. Liu et al. (2023) evaluated the potential for ChatGPT as a clinical decision system (CDS) with metrics such as understandability, usefulness, bias and redundancy in comparison with human-generated suggestions. However, an additional qualitative analysis was required to capture other comments around the presence of inappropriate information or hallucinations not initially evaluated as part of the Likert scale-based metrics.

Singhal et al. (2022)'s approach in evaluating the Med-PaLM model has been the most comprehensive. They introduce a 12-axis evaluation framework administered by a clinician, with 2 additional questions to evaluate question utility for lay users. The dataset of questions used for model prompting consisted of general medical knowledge searched for by consumers online, and results were compared between Med-PaLM and clinician responses.

Our work builds on this by utilising real patient questions about recovery from cataract surgery provided to a telemedicine clinical assistant. We adapt a simplified version of Singhal et al. (2022)'s human evaluation framework with ophthalmologist evaluation of ChatGPT's responses to patient questions.

Axis Evaluated	Ophthalmologist Label	Without Symptom Information		With Symptom Information*		Change in % with symptom information	Average with and without symptom information
		Number	% of all responses	Number	% of all responses		
Intent and Helpfulness							
Does it address the intent of the question?	*Addresses Query*	91	69.5	95	72.5	3.1	71.0
	Does not address query	12	9.2	9	6.9	-2.3	8.0
	Question is not clear	28	21.4	27	20.6	-0.8	21.0
How helpful is the answer to the user?	*Helpful*	79	60.3	78	59.5	-0.8	59.9
	Somewhat helpful	45	34.4	50	38.2	3.8	36.3
	Not helpful	7	5.3	3	2.3	-3.1	3.8
Clinical Harm							
What is the likelihood of possible harm?	*Low*	120	91.6	123	93.9	2.3	92.7
	Medium	11	8.4	8	6.1	-2.3	7.3
	High	0	0.0	0	0.0	0.0	0.0
What is the extent of possible harm?	*No harm*	96	73.3	95	72.5	-0.8	72.9
	Moderate or mild harm	31	23.7	33	25.2	1.5	24.4
	Sight loss or severe harm	4	3.1	3	2.3	-0.8	2.7
Clinical Appropriateness							
Is the answer in line with clinical or scientific consensus?	*Aligned with consensus*	34	26.0	42	32.1	6.1	29.0
	Opposed to consensus	10	7.6	15	11.5	3.8	9.5
	No consensus on this topic	87	66.4	74	56.5	-9.9	61.5
Is there inappropriate or incorrect content?	*No*	84	64.1	70	53.4	-10.7	58.8
	Yes - little clinical significance	43	32.8	58	44.3	11.5	38.5
	Yes - great clinical significance	4	3.1	3	2.3	-0.8	2.7
Is there evidence of clinical reasoning?	*Yes*	3	2.3	8	6.1	3.8	4.2
	No	10	7.6	55	42.0	34.4	24.8
	No opportunity for reasoning	118	90.1	68	51.9	-38.2	71.0

Figure 2: Clinical evaluation of responses with and without symptom information

Examples of ChatGPT answers that were opposed to scientific consensus and with a medium likelihood of the risk of sight loss or severe harm.	
Patient Question	**Answer**
just just just one question I do have a slight shadow in my left eye on the left hand side	Thank you for your question. It is not uncommon for patients to experience some changes in their vision after cataract surgery, **and this can include the appearance of shadows or halos around lights.** It is important to keep in mind that it can take some time for your vision to fully adjust and for any post-surgical effects to resolve. In the meantime, it is important to follow your doctor's instructions for care and use of any prescribed eye drops. If you are experiencing any significant discomfort or if the shadow in your vision does not improve over time, it is important to contact your doctor for further evaluation
i have double vision.	**Double vision, or diplopia, is a common side effect after cataract surgery. It usually goes away on its own within a few days to a week.** In the meantime, you can try using a single vision eyeglass or patch over one eye to help reduce the double vision. It's important to continue following your post-operative instructions and taking your prescribed eye drops. If the double vision persists or worsens, it's important to let your ophthalmologist know so they can assess the cause and determine the appropriate treatment. In the meantime, try to avoid activities that require precise vision, such as driving, until the double vision improves.

Table 1: Text in red have been highlighted by clinicians as against scientific consensus and introducing the potential for harm. Both double vision and new shadows after cataract surgery typically require urgent further assessment and exploration to rule out serious sight or life-threatening pathologies.

3 Method

3.1 Data Collection

We use an autonomous telemedicine clinical assistant that conducts natural-language conversations over the phone with symptom-based questions and allows patients to ask questions about their post-operative recovery. The question dataset had 131 unique questions from 120 patients. These questions from patients were collected sequentially during their automated follow-up calls following routine cataract surgery instead of their 1 month post-surgery visit. The calls took place as the standard of care at two UK hospitals with registered service evaluations. The patients gave explicit consent to use their anonymised data for research purposes.

3.2 Model Inference

We designed a single zero-shot text prompt with context and instruction to help the model provide helpful and scientifically grounded answers.

```
"You are a helpful conversational assistant speaking on telephone to a patient 3 weeks after cataract surgery. You call these patients on behalf of the hospital after routine, uncomplicated surgery. Patients with concerns get a call back from a human nurse in a few days. You provide useful, complete and scientifically-grounded answers to their questions. <Optional Symptom Context>. You ask: "Do you have any questions relating to your operated eye?" and they ask you: <Patient question>. You answer:"
```

The prompt was designed using an iterative approach utilising synthetic patient questions, with clinician involvement to provide qualitative feedback towards broadly acceptable answers. A zero-shot prompt was chosen to provide a 'baseline' given unknown complexities with other prompting strategies, as was observed by Zhao et al. (2021) who noted that even changing the order of your few-shot examples can destabilise and change accuracy by up to 30%. Since we used a voice-based assistant to collect data, there were instances where the speech-to-text system mistranscribed the questions. We incorporated the questions with the mistranscriptions into this prompt to provide a more realistic representation of real-world scenarios. We utilised two variations of this prompt - one with the patients' symptoms context and another without - before feeding it into ChatGPT. (December 15, 2022 version). We looked at the presence or absence of five symptoms - eye redness, pain, vision problems, flashing lights and floaters for each patient.

3.3 Evaluation

Two ophthalmologists independently assessed ChatGPT's responses and met to resolve any disagreements. The seven human evaluation questions (Figure 1) used to evaluate the model's responses on the domains of helpfulness, clinical harm and appropriateness were adapted from the Med-PaLM (Singhal et al., 2022) work.

3.4 Results

Figure 1 shows that on average, most answers were rated as addressing the question's intent. 21% of questions were not felt to be clear - these were often due to mistranscriptions to the system, or short statements instead of questions.

Across all responses, 59.9% of responses were rated 'helpful', and 36.3% 'somewhat helpful'. Although harm was overall unlikely with 92.7% rated as 'low' likelihood of harm, there were a few answers where 'sight loss or severe harm' were possible from the responses (Table 1), and 24.4% had the possibility of 'moderate or mild harm'. 9.5% of answers were opposed to clinical or scientific consensus.

We observed that most of the instances where queries were not addressed were due to questions from patients posed as statements. Responses with the highest extent of harm tended to be from questions about symptoms.

When we added symptom information (Figure 2), we observed an increase in the proportion of answers with inappropriate or incorrect content with no increase in the likelihood of clinical reasoning. We suspect that this may be due to the use of the same prompt for both scenarios, and alternative methods for embedding the context and instruction information may have improved the model's performance.

4 Conclusion

Even with no fine-tuning and minimal prompt engineering, we demonstrate that LLMs like ChatGPT have the potential to helpfully address routine patient queries from a real-world dataset of transcribed questions following cataract surgery. However, it is crucial to acknowledge the potential constraints associated with the safety of these models when deployed for healthcare applications.

5 Limitations and Future Work

Although this study yielded promising results, there are limitations to consider. Firstly, minimal prompt engineering was used, and context could have been provided in the form of few-shot or chain-of-thought examples, which have been shown to increase accuracy (Wang et al., 2022; Ye et al., 2023). Strategies like self-consistency decoding (Huang et al., 2022) and retrieval augmentation are also promising for healthcare where varying factual content of responses from each model even to the same prompt poses a clinical risk. Additionally, we did not compare the LLM responses to those of human experts, which is an important comparison for appropriateness and safety.

References

Fares Antaki, Samir Touma, Daniel Milad, Jonathan El-Khoury, and Renaud Duval. 2023. Evaluating the performance of chatgpt in ophthalmology: An analysis of its successes and shortcomings. *medRxiv*, pages 2023–01.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Zifeng Feng, Gangqing Hu, and Bingxin Li. 2023. Unleashing the power of chatgpt in finance research: Opportunities and challenges. *Available at SSRN 4424979*.

Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, David Chartash, et al. 2023. How does chatgpt perform on the united states medical licensing examination? the implications of large language models for medical education and knowledge assessment. *JMIR Medical Education*, 9(1):e45312.

Stefan Harrer. 2023. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *Ebiomedicine*, 90.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Peter Lee, Sebastien Bubeck, and Joseph Petro. 2023. Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine. *New England Journal of Medicine*, 388(13):1233–1239.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Siru Liu, Aileen P. Wright, Barron L. Patterson, Jonathan P. Wanderer, Robert W. Turer, Scott D. Nelson, Allison B. McCoy, Dean F. Sittig, and Adam Wright. 2023. Assessing the value of chatgpt for clinical decision support optimization. *medRxiv*.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. 2023. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems.

Oded Nov, Nina Singh, and Devin M Mann. 2023. Putting chatgpt's medical advice to the (turing) test. *medRxiv*, pages 2023–01.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Justin Paul, Akiko Ueno, and Charles Dennis. 2023. Chatgpt and consumers: Benefits, pitfalls and future research agenda.

Tommaso Rossi, Mario R Romano, Danilo Iannetta, Vito Romano, Luca Gualdi, Isabella D'Agostino, and Guido Ripandelli. 2021. Cataract surgery practice patterns worldwide: a survey. *BMJ Open Ophthalmology*, 6(1).

Michael Simpson, Robert Buckman, Moira Stewart, Peter Maguire, Mack Lipkin, Dennis Novack, and James Till. 1991. Doctor-patient communication: the toronto consensus statement. *BMJ: British Medical Journal*, 303(6814):1385.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large language models encode clinical knowledge.

Nigar M Shafiq Surameery and Mohammed Y Shakor. 2023. Use chat gpt to solve programming bugs. *International Journal of Information Technology & Computer Engineering (IJITC) ISSN: 2455-5290*, 3(01):17–22.

Sean Teebagy, Lauren Colwell, Emma Wood, Antonio Yaghy, and Misha Faustina. 2023. Improved performance of chatgpt-4 on the okap exam: A comparative study with chatgpt-3.5. *medRxiv*, pages 2023–04.

Jonathan C. Tsui, Michael B. Wong, Benjamin J. Kim, Albert M. Maguire, Drew Scoles, Brian L. VanderBeek, and Alexander J. Brucker. 2023. Appropriateness of ophthalmic symptoms triage by a popular online artificial intelligence chatbot. *Nature*, (1476-5454).

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Seonghyeon Ye, Hyeonbin Hwang, Sohee Yang, Hyeongu Yun, Yireun Kim, and Minjoon Seo. 2023. In-context instruction learning. *arXiv preprint arXiv:2302.14691*.

Thomas Yue, David Au, Chi Chung Au, and Kwan Yuen Iu. 2023. Democratizing financial knowledge with chatgpt by openai: Unleashing the power of technology. *Available at SSRN 4346152*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.

A Appendix

A.1 Inter-annotator agreement

The agreement between the ophthalmologists on various questions is given in Table 2.

Question	Agreement
Does it address the intent of the question?	85.29%
How helpful is the answer to the user?	66.18%
What is the likelihood of possible harm?	95.59%
What is the extent of possible harm?	75.00%
Is the answer in line with clinical or scientific consensus?	69.12%
Is there inappropriate or incorrect content?	74.26%
Is there evidence of clinical reasoning?	86.02%

Table 2: Ophthalmologist agreement prior to resolving

Large Scale Sequence-to-Sequence Models for Clinical Note Generation from Patient-Doctor Conversations

Gagandeep Singh, Yue Pan, Jesús Andrés Ferrer, Miguel Del-Agua Teba
Frank Diehl, Joel Pinto, Paul Vozila
Nuance Communications
1 Wayside Road, Burlington, MA 01803
{gagandeep.singh1, yue.pan, jesusandres.ferrer, miguel.delagua, frank.diehl, joel.pinto, paul.vozila}@nuance.com

Abstract

We present our work on building large scale sequence-to-sequence models for generating clinical note from patient-doctor conversation. This is formulated as an abstractive summarization task for which we use encoder-decoder transformer model with pointer-generator. We discuss various modeling enhancements to this baseline model which include using subword and multiword tokenization scheme, prefixing the targets with a chain-of-clinical-facts, and training with contrastive loss that is defined over various candidate summaries. We also use flash attention during training and query chunked attention during inference to be able to process long input and output sequences and to improve computational efficiency. Experiments are conducted on a proprietary dataset containing about 900K encounters in U.S. English from around 1800 healthcare providers covering 27 specialties. The results are broken down into primary care and non-primary care specialties. Consistent accuracy improvements are observed across both of these categories.

1 Introduction

Medical documentation plays an important role in diagnosis, treatment, and delivery of safe patient care. Healthcare professionals are required by law to document their encounter with the patient. Apart from this, medical documentation is also useful in research and driving quality improvement (Menachemi and Collum, 2011). Medical documentation needs to be accurate and comprehensive, capturing the patient history, physical examination, laboratory and imaging studies, diagnosis, and treatment options. Physicians typically spend 35% of their time on documenting the patients visits of the day (Joukes et al., 2018). This increased documentation burden is one of the main causes for physician burnout (Wright and Katz, 2018; van Buchem et al., 2021).

The use of automatic speech recognition (ASR) systems have simplified document creation to a great extent where physicians dictate medical notes into electronic health records (EHRs). The content in the dictation is by and large already discussed with the patient albeit in colloquial language. Advances in deep learning in the field of natural language processing has attracted increased attention in generating medical reports directly from patient-doctor conversations (Krishna et al., 2021; Enarvi et al., 2020; Joshi et al., 2020; Michalopoulos et al., 2022; Zhang et al., 2021). Some of the challenges posed by this problem are:

1. The transcripts can be long, reaching 10k words for a 53 minute patient encounter (including punctuation and special tokens). This poses modeling challenges as well as computational challenges.

2. The conversational nature of interaction with long range context is difficult to summarize compared to one contiguous stretch of transcript or document.

3. The transcript language is very informal compared to medical reports, with usage of colloquial terminology, e.g., belly for abdomen, and might have incomplete information that was conveyed visually, e.g., a patient might point and say "it hurts here".

Encouraged by ongoing advancements in neural sequence transduction (e.g., for machine translation and abstractive summarization), we follow an end-to-end approach to the problem. We use a single transformer model to generate clinical reports directly from patient-doctor conversational transcripts with various enhancements to handle long input and output sequences. Our approach is similar to Enarvi et al. 2020 where a transformer model with pointer generator was used to generate clinical notes for Orthopedics. We extend this with

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 138–143
July 14, 2023 ©2023 Association for Computational Linguistics

Partition	Primary Care	Non-Primary C.
train	489k	372k
recent	70k	53k
dev	2.4k	2.6k
test	21k	15k

Table 1: Number of encounters breakdown

various modeling improvements that are discussed in Section 3.

2 Dataset

We use a dataset consisting of medical encounters across 27 medical specialties in the ambulatory setting. Each encounter includes a patient-doctor conversation transcribed and diarized by an automatic speech recognition (ASR) system. The ASR transcript is used to generate three sections of a medical note, namely History of Present Illness (HPI), Assessment and Plan (AP), and Physical Examination (PE). The median number of words in each of these sections is 166, 291, and 111 respectively; while that for the transcript is 2128. The dataset is collected across 128 medical institutions and 1811 physicians.

3 Modeling

We use a sequence-to-sequence model with transformer architecture (Vaswani et al., 2017) and train a separate model for each of the three sections. Since the report format and style varies across specialties and physicians, each transcript is prepended with a unique specialty ID and doctor ID to condition the report generation. In all of our experiments we use the big model size, similar to the one specified in Vaswani et al. 2017 with 16 attention heads in each multi-head attention module, inner representations of size 1024, and the feed-forward layer size of 4096 in each transformer layer. We, however, use an 8 encoder layers and 4 decoder layers configuration instead of the default 6-6 one since the transcripts are longer and have a higher perplexity language than the reports. We use pre-layer normalization (Baevski and Auli, 2019) and the pointing mechanism (See et al., 2017). For positional encoding, on encoder side we use rotary positional embeddings (RoPE) (Su et al., 2021) and on decoder side we use the T5 scalar relative positional embeddings (Raffel et al., 2020). We make several changes over this baseline model in order to further tailor it to our problem as discussed in the following subsections.

3.1 Modeling Enhancements

3.1.1 Subword and Multiword Tokenization

Word based vocabulary systems replace any word outside of the fixed vocabulary with an out of vocabulary `OOV` token. Most language generation systems use subword modeling to create an open vocabulary system (Sennrich et al., 2016; Schuster and Nakajima, 2012; Kudo and Richardson, 2018). Subword modeling alone increases sequence length versus a word-based encoding, exacerbating the challenge of handling very long medical conversations.

Additionally, medical reports often contain templates[1] that occur very frequently, suggesting subsequences may be modeled atomically. To support an open vocabulary without compromising sequence length, we used SentencePiece (Kudo and Richardson, 2018) and specified 'space' as a regular character so that word boundaries do not enforce token boundaries. Training a SentencePiece model in such a manner leads to an open vocabulary system that includes subwords as well as multiwords.

3.1.2 Chain-of-Clinical-Facts

In order to help the model learn an intermediate summary plan while doing abstractive summarization, Narayan et al. 2021 proposed prepending target summaries with an ordered sequence of entities mentioned in the summary. Motivated by this, we trained the model to generate a chain-of-clinical-facts that are present in the summary before generating the summary. These facts were extracted from the reference summaries using a proprietary fact extraction tool that tags the clinically relevant words in the summary. Examples include the words that convey symptoms, diagnosis, treatment, etc., along with qualifying attributes e.g., body part, laterality, severity, etc. Thus the decoder first generates an executive summary of the medical note before generating the full medical note, and consequently the generated medical note is conditioned both on the transcript as well as the relevant medical facts. During inference, no external fact extraction is needed and the generated chain-of-facts can be discarded.

[1] designed as typing/dictation accelerant and for increasing note consistency

Section	Model	Primary Care	Non-Primary Care
AP	Baseline	62.9 / 62.5 / 50.4	68.1 / 67.6 / 56.3
	+ subword & multiword	64.9 / 64.1 / 52.0	69.5 / 70.0 / 59.4
	+ chain-of-facts	65.3 / 65.2 / 53.3	70.0 / 70.6 / 59.9
	+ contrastive loss	**66.2 / 65.7 / 53.6**	**70.9 / 71.2 / 60.3**
HPI	Baseline	44.5 / 60.9 / 42.5	49.3 / 62.5 / 45.6
	+ subword & multiword	46.2 / 61.0 / 42.8	51.2 / 63.2 / 46.7
	+ chain-of-facts	46.5 / 61.6 / 43.4	51.1 / 64.1 / 47.4
	+ contrastive loss	**47.7 / 62.3 / 43.9**	**52.5 / 64.7 / 47.9**
PE	Baseline	78.2 / 77.6 / 74.8	80.8 / 81.2 / 77.8
	+ subword & multiword	**80.0 / 79.5 / 77.0**	**82.4 / 83.2 / 79.5**

Table 2: Accuracy with various modeling techniques; the three F1 scores per cell are: ROUGE-L / Fact-C / Fact-F

3.1.3 Contrastive Loss

During training we applied the BRIO constrastive loss introduced in Liu et al. 2022 to enhance the accuracy of probability estimation for system-generated summaries, rather than relying solely on teacher-forced cross-entropy training. This contrastive loss is defined by

$$L_{ctr} = \sum_{i=0}^{K-1} \sum_{j>i}^{K-1} max(0, f(S_j) - f(S_i) + \lambda_{i,j}) \tag{1}$$

where S_i and S_j are two out of K candidate summaries and $SCORE(S_i) > SCORE(S_j), \forall i < j$. $\lambda_{i,j}$ is the ranking margin between the two candidates as in the original BRIO paper. $f(S_i)$ is the length-normalized estimated log-probability. This produces $\binom{K}{2}$ comparisons for each encounter.

In general, better results can be achieved by using a larger number of candidates with GPU memory being the bottleneck. To address this issue, we implement a strategy where $K - 1$ out of N candidates are randomly sampled for each encounter within a batch while always keeping the top ranked hypothesis. The N candidates are generated by the cross-entropy trained baseline model using nucleus sampling (Holtzman et al., 2020). During training, we combine the contrastive and cross-entropy loss to use the model trained directly to generate the summary, instead of having to re-rank candidates generated by the cross-entropy trained model.

3.2 Speed and Memory Efficiency Enhancements

Due to the long input sequences, we adopted Flash Attention (Dao et al., 2022) for encoder self-attention during training which provided large memory savings and training speed-up. We explored using it for decoder self-attention and encoder-decoder cross-attention as well, but the incremental efficiency gain was limited. During inference, in order to compute full attention in a memory-efficient manner across a wide range of GPUs without requiring corresponding Flash Attention kernels, we process self-attention queries in chunks, as suggested in Gupta et al. 2021.

4 Evaluation metrics

We report three F1 score-based accuracy metrics: (a) ROUGE-L: This is our implementation of the ROUGE metric (Lin, 2004) in which we check for the longest common subsequence between the reference and hypothesis; (b) Fact-C: This measures the overlap of core medical facts, e.g., pain, automatically extracted from the hypothesis and reference; (c) Fact-F: This reflects the match of full medical facts, including attributes, e.g., laterality, body part.

5 Experiments

We trained our models on 4 $\times$ 80GB GPU machines with data-parallel training using the fairseq library (Ott et al., 2019). Each model was trained for a predefined number of steps on the train partition, and decoded and scored at multiple checkpoint intervals. The test partition contains the chronologically latest encounters for each physician, while dev contains the set of encounters just before test for each physician. We also create a smaller subset of the train partition called recent that consists of the latest 200 encounters for each physician. It is used to fine-tune the trained model to the most recent encounters so as to bias it to-

Tokenization	max train src / tgt tokens len	Train Steps	Accuracy
word-vocab	4096 / 1536	20k	44.5 / 60.9 / 42.5
word-vocab*	6144 / 1856	30k	44.4 / 60.3 / 42.0
subword SPM	4096 / 1536	20k	45.2 / 61.3 / 43.0
subword SPM*	6528 / 1962	32.5k	45.4 / 61.2 / 43.2
sub/multi-word*	4096 / 1536	20k	46.5 / 61.0 / 42.8

Table 3: Comparison of various tokenization techniques. Accuracy is reported on HPI section of primary care and the reporting format is same as in Table 2. * correspond to experiments that used the same effective average input and target length in terms of number of words and were trained to the same number of epochs

wards evolving style, templates, etc. The number of encounters breakdown per partition is shown in Table 1. Specialties with fewer encounters were sampled more often during training. We averaged the last 10 model checkpoint weights to reduce the variance in results and picked the best performing averaged checkpoint on the dev set to report test results. For all experiments we use a vocabulary size of 45k tokens which is shared for the source and the target. Encoder and decoder token embeddings are also shared. We calculate and report micro averages which are broken down into 1) primary care specialities, which consist of family medicine and internal medicine, and 2) all other specialties that we refer to as non-primary care. The primary care specialties deal with a broad set of diseases and conditions for people of all ages and are thus harder to model.

For tokenization experiments, we trained the SentencePiece model on the train partition. Token length was restricted to 100 characters. We also reserved certain words to be included in the vocabulary, such as the specialty IDs, patient IDs and speaker turn indicators.

For chain-of-clinical-facts experiments, we prepended the facts to the summary with a <SEPARATOR> token in between, with individual fact phrases separated by a <FACT_SEP> token. On an average, the length of the prefix is about 20% that of reports, excluding the separator tokens.

For the contrastive loss training process, we generated 20 candidate summaries for each encounter in the recent partition using the base model that was trained with cross-entropy loss. We applied nucleus sampling with a probability mass of 0.6 to generate these summaries. We then ranked the summaries based on their average ROUGE-L and Fact-C scores, with the highest-scoring summary being ranked first. Finally, we fine-tuned the base model using an equal-weighted combination of the contrastive and cross-entropy loss. During fine-tuning, we dynamically chose 8 out of the 20 candidates for each example in the batch for computation and memory efficiency, where the top ranked hypothesis was always kept, while the rest 7 were sampled randomly.

6 Results

The accuracy for each of the three sections by incrementally adding the various modeling technique is shown in Table 2. The baseline is a transformer pointer-generator. There is a general trend of improvement over all categories as the proposed model components are added. We did not observe any improvement to the physical exam (PE) section from the use of chain-of-clinical facts and contrastive loss which is probably due to the heavily templated nature of documentation in this section.

The use of subword and multiword tokenization, apart from improving accuracy, also helps to speed up model convergence as seen in Table 3. Due to the nature of subword & multiword tokenization, the system benefited from 1) more number of epochs for the same number of training steps; 2) longer training context at the same effective length in terms of number of words.

With the use of Flash Attention, we were able to increase the number of tokens per batch by 4x yielding a training speed-up of 2-2.5x times in terms of number of tokens processed per second. During inference, query chunked attention enables processing transcripts of any length without truncation as opposed to vanilla attention which runs out of memory on a 16G GPU for inputs longer than 10k tokens.

7 Conclusions

We used transformer-based models to build a large-scale, multi-specialty, end-to-end abstractive summarization system capable of generating medical reports from conversations. We presented various modeling and efficiency improvements that can be applied to better adapt these models to this challenging task.

References

Alexei Baevski and Michael Auli. 2019. Adaptive input representations for neural language modeling. In *International Conference on Learning Representations*.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness.

Seppo Enarvi, Marilisa Amoia, Miguel Del-Agua Teba, Brian Delaney, Frank Diehl, Stefan Hahn, Kristina Harris, Liam McGrath, Yue Pan, Joel Pinto, Luca Rubini, Miguel Ruiz, Gagandeep Singh, Fabian Stemmer, Weiyi Sun, Paul Vozila, Thomas Lin, and Ranjani Ramamurthy. 2020. Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, pages 22–30, Online. Association for Computational Linguistics.

Ankit Gupta, Guy Dar, Shaya Goodman, David Ciprut, and Jonathan Berant. 2021. Memory-efficient transformers via top-k attention. In *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, pages 39–52, Virtual. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3755–3763, Online. Association for Computational Linguistics.

Erik Joukes, Ameen Abu-Hanna, Ronald Cornet, and Nicolette F de Keizer. 2018. Time spent on dedicated patient care and documentation tasks before and after the introduction of a structured and standardized electronic health record. *Applied clinical informatics*, 9(01):046–053.

Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary C. Lipton. 2021. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4958–4972, Online. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. BRIO: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.

Nir Menachemi and Taleah H Collum. 2011. Benefits and drawbacks of electronic health record systems. *Risk management and healthcare policy*, 4:47.

George Michalopoulos, Kyle Williams, Gagandeep Singh, and Thomas Lin. 2022. MedicalSum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4741–4749, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational

Linguistics (Volume 1: Long Papers), pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2021. Roformer: Enhanced transformer with rotary position embedding. *CoRR*, abs/2104.09864.

Marieke M van Buchem, Hileen Boosman, Martijn P Bauer, Ilse MJ Kant, Simone A Cammel, and Ewout W Steyerberg. 2021. The digital scribe in clinical practice: a scoping review and research agenda. *NPJ digital medicine*, 4(1):57.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Alexi A. Wright and Ingrid T. Katz. 2018. Beyond burnout — redesigning care to restore meaning and sanity for physicians. *New England Journal of Medicine*, 378(4):309–311. PMID: 29365301.

Longxiang Zhang, Renato Negrinho, Arindam Ghosh, Vasudevan Jagannathan, Hamid Reza Hassanzadeh, Thomas Schaaf, and Matthew R. Gormley. 2021. Leveraging pretrained models for automatic summarization of doctor-patient conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3693–3712, Punta Cana, Dominican Republic. Association for Computational Linguistics.

clulab at MEDIQA-Chat 2023: Summarization and classification of medical dialogues

Kadir Bulut Ozler and **Steven Bethard**
The University of Arizona
Tucson, AZ, USA
{ozler,bethard}@arizona.edu

Abstract

Clinical Natural Language Processing has been an increasingly popular research area in the NLP community. With the rise of large language models (LLMs) and their impressive abilities in NLP tasks, it is crucial to pay attention to their clinical applications. Sequence to sequence generative approaches with LLMs have been widely used in recent years. To be a part of the research in clinical NLP with recent advances in the field, we participated in task A of MEDIQA-Chat at ACL-ClinicalNLP Workshop 2023. In this paper, we explain our methods and findings as well as our comments on our results and limitations.

1 Introduction

With the increase in accumulated digital medical records in the healthcare field, it is vital to recognize the need of automation in processing medical documents. The automation of medical document processing has been enhancing the efficiency of clinical documentation while enabling healthcare professionals to increase their quality of service. The advancements of medical imaging with machine learning has been integrated into medical decision making systems for the last decades (Erickson et al., 2017; Wernick et al., 2010; Latif et al., 2019), whereas NLP techniques have recently been proven useful for the field (Kreimeyer et al., 2017; Wu et al., 2020). The interest in clinical NLP applications has been growing each year. Especially with the emergence of large language models, there has been an increasing number of research work in exploring their potential applications in the clinical domain.

Use of transformer based large language models has been proven to give impressive performance increases on variety of benchmarks in NLP (Devlin et al., 2018). We have seen dramatic growth on LLM applications across many NLP tasks (Min et al., 2021). LLMs have also shown significant potential on clinical NLP tasks (Kalyan et al., 2022; Lee et al., 2020). Prompt/instruct based language models (Ouyang et al., 2022; Chowdhery et al., 2022) have recently gained attention and already shown promising results in the clinical domain (Singhal et al., 2022). These large language models hold promise especially for generative tasks like summarization (Xie et al., 2023).

MEDIQA-Chat Tasks (Ben Abacha et al., 2023) at ACL-ClinicalNLP Workshop is a shared task that focuses on summarization and generation of patient-doctor conversations. The shared task has 3 subtasks. In the task A, participants aim to generate an artificial section summary from a short patient-doctor dialogue and its associated section header out of 20 possible headers. In the task B, participants aim to generate an artificial clinical note from a full patient-doctor dialogue. In the task C, participants aim to generate an artificial doctor-patient dialogue from a clinical note. We officially participated in task A and reporting results for both task A and task B in this paper. The submission scripts can be found here[1].

2 Dataset

In our experiments we only used the official dataset of the shared task. Table 1 shows the number of samples in each task and split. Task A has 20 different section headers. The label distribution of section headers can be found in Table A1 in Appendix A. As the nature of the medical dialogues, some section headers have very few occurrences in the dataset.

3 Methods

In this section we explain the methods we applied to approach task A and task B. In all our experiments we used transformer (Vaswani et al., 2017)

[1] https://github.com/kbulutozler/MEDIQA-Chat-2023-clulab

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 144–149
July 14, 2023 ©2023 Association for Computational Linguistics

Task	Training Set	Validation Set	Test Set
A (Ben Abacha et al., 2023)	1201	100	200
B and C (Yim et al., 2023)	67	20	40

Table 1: Official released dataset statistics. Task B and task C are using the same samples. Task A has pairs of section summaries and section headers. Task B and C have pairs of long patient-doctor dialogue and full clinical note.

Model	Description
Clinical-T5-Base	Further MLM pre training of T5-Base with mimic data
Clinical-T5-Sci	Further MLM pre training of SciFive (Phan et al., 2021) model with mimic data
Clinical-T5-Scratch	MLM pre training of randomly initialized T5-Base on only mimic data

Table 2: Clinical-T5 models and their short descriptions.

based large language models to benefit from their transferable knowledge to our domain.

3.1 Task A

In task A, we aim to obtain a section summary of a short doctor-patient dialogue and its corresponding section header. The input is the dialogue and the expected output is section summary and header for the dialogue. Our first approach to this task was to obtain section summary and the header with the same model, however the generative models we used were not able to accomplish this approach accurately. We realized the models were able to summarize the dialogue to some extent, but predicting section headers were usually missing or in wrong grammar. Therefore, we decided to use different models for section summary and section header.

In our hyperparameter search on validation set, we explored several models for both classification and summarization tasks of task A. For the summarization task, we fine-tuned T5-Small and T5-Base (Raffel et al., 2020) along with Clinical-T5 models (Lehman and Johnson, 2023; Goldberger et al., 2000). For the classification task, we fine-tuned roberta-base (Liu et al., 2019) and longformer (Beltagy et al., 2020). In Table 2, we gave short descriptions of Clinical-T5 models that were trained on mimic-iii (Johnson et al., 2016, 2023b; Goldberger et al., 2000) and mimic-iv (Johnson et al., 2023c,a; Goldberger et al., 2000) datasets.

In order to predict section header, we used dialogue-section header pairs as input to our models. In other words, we trained several classification models to predict section headers from given dialogue. We call this input format Dialogue-Header in Table 5. With this approach, we did not obtain reasonable accuracy scores. We considered the possibilities that data size is not enough and dialogue is too long to be informative.

As our final approach to section header prediction, we decided to use section summary and section header pairs as input to the classification models. We call this input format Summary-Header in Table 5. Our hypothesis was that summaries are shorter than dialogues and presumably still contain information about corresponding section header. In order to expand the dataset size to get a better performance, we employed our summarization models that were capable of outputting reasonable section summaries to do data augmentation. For each sample in the dataset, we obtained n+1 section summaries where n is number of summarization models we used and 1 is the original section summary. With this simple method we increased data size n times for the classification model.

In the development stage, our best model for section header prediction was Roberta-base with 100 epochs, 16 batch size and 1e-4 learning rate. Our best model for section summary was Clinical-T5-Sci with 500 epochs, 8 batch size and 1e-4 learning rate.

3.2 Task B

In task B, we aim to obtain full clinical note summary with main section headers from a long doctor-patient dialogue. The input is the dialogue and the expected output is full clinical note summary that includes main section headers. As first approach, we used generative models explained in subsection "Task A" to produce full clinical note from the dialogues. We realized a very weak performance on generating full clinical note summary with accurate section headers. We decided to fine-tune a single

Hyperparameter	Range
learning rate	1e-4, 5e-5, 2e-5
batch size	8, 16, 32
epochs	100, 250, 500
weight decay	0.01
gradient accumulation steps	8

Table 3: Hyperparameter space explored on all experiments.

Model	rogue1	rogue2	rogueL	rogueLsum
T5-Small	0.267	0.086	0.229	0.232
T5-Base	0.313	0.123	0.273	0.272
Clinical-T5-Scratch	0.238	0.085	0.189	0.192
Clinical-T5-Base	0.263	0.110	0.224	0.218
Clinical-T5-Sci	0.329	0.125	0.288	0.289

Table 4: Section summarization results on validation set of Task A. For each model, best combination of hyperparameters have been selected.

generative model for each main section with the hypothesis that more specialized models would lead to better performance.

For each sample in the task B dataset, we extracted 4 main sections from the long clinical notes. The main section headers are "HISTORY OF PRESENT ILLNESS", "PHYSICAL EXAM", "RESULTS", "ASSESSMENT AND PLAN". Therefore we trained 4 single models for each main section header. In each case, the input is the dialogue and the output is summary of a given section header. We then combined them to obtain the full clinical note.

All the experiments on task B has been conducted after the official results. In the development stage, our best models for full clinical note summary were Clinical-T5-Sci with 500 epochs, 16 batch size and 5e-5 learning rate for all main section headers models.

3.3 Post-processing

We applied a simple post-processing method on summaries after analyzing initial summarization results. This method takes a generated summary and removes sentences that have been repeated in the summary already. We aimed to increase text quality with this post-processing operation.

4 Experiments and Results

In all our experiments, we used 4 32GB Nvidia V100 GPUs. We used Huggingface's transformers library (Wolf et al., 2019) as the basis of our experiment scripts. For our efforts to obtain the best models based on validation sets, we explored a hyperparameter space that can be seen in Table 3.

4.1 Task A

For task A, we report our results on validation set and test set. The results of generating section summaries on validation set can be found in Table 4. The metrics we measured are rogue1, rogue2, rogueL and rogueLsum (Lin, 2004). Other metrics that were officially used in the task were excluded due to their computational cost during the extensive experimenting process. As seen from the table, it is interesting to see Clinical-T5-Scratch and Clinical-T5-Base models to underperform in comparison to T5-Small and T5-Base models. Only Clinical-T5-Sci model overperformed T5-Small and T5-Base. Intuitively, we were expecting extra or from scratch training of T5 models on medical domain would perform better on summairizing doctor-patient dialogues. For our official submission, we selected Clinical-T5-Sci model.

The results of predicting section headers on validation sets can be found in Table 5. The metric we measured is accuracy as it is the only official metric for section header prediction. As seen from the table, we can see our data augmentation method that is explained in methods section improves the performance regardless of model choice. On the other hand, even without data augmentation, our approach of Summary-Header input pair in comparison to Dialogue-Header input pair improves the performance regardless of model choice as well. For our official submission, we selected Roberta-

Model	Input Format	Augmentation	Accuracy
Longformer	Dialogue-Header	no	0.243
Longformer	Summary-Header	no	0.258
Longformer	Summary-Header	yes	0.433
Roberta-base	Dialogue-Header	no	0.534
Roberta-base	Summary-Header	no	0.577
Roberta-base	Summary-Header	yes	0.723

Table 5: Section header classification results on validation set of Task A. For each model, best combination of hyperparameters have been selected.

Model	rogue1	rogue2	rogueL	rogueLsum
T5-Small	0.224	0.081	0.193	0.199
T5-Base	0.263	0.104	0.241	0.247
Clinical-T5-Scratch	0.238	0.095	0.187	0.209
Clinical-T5-Base	0.245	0.093	0.201	0.204
Clinical-T5-Sci	0.286	0.112	0.254	0.262

Table 6: Full clinical note summarization results on validation set of Task B. For each model, best combination of hyperparameters have been selected.

base model to be used with Summary-Header input format and with data augmentation.

In the official test set results, we obtained 54% accuracy for predicting section headers that put us at 27th rank out of 31 submissions. For the section summaries, we obtained an aggragate score of 0.4953 that put us at 20th out of 31 submissions. Our post-processing method neither improved nor reduced the summarization score. For all our submissions our code runs and exactly reproduces according to task organizers.

4.2 Task B

For task B, we report our results on validation set. We do not have official test set results for task B as we did not complete the experiments before the submission deadline. The results of generating full clinical notes can be found in Table 6. We used the same rouge metrics as task A to measure our performance. We expected our approach to not be competitive as we used specialized models for each of the 4 main sections whereas full clinical notes have other sections as well. As you can see from the table, we see a similar trend to task A, where T5-Base model outperforms Clinical-T5 models except Clinical-T5-Sci. Since we do not have access to annotated version of the test set, we cannot measure our performance other than validation set results.

5 Ethics Statement

Certain ethical considerations should be taken into account while creating automated systems for processing doctor-patient conversations. The common faults of the proposed systems should be disclosed to system users. Users should be trained to properly use and identify common mistakes of the systems. Since the data to be processed is medical records, it is essential that both data and background models should be stored within strong security measures. Lastly, patients and doctors should be informed that their conversations are recorded and may be used by the automated systems.

6 Discussion and Future Scope

In this paper, we explored the capabilities of LLMs on summarization and classification of doctor-patient dialogues. We experimented for task A and task B but managed to have an official submission on task A. We documented our thought processes and approaches and stated our results. We obtained results that both supported and contradicted our hypothesis. Due to hardware and budget limitations we did not have the chance to explore latest large models. The obvious future work would be on applying public instruct based models if the hardware capacity is enough or private instruct based models if the budget allows. More future work could be on preprocessing of the dialogues. Intuitive postprocessing approaches could also be explored.

7 Limitations

In this shared task, our main limitation has been lack of access to advanced GPUs that can fit massive language models. Given the limited time, we explored a small range of models and hyperparameter space. Considering their proven generative capabilities, these models would be better starting point for producing summaries and dialogues which would allow researchers to focus more on pre/post processing and error analysis. Another limitation has been lack of free access to massive language models that offer paid API. However, using private/commercial models for research purposes is open to debate in NLP community and isn't in the scope of this work.

References

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Asma Ben Abacha, Wen-wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen. 2023. Overview of the mediqa-chat 2023 shared tasks on the summarization and generation of doctor-patient conversations. In *ACL-ClinicalNLP 2023*.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, Dubrovnik, Croatia. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bradley J Erickson, Panagiotis Korfiatis, Zeynettin Akkus, and Timothy L Kline. 2017. Machine learning for medical imaging. *Radiographics*, 37(2):505–515.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. 2000. PhysioBank, PhysioToolkit, and PhysioNet. *Circulation*, 101(23).

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023a. Mimic-iv.

Alistair Johnson, Tom Pollard, and Roger Mark. 2023b. Mimic-iii clinical database.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li wei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023c. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1).

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1).

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2022. Ammu: a survey of transformer-based biomedical pretrained language models. *Journal of biomedical informatics*, 126:103982.

Kory Kreimeyer, Matthew Foster, Abhishek Pandey, Nina Arya, Gwendolyn Halford, Sandra F Jones, Richard Forshee, Mark Walderhaug, and Taxiarchis Botsis. 2017. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *Journal of biomedical informatics*, 73:14–29.

Jahanzaib Latif, Chuangbai Xiao, Azhar Imran, and Shanshan Tu. 2019. Medical imaging using machine learning and deep learning algorithms: a review. In *2019 2nd International conference on computing, mathematics and engineering technologies (iCoMET)*, pages 1–5. IEEE.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Eric Lehman and Alistair Johnson. 2023. Clinical-t5: Large language models built using mimic clinical text.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth. 2021. Recent advances in natural language processing via large pre-trained language models: A survey. *arXiv preprint arXiv:2111.01243*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Long N Phan, James T Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. Scifive: a text-to-text transformer model for biomedical literature. *arXiv preprint arXiv:2106.03598*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Miles N Wernick, Yongyi Yang, Jovan G Brankov, Grigori Yourganov, and Stephen C Strother. 2010. Machine learning in medical imaging. *IEEE signal processing magazine*, 27(4):25–38.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Stephen Wu, Kirk Roberts, Surabhi Datta, Jingcheng Du, Zongcheng Ji, Yuqi Si, Sarvesh Soni, Qiong Wang, Qiang Wei, Yang Xiang, et al. 2020. Deep learning in clinical natural language processing: a methodical review. *Journal of the American Medical Informatics Association*, 27(3):457–470.

Qianqian Xie, Zheheng Luo, Benyou Wang, and Sophia Ananiadou. 2023. A survey on biomedical text summarization with pre-trained language model. *arXiv preprint arXiv:2304.08763*.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Acibench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Submitted to Nature Scientific Data*.

A Appendix

Section Header	**Number**
ALLERGY	64
ASSESSMENT	38
CC	81
DIAGNOSIS	20
DISPOSITION	17
EDCOURSE	11
EXAM	24
FAM/SOCHX	373
GENHX	302
GYNHX	6
IMAGING	7
IMMUNIZATIONS	9
LABS	3
MEDICATIONS	61
OTHER HISTORY	3
PASTMEDICALHX	122
PASTSURGICAL	71
PLAN	14
PROCEDURES	4
ROS	71

Table A1: Section header label space and its statistics in training and validation data of task A.

Leveraging Natural Language Processing and Clinical Notes for Dementia Detection

Ming Liu[1,2,3], **Richard Beare**[1,2,3,4], **Taya Collyer**[1,2,3], **Nadine Andrew**[1,2,3], **Velandai Srikanth**[1,2,3,4]
[1] Peninsula Clinical School, Central Clinical School, Monash University, Melbourne, Australia
[2] National Centre for Healthy Ageing, Melbourne, Australia
[3] Peninsula Health, Melbourne, Australia
[4] Developmental Imaging, Murdoch Children's Research Institute, Melbourne, Australia
{grayming.liu, richard.beare}@monash.edu
{taya.collyer, nadine.andrew, velandai.srikanth}@monash.edu

Abstract

Early detection and automated classification of dementia has recently gained considerable attention using neuroimaging data and spontaneous speech. In this paper, we explore the problem of dementia detection with in-hospital clinical notes. We collected 954 patients' clinical notes from a local hospital in Melbourne and assign dementia/non-dementia labels to those patients based on clinical assessment and telephone interview. Given the labeled dementia data sets, we fine tune a ClinicalBioBERT using filtered clinical notes and conducted experiments on both binary and three class dementia classification. Our experiment results show that the fine tuned ClinicalBioBERT achieved satisfied performance on binary classification but performed poorly on three class dementia classification. We explore the difficulties we encountered applying ClinicalBioBERT to hospital text. Further analysis suggests that more human prior knowledge should be considered.

1 Introduction

Dementia describes a collection of symptoms that are caused by disorders affecting the brain. The global burden of dementia is large and expected to triple by 2050 in the absence of a treatment (Patterson, 2018). The application of deep learning to early detection and automated classification of dementia has recently gained considerable attention (Jo et al., 2019; Reuben et al., 2017), as rapid progress in neuroimaging techniques has generated large-scale multimodal neuroimaging data. The ADReSS challenge (Luz et al., 2020) released a benchmark dataset of spontaneous speech, which is acoustically pre-processed and balanced in terms of age and gender, defining a shared task through which different approaches to dementia recognition in spontaneous speech can be compared, several speech classification models were used for dementia detection, in which different types of linguistic features were extracted and fed into traditional statistical models. This study is an interesting proof of concept, with fewer than 100 patients. More recent studies (Calzà et al., 2021; Farzana et al., 2022) measured the impact of linguistic features (e.g. verbal disfluency tags) on dementia detection.

Dementia can be an underlying cause of hospital admissions, for example due to increased rates of falls in dementia sufferers. However the diagnosis associated with the admission will be a fracture, rather than dementia. In this paper, we test the possibility of early detection for dementia patients based on the in-hospital clinical notes. Specifically, we collected 954 patients' clinical notes from Melbourne Frankston hospital [1] and assign dementia/non-dementia/uncertainty labels. Given the labeled dementia data sets, we develop a deep learning model based on ClinicalBioBERT (Alsentzer et al., 2019). We experiment with both the binary (dementia/non-dementia) and the coarse (dementia/non-dementia/uncertainty) settings, and find that ClinicalBioBERT works well in the binary setting but performs poorly on the coarse setting, at the same time it still suffers from the low annotation problem, and the embedding representation is not effective as the structured representation (e.g. UMLS concept representation). "Poor" in this context is in comparison to traditional statistical and machine learning classifiers (not discussed here). Our main contributions are:

- We collected clinical text from a local hospital and provided a labeled data set for dementia detection.
- We developed a deep neural model based on ClinicalBioBERT and evaluated its performance on both binary and three-class coarse level dementia prediction, suggesting it works

[1] https://www.peninsulahealth.org.au/locations/frankston/

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 150–155
July 14, 2023

well on the binary setting but performs poorly on the coarse setting.

- We analyzed the representation power of the fine tuned ClinicalBioBERT, and find that the UMLS concept representation is stronger than the embedding one. As far as we are aware, our work is the first application to leverage deep models and clinical notes for dementia disease classification.

2 Dementia Dataset Construction

In this section, we describe how we collected the medical notes and acquired the gold-standard labels, we also show some of the basic data statistics.

2.1 Dataset collection and labeling

We recruited patients from two sources: i) a Cognitive Dementia and Memory Service (CDAMS) and ii) random selection based on attendance at the local health service. Patients attending CDAMS were split into two groups: those with a clinical diagnosis of dementia (*1a*) and those without (*1b*). Patients in group *1b* may have received a different diagnosis or not completed their assessment. Patients in group 2 were screened with the Telephone Interview for Cognitive Status (TICS-M, Australian version), with those scoring in the population average or better band after adjustment for age, sex and education (cohort *2a*) considered as free of dementia and those scoring below the average considered as uncertain (cohort *2b*). We collected documents from the in-patient electronic health record for a total of 954 patients. Table 1 shows the number of patients in each cohort. It can be seen there are much more patients in cohort *1b*, which is around half of all the patients. Also, we notice cohort 1 has more than two times patients than cohort 2, this imbalance may have some impact on later model development and cause low specificity issues.

2.2 Dataset statistics

Document Types There are various document types for the patients, including but not limited to patient demographics, medications, vital signs, past medical history description, radiology report and progress note. We noticed that the progress notes were the majority (24.89%) types for the patients, as shown in Table 1, patients in cohort 1 had more than 2% progress notes than that in cohort 2. The *1a* cohort has largest number of progress notes (26.61%), which is reasonable as those patients may have more times of visits than other groups.

Document Counts and Length We also calculated the statistics of document counts and length for each cohort. As shown in Table 2, document counts for patients from different cohorts vary significantly, while the document average lengths from the four cohorts is more or less similar. More specifically, patients in cohort 1 tend to have around 4 times as many documents (283) as those in cohort 2a (66). Patients in cohort 2a had fewer documents, because the randomly selected patients were usually less complex and had fewer admissions than cohort 1. However, we cannot use document count as an input feature for later statistical modelling as other complex disorders are likely to have similar document counts to dementia patients.

An example We show a progress note with demographic information removed for a patient from cohort *1a* as the following: *[Progress Note: Pt ambulated to toilet Independently with x1 assist While coming out from Toilet ,pt become agitated and aggressive towards author T/L involved and pt stating his Meds is not given Though Writer mentioned this matter to Treating Dr earlier, couldn't chart the meds as he hasn't had the list of meds Informed to Treating Dr and NIC Contacted wife over the phone and treating Dr spoke to her Pt become calm and has had Meds as per MAR.]* We notice three important characteristics for such clinical text: i) abbreviations, ii) spelling errors - clinical staff complete documents under time pressure and spelling errors are common. iii) Long distance context, as the history notes may also needed to give full interpretation for the current text.

3 Methodology

In this section, we describe the development of classification models using the data sets described above. There are two aspects to consider before the model development. First, what is the classifier's granularity? There could be three levels of input, i.e., sentence, document and patient (multi-document) level. Typically, it is more challenging to achieve high performance when the input text is longer. However, developing a sentence or single document level classification model requires fine grained annotation, which is often time consuming and expensive in the medical setting. Meanwhile,

Cohort	Description	Patient counts	Progress note Pct.
1a	*Diagnosed as dementia in CDAMS, Positive*	245	26.61%
1b	*No final diagnosis in CDAMS, Uncertain*	419	24.10%
2a	*Diagnosed as non-dementia via TICS, Negative*	196	22.94%
2b	*No diagnosis via TICS, Uncertain*	99	23.02%
Total	-	959	24.89%

Table 1: Patient statistics

Cohort	Max	Min	Mean	Std	Median
1a	2150 (4750)	1 (170)	283 (871)	356 (319)	125 (847)
1b	3770 (4678)	1 (239)	241 (898)	410 (308)	79 (887)
2a	737 (2586)	1 (393)	66 (813)	90 (269)	34 (775)
2b	347 (2829)	1 (400)	74 (893)	79 (340)	48 (822)
All	3770 (4750)	1 (170)	199 (873)	340 (309)	67 (844)

Table 2: The statistics for document counts (document length) in each cohort.

multi-instance learning may further improve the complexity in the prediction stage. Therefore, we aim to develop a patient level classification model directly. Second, what types of Machine Learning models can be used? We consider the recent deep neural models (e.g. BERT), but since BERT is pre-trained from generic text, we will fine tune a ClinicalBioBERT (Alsentzer et al., 2019) [2] due to its domain similarity.

Clinical Note Filtering and Compression For the BERT based classification model, we choose ClinicalBioBERT as the pre-trained LM, and fine tune it with the medical text from each patient. However, as most BERT based models can only take 512 tokens as the maximum input, it is necessary to compress each patient's notes within that length. We consider several strategies: The first one is to filter out the notes where there are structured notes, as these structure information are often progress notes and not disease specific. The second strategy is to annotate some key sentences and build a sentence level classifier, and use the classifier to filter and shorten the clinical notes. However, it is expensive and requires further human annotation. The third strategy is truncation based on the latest notes, as in table 2, we show the average clinical note length for all patients is 873, in our text pre-processing stage we notice there are at least 10 annotated UMLS concepts for a clinical note if the number of tokens in it is over 873. Therefore, we use a simple and realistic heuristic by keeping those medical notes in which there are at least 10 UMLS concepts, and aggregate the latest notes to represent the patient note summary.

Fine tune ClinicalBioBERT After getting the patient note summary, we pair those summaries with their cohort labels and fine tune ClinicalBioBERT. We add the [CLS] token at the beginning of the patient note summary and use it as the hidden representation. During fine tuning, we update all the transformer layers and use Adam(Kingma and Ba, 2014) as the optimizer.

4 Experiments

Experiment Setup We experimented with two classification schemes: binary and three-class. For the binary case, only the medical notes for patients from cohorts *1a* and *2a* were selected, which is an exact binary classification (1a v.s. 2a) setting. In contrast, in the three-class setting, we regard the *1b* and *2b* cohorts as the uncertainty group, which returns the three class (1a v.s. 2a v.s. uncertain) setting. For the ClinicalBioBERT model, we keep the default settings and trained 20 epochs until convergence. Like other biomedical settings, we use accuracy, precision, recall and Micro F1 as the evaluation metrics. We also add a keyword method as the naive baseline, where we use a pre-recognized 245 UMLS concept names as a keyword list, these concepts are recognized by human experts to correlated with dementia. If any of those concept names appear in the document, we give a prediction of

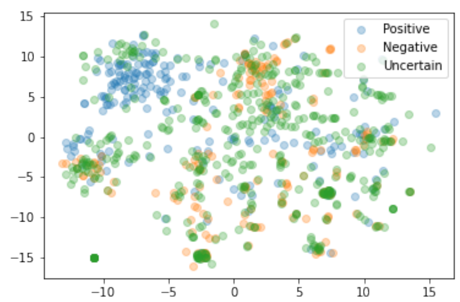
[2]The ClinicalBioBERT model was trained on all notes from MIMIC-III, a database containing electronic health records from ICU patients at the Beth Israel Hospital in Boston, MA. Model can be found from https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT

Models	Accuracy	Precision	Recall	F1
Keyword-based (binary)	0.453 (0.021)	0.469 (0.023)	0.482 (0.025)	0.475 (0.021)
Keyword-based (coarse)	0.398 (0.051)	0.382 (0.052)	0.393 (0.043)	0.387 (0.052)
ClinicalBioBERT (binary)	**0.810 (0.041)**	**0.832 (0.051)**	**0.801 (0.052)**	**0.814 (0.050)**
ClinicalBioBERT (coarse)	0.458 (0.045)	0.449 (0.043)	0.406 (0.046)	0.381 (0.045)

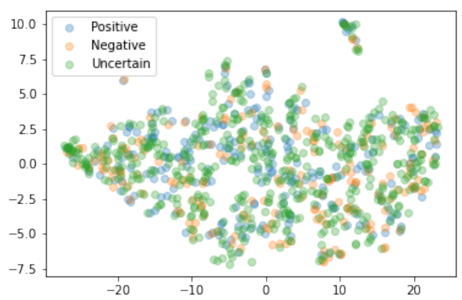
Table 3: Binary/coarse dementia classification results for the fine Tuned ClinicalBioBERT model, the binary classification are for (1a vs 2a), the coarse classification are for (1a vs 2a vs 1b, 2b) The numbers in the parenthesis show the standard deviation of the ten runs.

positive dementia, otherwise negative.

Results We perform 10-fold cross validation on the selected patients given the binary and three class setting. Table 3 shows the key results. In general, we find that the fine tuned ClinicalBioBERT performs well in the binary setting with a 0.81 accuracy, but it dropped significantly in the three class coarse setting. Meanwhile, it is shown that the performance of both models decreased around 20% from the binary to the three class setting.



(a) t-SNE for UMLS concept representation



(b) t-SNE for ClinicalBioBERT representation

Figure 1: We apply t-SNE for the 954 patients' feature representation with (a) 245 UMLS concepts and (b) the [CLS] embedding of the note summaries. In general, the UMLS concept representation distinguish positive and negative dementia patients better.

5 Analysis

Even though BERT based models show superior performance on most generic text classification tasks, the fine tuned ClinicalBioBERT does not exhibit satisfied results in the coarse setting in this study. We anticipate three reasons: First, the clinical notes are too long for ClinicalBioBERT to encode, since the standard BERT models can only take an input length of 512 tokens. Meanwhile, the dementia related text spans are quite sparse in the clinical notes, further text compression and selection heuristics are required. Furthermore, the BERT based modeling techniques cannot leverage expert prior knowledge, which in this study are the filtered UMLS concepts. To validate our hypothesis, We apply t-SNE for the 954 patients' feature representation with either the 245 UMLS concepts or the [CLS] embedding of the clinical note summary. As shown in figure 1 (a), the UMLS concept representation is more meaningful, as those positive dementia patients can be easily separated with those negative patients, while in figure 1 (b) there is no clear representation patterns for these three classes.

6 Related Work

Clinical text representation and classification When clinical text classification is used for disease detection tasks, it varies a lot from generic text classification: (i) Traditional text classification tasks take both precision and recall as the system measurement, while recall is considered to be top priority in most medical text classification tasks (Spasic et al., 2020) because doctors would never like to miss the information of any "likely" infected patients. That is, the system is being used to screen latent potential candidates. (ii) Annotation cost is higher in the medical domain (Wei et al., 2019) because professional skills from medical experts are needed. In common text annotation tasks, it is not necessary to hire highly skilled people and even crowd sourcing can be used. (iii) The text in the medical domain contains a lot of abbreviations, jargon and acronyms for different medical concepts (Xu et al., 2007). (iv) There are patient records which are sequential and correlated within each other. A patient can have multiple reports, in which each report is the description of a specific time period. The classification for these multiple reports varies based on time, so there should be some level consistency. ClinicalXLNet (Huang et al., 2019) was recently developed to model such sequential clinical text. (v) Medical text for a patient can come from different sources (Yang and Wu, 2021) such as CT scans, blood scans and operation reports, etc.

Disease detection with NLP Before this study, we have previously explored automatic fungal disease detection with radiology reports (Liu et al., 2016, 2017; Baggio et al., 2019) and showed the effectiveness of various NLP models on clinical notes. Even though deep learning has revolutionized the ML applications, Sheikhalishahi et al. (2019) reviewed the ML models on chronic diseases with clinical notes and showed that more than 90% of the methods still relied on statistical models. Wang et al. (2020) conducted a systematic evaluation of NLP in medicine over the past 20 yeears, they showed that cancer (24.94%) was the most common subject area in NLP-assisted medical research on diseases, with breast cancers (23.30%, 24/103) and lung cancers (14.56%) accounting for the highest proportions of studies.

Dementia detection The application of deep learning to early detection and automated classification of dementia has recently gained considerable attention (Jo et al., 2019), as rapid progress in neuroimaging techniques has generated large-scale multimodal neuroimaging data. The ADReSS challenge (Luz et al., 2020) released a benchmark dataset of spontaneous speech, which is acoustically pre-processed and balanced in terms of age and gender, defining a shared task through which different approaches to dementia recognition in spontaneous speech can be compared. More recently, Farzana et al. (2022) measured the impact of verbal disfluency tags on denmentia detection.

Biomedical language models Most biomedical language models are pre-trained with BERT (Devlin et al., 2018) and related clinical text. For example, the ClinicalBioBERT (Alsentzer et al., 2019) model was trained on all notes from MIMIC-III (Johnson et al., 2016), a database containing electronic health records from ICU patients at the Beth Israel Hospital in Boston, MA. MedBERT (Rasmy et al., 2021) was pretrained on a structured EHR dataset of 28,490,650 patients.

7 Conclusion

In this work, we collected clinical text from a local hospital and leveraged deep neural models for dementia detection. We fine tuned a ClinicalBioBERT and evaluated its their performance on dementia classification, experiment results showed that the fine tuned model works well on binary dementia classification but fails on three class dementia classification. As for the future work, we will leverage more human prior knowledge and experiment with both statistical and deep neural models. Also, more structured patient representation using knowledge graphs will be considered.

8 Limitation

There are a few limitations of this study: First, the patient sample size for the validation cohorts was limited to 954 patients from a local hospital. As annotation in the medical setting is expensive and time consuming, we only get patient level labels and cannot pay the effort for document level annotations. The size and diversity of the data sample could be improved by collecting clinical notes for patients from other hospitals in different age groups and of similar clinical complexity. We did not perform cross label check for the sampled patients, as there is a large number of uncertain patients, among those patients there are still ones who suffer from dementia but not diagnosed. Second, more statistical models can bee developed. At the moment we only tried a keyword based model and a deep neural models. Traditional statistical models like Logistic Regression with biomedical concept features can also be considered. Furthermore, our study would have benefited from more model interpretability and human error analysis on the classifier predictions. We have plans to extend our current work with the above mentioned directions.

Acknowledgements

This work was supported by the Australian Medical Research Future Fund (RRDHI000088).

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Diva Baggio, Trisha Peel, Anton Y Peleg, Sharon Avery, Madhurima Prayaga, Michelle Foo, Gholamreza Haffari, Ming Liu, Christoph Bergmeir, and Michelle Ananda-Rajah. 2019. Closing the gap in surveillance and audit of invasive mold diseases for antifungal stewardship using machine learning. *Journal of Clinical Medicine*, 8(9):1390.

Laura Calzà, Gloria Gagliardi, Rema Rossini Favretti, and Fabio Tamburini. 2021. Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia. *Computer Speech & Language*, 65:101113.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Shahla Farzana, Ashwin Deshpande, and Natalie Parde. 2022. How you say it matters: Measuring the impact of verbal disfluency tags on automated dementia detection. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 37–48.

Kexin Huang, Abhishek Singh, Sitong Chen, Edward T Moseley, Chih-ying Deng, Naomi George, and Charlotta Lindvall. 2019. Clinical xlnet: Modeling sequential clinical notes and predicting prolonged mechanical ventilation. *arXiv preprint arXiv:1912.11975*.

Taeho Jo, Kwangsik Nho, and Andrew J Saykin. 2019. Deep learning in alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data. *Frontiers in aging neuroscience*, 11:220.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ming Liu, Gholamreza Haffari, and Wray Buntine. 2016. Learning cascaded latent variable models for biomedical text classification. In *Proceedings of the Australasian Language Technology Association Workshop 2016*, pages 128–132.

Ming Liu, Gholamreza Haffari, Wray Buntine, and Michelle Ananda-Rajah. 2017. Leveraging linguistic resources for improving neural text classification. In *Proceedings of the australasian language technology association workshop 2017*, pages 34–42.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2020. Alzheimer's dementia recognition through spontaneous speech: the adress challenge. *arXiv preprint arXiv:2004.06833*.

Christina Patterson. 2018. World alzheimer report 2018.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):1–13.

David B Reuben, Andrew S Hackbarth, Neil S Wenger, Zaldy S Tan, and Lee A Jennings. 2017. An automated approach to identifying patients with dementia using electronic medical records. *Journal of the American Geriatrics Society*, 65(3):658–659.

Seyedmostafa Sheikhalishahi, Riccardo Miotto, Joel T Dudley, Alberto Lavelli, Fabio Rinaldi, Venet Osmani, et al. 2019. Natural language processing of clinical notes on chronic diseases: systematic review. *JMIR medical informatics*, 7(2):e12239.

Irena Spasic, Goran Nenadic, et al. 2020. Clinical text data in machine learning: systematic review. *JMIR medical informatics*, 8(3):e17984.

Jing Wang, Huan Deng, Bangtao Liu, Anbin Hu, Jun Liang, Lingye Fan, Xu Zheng, Tong Wang, Jianbo Lei, et al. 2020. Systematic evaluation of research progress on natural language processing in medicine over the past 20 years: bibliometric study on pubmed. *Journal of medical internet research*, 22(1):e16816.

Qiang Wei, Yukun Chen, Mandana Salimi, Joshua C Denny, Qiaozhu Mei, Thomas A Lasko, Qingxia Chen, Stephen Wu, Amy Franklin, Trevor Cohen, et al. 2019. Cost-aware active learning for named entity recognition in clinical text. *Journal of the American Medical Informatics Association*, 26(11):1314–1322.

Hua Xu, Peter D Stetson, and Carol Friedman. 2007. A study of abbreviations in clinical notes. In *AMIA annual symposium proceedings*, volume 2007, page 821. American Medical Informatics Association.

Bo Yang and Lijun Wu. 2021. How to leverage multimodal ehr data for better medical predictions? *Conference on Empirical Methods in Natural Language Processing*.

Automated Orthodontic Diagnosis from a Summary of Medical Findings

Takumi Ohtsuka†, Tomoyuki Kajiwara†, Chihiro Tanikawa‡
Yuujin Shimizu‡, Hajime Nagahara‡, Takashi Ninomiya†
†Ehime University ‡Osaka University
{ohtsuka@ai., kajiwara@, ninomiya@}cs.ehime-u.ac.jp
{ctanika@dent, yjnshimizu@dent, nagahara@ids}.osaka-u.ac.jp

Abstract

We propose a method to automate orthodontic diagnosis with natural language processing. It is worthwhile to assist dentists with such technology to prevent errors by inexperienced dentists and to reduce the workload of experienced ones. However, text length and style inconsistencies in medical findings make an automated orthodontic diagnosis with deep-learning models difficult. In this study, we improve the performance of automatic diagnosis utilizing short summaries of medical findings written in a consistent style by experienced dentists. Experimental results on 970 Japanese medical findings show that summarization consistently improves the performance of various machine learning models for automated orthodontic diagnosis. Although BERT is the model that gains the most performance with the proposed method, the convolutional neural network achieved the best performance.

1 Introduction

To make a proper orthodontic diagnosis, dentists need a wealth of knowledge and experience. Therefore, inexperienced dentists may overlook patient problems. Artificial intelligence technologies, such as automatic diagnosis, are promising for preventing such errors by dentists (Shimizu et al., 2022). Even for experienced dentists, automatic diagnosis technology can contribute in terms of workload reduction and improved efficiency. Therefore, this study focuses on automatic diagnosis from medical findings texts written by dentists.

While computer vision technologies for orthodontic applications, such as landmark identification from cephalometric X-rays (Kunz et al., 2020) and tooth segmentation on 3D dental surfaces captured by intraoral scanners (Lian et al., 2019), have been actively studied, research on natural language processing (NLP) technologies in this area has been limited. The only previous work (Shimizu et al., 2022) applying NLP to automatic diagnosis from medical findings relied on bag-of-words (BoW) feature extraction and support vector machine (SVM) classification without the benefit of deep-learning (DL), which has been successfully applied to a variety of tasks in recent years. We hypothesize that this is due to the frequent use of technical terms not covered by even powerful pre-trained models, as well as the long documents in which the medical findings are written in an inconsistent style. Specifically, the medical findings in this study average 1,886 tokens, with a maximum of 6,379 tokens, and contain many incomplete sentences, such as bullet points.

To solve the problems of document length and style inconsistency, we utilize a short summary of medical findings written by dental specialists for automatic diagnosis. In contrast to the original medical findings, our summary is about 90% shorter (179 tokens on average) and consists only of complete sentences. With these advantages, summary text facilitates feature extraction from documents by encoders in DL models.

To evaluate the effectiveness of automatic diagnosis from a summary of medical findings, we experimented with 970 Japanese medical findings. Experimental results on DL models of recurrent neural networks (RNN), convolutional neural networks (CNN), self-attention networks (SAN), and BERT (Devlin et al., 2019), a pre-trained SAN with masked language modeling objectives, showed that the CNN model achieved the best performance. Furthermore, the performance of the SVM and DL models was consistently improved when utilizing the summaries compared to the original medical findings. In particular, BERT was the best performance improvement with the proposed method.

2 Related Work

With the development of DL technologies, many medical applications including the field of orthodontics are being addressed.

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 156–160
July 14, 2023 ©2023 Association for Computational Linguistics

2.1 Medical Applications of NLP

One of the medical applications of NLP is automated question-answering (QA) (Nguyen, 2019). emrQA[1] (Pampari et al., 2018) and emrKBQA[2] (Raghavan et al., 2021) are large-scale corpora automatically created from English electronic medical records for QA in the clinical domain.

The MedWeb task in NTCIR-13[3] (Wakamiya et al., 2017) targeted user-generated text on social media, which is more accessible than electronic medical records. This competition addressed disease classification in three languages: English, Chinese, and Japanese.

In recent years, medical language processing related to COVID, a worldwide epidemic, has also been actively studied. Examples include COVID-QA[4] (Möller et al., 2020) for question answering, COVID-Q[5] (Wei et al., 2020) for question classification, and COVID-19 Real World Worry Dataset[6] (Kleinberg et al., 2020) for emotional analysis.

As described above, medical applications of NLP are being studied in a variety of languages and tasks. However, there are few efforts to apply NLP in the field of orthodontics. In particular, there is no application of DL-based NLP other than this study.

2.2 Orthodontic Applications of DL Models

Many applications of deep learning models in the field of orthodontics are computer vision technologies. Kunz et al. (2020) utilized CNN models to identify landmarks in cephalometric X-rays. They reported that training with 1792 images resulted in CNN models achieving nearly the same quality as experienced examiners. Lian et al. (2019) proposed MeshSNet, which performs tooth segmentation on 3D dental surfaces captured by intraoral scanners.

While DL-based computer vision models have been actively applied to the field of orthodontics, there is no previous study of DL-based NLP. The only previous work applying NLP to the field of orthodontics (Shimizu et al., 2022) has addressed automatic diagnosis from medical findings. They perform automatic diagnosis by feature extraction with BoW and classification with SVM, and do not benefit from recent DL technologies.

3 Proposed Method

We utilize deep learning-based document classification models for automatic diagnosis from medical findings in the field of orthodontics. We first describe these models in Section 3.1. Since medical findings are difficult to classify as they are, our proposed method instead utilizes a short summary of them, written in a consistent style. This is explained in Section 3.2.

3.1 DL-based Document Classification Models

In recent years, DL-based models have been widely used in NLP tasks, including document classification. In this study, we apply four types of DL models, including RNN, CNN, SAN, and BERT (Devlin et al., 2019), for automatic diagnosis from medical findings in the field of orthodontics.

Recurrent Neural Network (RNN) is one of the neural networks that deal with time-series data by recursively processing input data. In NLP, sentences are segmented into tokens as a preprocessing step, and the RNN processes the tokens in order from the beginning of the sentence. Since the original RNN is not good at long-term memory, extensions to BiLSTM, which uses LSTM cells and receives sentences in both directions, have improved performance on many tasks such as pronoun prediction (Stymne et al., 2017) and dependency parsing (Falenska and Kuhn, 2019). Nevertheless, it is difficult to achieve high performance for a very long series exceeding 1,000 tokens (Li et al., 2018). This study also employs the BiLSTM model as an RNN.

Convolutional Neural Network (CNN) is one of the neural networks that utilizes convolutional filters and pooling layers to extract features from input data as regions rather than points. While this is a model typically used for computer vision, it is also known to be effective in NLP, such as text classification (Kim, 2014). Instead of convolving n neighboring pixels as a region in computer vision, CNN acquires an n-gram representation by convolving n continuous tokens in NLP.

Self Attention Network (SAN) is another neural network that deals with series by learning contextualized token embeddings instead of aggregating

[1] https://github.com/panushri25/emrQA/
[2] https://github.com/emrQA/emrKBQA
[3] http://research.nii.ac.jp/ntcir/permission/ntcir-13/perm-en-MedWeb.html
[4] https://github.com/deepset-ai/COVID-QA
[5] https://github.com/JerryWei03/COVID-Q
[6] https://github.com/ben-aaron188/covid19worry

information like RNN and CNN. It was originally proposed as an encoder-decoder neural network for machine translation (Vaswani et al., 2017), but is also used for text classification with an encoder only. BERT (Devlin et al., 2019), which pre-trained SAN model for the objective of masked language modeling with a large-scale corpus, has remarkable performance on a number of NLP tasks through fine-tuning on the target task. Furthermore, models such as SciBERT (Beltagy et al., 2019) in the scientific domain and ClinicalBERT (Alsentzer et al., 2019) in the medical domain, which are pre-trained to focus on the desired domain, achieve even higher performance on specific tasks. Unfortunately, there are no pre-trained masked language models specific to the orthodontic domain, thus we utilize a SAN model that is trained from scratch and a general-purpose BERT that is pre-trained on Wikipedia. Note that BERT is limited to a maximum of 512 input tokens to balance memory usage and performance.

3.2 Utilizing a Summary of Medical Findings

Although the DL models described in the previous section are widely used in recent NLP tasks, it is difficult to handle texts longer than 1,000 tokens due to the difficulty of learning extreme long-term dependencies and limitations in memory usage. The medical findings in the field of orthodontics that we deal with in this study are very long documents, with an average of 1,886 tokens and a maximum of 6,379 tokens, as shown in Table 1. This is too long a text to be handled by BiLSTM or BERT.

Not only do DL models suffer from the text length, but also from inconsistencies in the writing style of the medical findings. These writing styles vary for each dentist who writes. Writing style issues include the use of incomplete sentences with bullets and indentation with spaces and tabs. While they improve visual clarity for human readers, they are noise to NLP models because such information is removed in pre-processing steps in many cases. Especially for pre-trained models such as BERT, these incomplete sentences, with different characteristics from the pre-training corpus, may seriously impair performance.

To address these problems, we propose to utilize its short summary with a consistent writing style instead of the original medical findings. These summaries are manually written by experienced dentists and are written in complete sentences without the use of bullets, indentations, or other decorations. As shown in Table 1, these summaries consist of an average of 179 tokens, that is, about 10% of the length of the original medical findings. Furthermore, even the longest summaries are not affected by the limit on the maximum length of input tokens in BERT. We assigned these summaries to all medical findings in our dataset. Compared to the noisy and lengthy original medical findings, these summaries are expected to improve the performance of DL-based document classification models.

	Original	Summary
Avg.	1,886	179
Max.	6,379	467
Min.	312	61

Table 1: Number of tokens for each document.

4 Evaluation

To evaluate the effectiveness of the proposed method, automatic diagnosis is performed from the original medical findings or a summary of them, and their performance is compared. We treat this task as a multi-label document classification.

4.1 Setting

In our experiment, we use documents of medical findings in Japanese for 970 patients who visited for orthodontic treatment. This dataset includes the text of the medical findings written by the dentist in charge, as well as the patient's facial and X-ray images. However, utilizing these images remains our future work, and we only use text in this study. We have assigned a short summary, described in Section 3.2, to every medical finding in this dataset. Each medical finding is also assigned multiple labels corresponding to the patient's medical condition. There are a total of 322 labels, with each patient having an average of 12 labels.

As text preprocessing, line feed characters were removed and full-width alphanumeric characters were normalized to half-width. We used Sudachi[7] (Takaoka et al., 2018) for word segmentation, except for the BERT model, for which a specific subword segmenter is provided. For evaluation, we used 5-fold cross-validation. The evalua-

[7] https://github.com/WorksApplications/SudachiPy

	RNN	CNN	SAN	BERT
Number of dimensions of embedding layer	256	256	256	768
Number of dimensions of hidden layers	256	256	512	768
Number of hidden layers	1	1	2	12
Dropout rate	0.2	0.2	0.1	0.1
Batch size	64	64	16	32

Table 2: Hyperparameters of deep learning models.

	Type of medical findings	
Models	Original	Summary
BoW+SVM	0.41	0.46
RNN	0.20	0.31
CNN	**0.44**	**0.48**
SAN	0.29	0.38
BERT	0.27	0.43

Table 3: Experimental results (F1-score).

tion metric used was the F1-score.

For the document classification model, we evaluate four DL models, including RNN, CNN, SAN, and BERT[8] (Devlin et al., 2019), described in Section 3.1, as well as SVM used in the previous work (Shimizu et al., 2022). The baseline model, denoted as BoW+SVM, employs the Binary Relevance method (Tsoumakas and Katakis, 2007) to train a binary classification for each label and utilizes the RBF kernel for SVM.[9] Our DL models use Adam (Kingma and Ba, 2015) as an optimizer. Other hyperparameters are listed in Table 2.

4.2 Result

Table 3 shows the experimental results. Deep learning models other than CNN suffer from document length and style inconsistencies, resulting in significantly poorer performance than the existing model of BoW+SVM.

When a short document summarized by the dentist is used in place of the original medical findings, the F1-scores for all models consistently improve. Notably, the performance of the pre-trained BERT has improved the most substantially. We believe this is due to the use of complete sentences that are consistent with the pre-training corpus and the elimination of information lost owing to the constraint of the maximum sentence length. These experimental results show that short summaries written in a consistent style are effective in improving the performance of automatic diagnosis in the field of orthodontics.

We found that the CNN model achieved the best performance for both the original document and summary inputs. Since medical findings often contain technical terms consisting of multiple tokens, we believe that a CNN model capable of capturing n-gram features through convolution would be suitable for this task.

5 Conclusion

In this study, we improved the performance of automatic diagnosis in orthodontic treatment by utilizing a short document that was manually summarized from medical findings by dentists. Experimental results on Japanese datasets show that the proposed method consistently improves the performance of various DL models. Among them, our CNN model outperformed the existing model and updated the state-of-the-art performance.

Our future work includes the automatic generation of summaries and the development of multimodal automatic diagnosis taking into account image information. Although this study utilized summaries of medical findings manually generated by experienced dentists, there is a substantial cost to creating such a dataset. It is desirable to develop an automatic diagnostic system that reduces the workload on dentists by automatically generating summaries. In addition, our dataset includes both facial and X-ray images. This allows us to develop multimodal models that incorporate findings from the field of computer vision, which are actively studied. Multimodal automatic diagnostic systems that combine both image and linguistic information in a complementary manner are expected to have higher performance.

[8] https://huggingface.co/cl-tohoku/bert-base-japanese-v2

[9] https://scikit-learn.org/

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3615–3620.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Agnieszka Falenska and Jonas Kuhn. 2019. The (Non-)Utility of Structural Features in BiLSTM-based Dependency Parsers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 117–128.

Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference for Learning Representations*.

Bennett Kleinberg, Isabelle van der Vegt, and Maximilian Mozes. 2020. Measuring Emotions in the COVID-19 Real World Worry Dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Felix Kunz, Angelika Stellzig-Eisenhauer, Florian Zeman, and Julian Boldt. 2020. Artificial Intelligence in Orthodontics. *Journal of Orofacial Orthopedics / Fortschritte der Kieferorthopädie*, 81(1):52–68.

Shuai Li, Wanqing Li, Chris Cook, Ce Zhu, and Yanbo Gao. 2018. Independently Recurrent Neural Network (IndRNN): Building a Longer and Deeper RNN. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition*.

Chunfeng Lian, Li Wang, Tai-Hsien Wu, Mingxia Liu, Francisca Durán, Ching-Chang Ko, and Dinggang Shen. 2019. MeshSNet: Deep Multi-scale Mesh Feature Learning for End-to-end Tooth Labeling on 3d Dental Surfaces. In *Proceedings of the 22th International Conference on Medical Image Computing and Computer Assisted Interventions*, pages 837–845.

Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. 2020. COVID-QA: A Question Answering Dataset for COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Vincent Nguyen. 2019. Question Answering in the Biomedical Domain. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 54–63.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A Large Corpus for Question Answering on Electronic Medical Records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368.

Preethi Raghavan, Jennifer J Liang, Diwakar Mahajan, Rachita Chandra, and Peter Szolovits. 2021. emrKBQA: A Clinical Knowledge-base Question Answering Dataset. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 64–73.

Yuujin Shimizu, Chihiro Tanikawa, Tomoyuki Kajiwara, Hajime Nagahara, and Takashi Yamashiro. 2022. The Validation of Orthodontic Artificial Intelligence Systems That Perform Orthodontic Diagnoses and Treatment Planning. *European Journal of Orthodontics*, 44(4):436–444.

Sara Stymne, Sharid Loáiciga, and Fabienne Cap. 2017. A BiLSTM-based System for Cross-lingual Pronoun Prediction. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 47–53.

Kazuma Takaoka, Sorami Hisamoto, Noriko Kawahara, Miho Sakamoto, Yoshitaka Uchida, and Yuji Matsumoto. 2018. Sudachi: A Japanese Tokenizer for Business. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 2246–2249.

Grigorios Tsoumakas and Ioannis Katakis. 2007. Multi-Label Classification: An Overview. *International Journal of Data Warehousing and Mining*, 3(3):1–13.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000—-6010.

Shoko Wakamiya, Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, and Eiji Aramaki. 2017. Overview of the NTCIR-13 MedWeb Task. In *In Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies*, pages 40–49.

Jerry Wei, Chengyu Huang, Soroush Vosoughi, and Jason Wei. 2020. What Are People Asking about COVID-19? A Question Classification Dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Harnessing the Power of BERT in the Turkish Clinical Domain: Pretraining Approaches for Limited Data Scenarios

Hazal Türkmen, Oğuz Dikenelli
Department of Computer Engineering, Ege University, 35100 Izmir, Türkiye
{hazal.turkmen, oguz.dikenelli}@ege.edu.tr

Cenk Eraslan, Mehmet Cem Çallı, Süha Süreyya Özbek
Department of Radiology, Ege University, 35100 Izmir, Türkiye
{cenk.eraslan, cem.calli, sureyya.ozbek}@ege.edu.tr

Abstract

Recent advancements in natural language processing (NLP) have been driven by large language models (LLMs), thereby revolutionizing the field. Our study investigates the impact of diverse pre-training strategies on the performance of Turkish clinical language models in a multi-label classification task involving radiology reports, with a focus on overcoming language resource limitations. Additionally, for the first time, we evaluated the simultaneous pre-training approach by utilizing limited clinical task data. We developed four models: TurkRadBERT-task v1, TurkRadBERT-task v2, TurkRadBERT-sim v1, and TurkRadBERT-sim v2. Our results revealed superior performance from BERTurk and TurkRadBERT-task v1, both of which leverage a broad general-domain corpus. Although task-adaptive pre-training is capable of identifying domain-specific patterns, it may be prone to overfitting because of the constraints of the task-specific corpus. Our findings highlight the importance of domain-specific vocabulary during pre-training to improve performance. They also affirmed that a combination of general domain knowledge and task-specific fine-tuning is crucial for optimal performance across various categories. This study offers key insights for future research on pre-training techniques in the clinical domain, particularly for low-resource languages.

1 Introduction

Language models have undergone a significant transformation in the field of natural language processing, demonstrating exceptional capabilities in executing tasks with minimal guidance. This shift can be attributed to pivotal milestones, such as word2vec (Mikolov et al., 2013), which replaced feature engineering methods with deep learning-based representation learning. Furthermore, the emergence of contextualized word embeddings with ELMo has led to the development of (Peters et al., 1802) pre-trained transformer-based models such as BERT (Devlin et al., 2018), GPT (Radford et al., 2018), T5 (Raffel et al., 2020), and BART (Lewis et al., 2019).

Recent advancements in large language models (LLMs) have led to the development of models with parameter sizes exceeding hundred billion, including the GPT (Generative Pre-trained Transformer) series (Radford et al., 2018, 2019b,a; Ouyang et al., 2022), such as ChatGPT and GPT-4 (OpenAI, 2023), which are pre-trained on massive datasets. However, research focusing on LLMs architectures within specialized domains characterized by limited resources is scarce. A range of approaches for developing language models exists to address the issue of limited language resources, including simultaneous pretraining with in-domain data (Wada et al., 2020) and domain-adaptive pretraining by fine-tuning an existing generic language model with in-domain data (Gururangan et al., 2020). The choice of pre-training technique depends on the specific task data and available resources, but determining the optimal utilization of limited clinical task data in pretraining and selecting the most suitable data for pretraining methods remain open questions. This study aimed to assess and contrast different techniques using a limited task corpus for pretraining BERT models in the Turkish clinical domain, a low-resource setting. We introduce two pre-trained language model families, TurkRadBERT-sim and TurkRadBERT-task, each comprising two models for the clinical domain in the Turkish language. These models explore the effects of different corpus selections that combine small task-related corpora and pretraining strategies in the Turkish clinical domain. The TurkRadBERT-sim pre-trained model family, developed via simultaneous pre-training (Wada et al., 2020), involves a balanced combination of two distinct corpora: one general and one limited task-specific. Both corpora were upsampled to create pretraining instances, resulting in robust neural language models.

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 161–170
July 14, 2023

The TurkRadBERT-task pretrained model family, developed via task-adaptive pre-training, involves an additional pretraining stage where the model is adaptively pretrained on a smaller, task-specific dataset following the initial pretraining. We also created a labeled dataset for multi-label document classification using head CT radiology reports to evaluate the models. The main contributions can be listed as follows:

- While simultaneous pretraining has previously been explored with limited biomedical data in the work of (Wada et al., 2020), our study shifts the focus towards applying this approach to limited clinical Turkish radiology data for the first time. We conducted an evaluation of simultaneous pretraining, incorporating limited clinical task radiology data, and compared it with task-adaptive pretraining through continual pre-training. This novel comparison provides valuable insights into the efficacy of these methods in the context of limited clinical radiology data, highlighting their potential in specialized domains.

- We created small task-related corpora, including Turkish head CT radiology reports by Ege University Hospital. Then, we built four pretrained clinical language models, for the first time, using Turkish head CT radiology reports, Turkish general corpus, and Turkish biomedical corpara, including Turkish medical articles (Türkmen et al., 2022) and Turkish radiology theses (Türkmen et al., 2022).

- We developed a multi-label document classification task aimed at identifying the presence or absence of 12 clinically significant observations, as well as a "no findings" label indicating no observations, within head CT radiology reports for the purpose of evaluating language models. To the best of our knowledge, there are no existing multi-label document classification studies in the Turkish clinical domain.

2 Related Work

To optimize natural language processing models for specialized domains, various studies have explored different approaches to adapt general BERT models for the biomedical domain. BioBERT (Lee et al., 2020), an early attempt to adapt general BERT models to the biomedical domain, employed continual pretraining to enhance performance. Initialized from the general BERT model, BioBERT was further trained on PubMed abstracts and full-text articles, yielding an improved performance for tasks such as named entity recognition, relation extraction, and question answering. Similarly, ClinicalBERT (Alsentzer et al., 2019), a domain-specific language model, was created using continual pretraining with MIMIC data, demonstrating its effectiveness in improving clinical task performance.

Other studies have explored continual pretraining for biomedical language models, such as SciBERT (Beltagy et al., 2019) and BlueBERT (Beltagy et al., 2019), which were pretrained on a mix of biomedical and general domain corpora. An alternative approach, pretraining from scratch, focuses exclusively on in-domain data, without relying on a generic language model. This method has been effective in creating models, such as PubMedBERT (Gu et al., 2021), which is pretrained solely on PubMed abstracts. Comparisons between the two pretraining methods reveal that continual pretraining often leads to more successful transfers from general to specialized domains. For example, one study proposed four BERT models (Bressem et al., 2020), two pretrained on German radiology free-text reports (FS-BERT and RAD-BERT), and two based on open-source models (MULTI-BERT and GER-BERT). The FS-BERT model, which used the pretraining from scratch approach, performed poorly compared to the other models, suggesting that domain-specific corpora alone might be insufficient for learning proper embeddings. Another study developed RadBERT (Yan et al., 2022), a set of six transformer-based language models pretrained on radiology reports with various language models for initialization, to explore their performance in radiology NLP applications.

Although pretraining BERT models can improve performance across various biomedical NLP tasks, they require significant domain-specific data. Biomedical text data are often limited and scattered across various sources, and few publicly available medical databases are written in languages other than English. This creates a high demand for effective techniques that can work well even with limited resources. One solution to this problem is the simultaneous pre-training technique proposed in (Wada et al., 2020), which up-samples a limited domain-specific corpus and uses it for pre-training

Corpus	Size (GB)	N tokens	Domain
General Turkish Corpus	35	4,404,976,662	General
Turkish Biomedical Corpus	0,48	60,318,554	Biomedical
Turkish Electronic Radiology Theses	0,11	15,268,779	Radiology
Head CT Reports	0.036	4,177,140	Clinical Radiology

Table 1: Corpora statistics

in a balanced manner with a larger corpus. Using small Japanese medical article abstracts and Japanese Wikipedia texts, the authors created a simultaneous pretrained BERT model, ouBioBERT. The study confirmed that their Japanese medical BERT model performed better than conventional baselines and other BERT models in a medical Japanese document classification task. However, they did not focus on applying the simultaneous pre-training approach to limited clinical task radiology data. Building upon this work, our study shifts the focus towards applying the simultaneous pre-training approach to limited clinical task data for the first time. To overcome the limitations of the limited resources problem, many researchers have explored the benefits of continued pretraining on a smaller corpus drawn from the task distribution as task-adaptive pre-training (Gururangan et al., 2020; Schneider et al., 2020). In addition, (Turkmen et al., 2022) previously demonstrated that their biomedical BERT models, the BioBERTurk family, which were continuously pre-trained on a limited Turkish radiology thesis corpus, exhibited improved performance in clinical tasks. However, the authors also highlighted the potential ineffectiveness of domain incompatibility when evaluating Turkish language models, emphasizing the need for a closer alignment between domain-specific data and evaluation tasks.

3 Materials and Methods

In this section, we provide a concise overview of the pre-training methods employed for the development of Turkish clinical language models and the characteristics of the corpora used in this process. We developed four Turkish clinical language models, leveraging the BERT-base architecture and constrained language resources by employing two pre-training strategies: simultaneous pre-training and continual pre-training, referred to as task-adaptive pretraining. Two models, referred to as the TurkRadBERT-sim family, were developed by employing simultaneous pre-training techniques that combined general, biomedical, and clinical task corpora, while utilizing distinct vocabularies. In contrast, two models, the TurkRadBERT-task family, were developed by employing task-adaptive pretraining using the task corpus. To construct these clinical models, we employed four distinct corpora: the Turkish biomedical corpus compiled from open-source medical articles (Türkmen et al., 2022), Turkish electronic radiology theses corpus (Türkmen et al., 2022), Turkish web corpus (Schweter, 2020), and newly created Turkish radiology report corpus, which is a limited task corpus. While all corpora were utilized in simultaneous pre-training, only Turkish radiology reports were used in task-adaptive pre-training. Subsequently, the clinical language models were fine-tuned on a downstream NLP task within the Turkish clinical domain. Finally, the clinical language models were compared to the general Turkish domain BERT model, BERTurk (Schweter, 2020), and the BioBERTurk variant (Turkmen et al., 2022), which was continually pretrained on Turkish radiology theses.

3.1 Pre-training Strategies

The BERT framework (Devlin et al., 2018) consists of two phases: pretraining and fine-tuning. During pre-training, BERT is trained on large-scale plain text corpora, such as Wikipedia, whereas in the fine-tuning phase, it is initialized with the same pre-trained weights and then fine-tuned using task-specific labeled data, such as sentence pair classification. BERT employs two unsupervised tasks during the pre-training phase: Masked Language Model (MLM) and Next Sentence Prediction (NSP). In MLM tasks, a certain percentage of input tokens is randomly masked, and the model predicts the masked tokens in a sentence, as described in the Cloze task (Taylor, 1953). For the NSP, the model predicts whether the second sentence follows a consecutive sentence in the dataset.

In our study, we implemented several modifications to the BERT architecture for simultaneous pretrain-

ing (Wada et al., 2020), our first technique. This pre-training approach posits that training the BERT model using both large and small corpora can prevent overfitting issues caused by limited medical data. To accurately feed inputs into the model, we followed a procedure from the same study (Wada et al., 2020). We divided the small medical corpus and large general corpus into smaller documents of equal size, and combined them to create structured inputs. This approach mitigates potential overfitting resulting from the limited data size by increasing the frequency of pre-training for MLM instances containing small medical data. In accordance with the same study by (Wada et al., 2020), we utilized domain-specific generated text and the Wordpiece algorithm to generate a domain-specific vocabulary, which is referred to as amplified vocabulary in their research. Thus, we examined the impact of the domain-specific vocabulary
Simultaneous pretraining enables the model to learn language representations by training on large-scale texts. However, this approach is expensive owing to the extensive amount of data involved. Finally, we implemented the task-adaptive pretraining method (Gururangan et al., 2020) using only a small amount of clinical task data. This technique is less resource-intensive than the others. In contrast to the aforementioned pre-training methods, we developed different BERT models based on model initialization for task-adaptive pre-training, using the existing BERT vocabulary instead of creating a new one.

3.2 Data Sources for Model Development

In the development of various language models, multiple corpora were utilized to ensure that the models were well suited to the specific domain and task at hand. The selection of appropriate corpora is crucial to the performance of language models, as it directly influences their understanding of domain-specific language patterns, structures, and vocabularies. The corpora used are summarized in Table 1 and listed below:
Head CT Reports: We collected 40,306 verified Turkish radiology reports pertaining to computed tomography (CT) examinations for patients aged 8 years and above from the neurology and emergency departments at Ege University Hospital between January 2016 and June 2018. Prior to data analysis, reports containing fewer than 100 characters were excluded, and newline characters and radiology-specific encodings were removed for consistency. All text data underwent de-identification and duplicate removal. Following preprocessing, 2,000 reports were randomly selected for the head CT annotation task, and the remaining data (approximately 36 MB) was reserved for pre-training techniques.
General Turkish Corpus: This corpus, which was used in the development of the BERTurk model, contains a large collection of Turkish text data (approximately 35 GB). This serves as a foundation for training language models to understand Turkish language patterns.
Turkish Biomedical Corpus: A domain-specific corpus (Türkmen et al., 2022) consisting of full-text articles collected from Dergipark, a platform hosting periodically refereed biomedical journals in Turkey.
Turkish Electronic Radiology Theses: A unique corpus of open-domain Ph.D. theses (Türkmen et al., 2022) conducted in radiology departments of medical schools obtained from the Turkish Council of Higher Education's website.

3.3 Data preparation

The first phase after data understanding is transforming the text to the BERT-supported inputs, namely tokenization. All engineering processes to be fed into BERT were designed for Google Cloud TPUs and implemented using CPU core i8. Furthermore, Wordpiece algorithm was used to generate vocabulary for tokenization in both pre-training methods due to the success in morphologic-rich languages such as Turkish (Toraman et al., 2023). Each vocabulary config file is the same as BERTurk for a fair comparison. We implemented the tokenizer library from Huggingface [1] to build BERT's vocabulary in simultaneous pre-training. For continual pre-training, we used existing BERT's vocabulary for continual pre-training instead of creating a new one. After this process, we used *create_pretraining_data.py* script provided by the Google AI Research team [2] to convert all documents into TensorFlow examples compatible with TPU devices.

3.4 Pretraining setup

We followed BERT-base architecture consisting of 12 layers of transformer blocks, 12 attention heads,

[1] https://huggingface.co/docs/tokenizers/python/latest/
[2] https://github.com/google-research/bert

and 110 million parameters for all pre-training strategies. All models were also generated using the same hyperparameters (see Appendix B, Table 5) and were trained with open-source training scripts available in the official BERT GitHub repository using V3 TPUs with 32 cores from Google Cloud Compute Services [3].

3.5 Developed Language Models

The simultaneous pre-training technique is the first pre-training method we implemented to utilize a small in-domain corpus. Moreover, the first step in simultaneous pre-training is choosing data for small and large corpus data. We produced different TurkRadBERT-sim models according to vocabulary usage. The distinction between the two models lies in their vocabulary use; the first model leverages an amplified, domain-specific vocabulary, whereas the latter adopts the BERTurk vocabulary.

TurkRadBERT-sim v1 employed a large Turkish general corpus (35 GB) used for developing BERTurk, alongside a mixed Turkish biomedical corpus, Turkish Electronic Radiology Theses, and Turkish Head CT Reports as smaller counterparts. Excluding the data utilized for labeling (approximately 6 MB), the head CT reports were not used as a standalone small corpus for pre-training due to their limited size (30 MB) compared to other corpora. Furthermore, experimental results suggested that simultaneous training with such a data size did not yield significant outcomes in radiology report classification. To address this, we combined the small-sized corpus to match the large one, creating pre-training instances. The model also employed an amplified vocabulary, built from the generated corpus, for simultaneous pre-training.

TurkRadBERT-sim v2 was also based on the BERT-base architecture and was pre-trained simultaneously. The model used the same corpus as v1 during pretraining. The difference was that the general domain vocabulary was used to observe the effect of the domain-specific vocabulary.

The last pre-training method is task-adaptive pre-training on radiology reports (30 MB). We developed two different BERT models according to the model initialization.

TurkRadBERT-task v1 used a general domain language model for Turkish, BERTurk for model initialization and then carried out continual pre-training as a task-adaptive pre-training method. Vocabulary was also inherited from BERTurk.

TurkRadBERT-task v2 used a Turkish biomedical BERT model, BioBERTurk variant(Turkmen et al., 2022), which was further pre-trained on Turkish electronic theses for model initialization. This Turkish biomedical BERT was chosen because it achieved the best score in classification radiology reports (Turkmen et al., 2022). For tokenization, the model again inherited from the general domain.

4 Supervision Task

4.1 Multi-label CT radiology reports classification

We developed a multi-label document classification task using 2000 Turkish head CT reports mentioned in Section 3.2. This was necessary as there was no shared task for clinical documents in Turkish. Our dataset has 20618 sentences and 249072 tokens. The objective of the document level classification task is to identify the existence of clinically significant observations in a radiology report that is presented in free-text format. These are 'Intraventricular' ,'Gliosis', 'Epidural', 'Hydrocephalus', 'Encephalomalacia', 'Chronic ischemic changes', 'Lacuna', 'Leukoaraiosis', 'Mega cisterna magna' ,'Meningioma', 'Subarachnoid Bleeding', 'Subdural', 'No Findings'. The classification process involves reviewing sentences within the report and categorizing them into one of two classes: positive or negative. The 13th observation, "No Findings", indicates the absence of any findings. Those 12 labels were selected to indicate major and relatively common clinical pathologies possible to be detected in a pre-contrast cranial computerized tomography (CT) examination. Moreover, the 12 labels used in the study also are not vague radiologic findings, but definite clinical pathologies. Therefore, no hedging was performed regarding these categories radiology experts labeled the dataset at document level according to this annotation schema. The annotation process unfolded in three stages, involving three experienced radiologists (C.E, M.C.C, and S.S.O). In each stage, two annotators (C.E, M.C.C) independently labeled a portion of the reports. Subsequently, the third annotator examined these annotations to detect any discrepancies. At the conclusion of each stage, all three annotators reached a consensus by generating mutually agreed-upon annotations. A spreadsheet file was utilized to facilitate the annotation task for

[3]https://cloud.google.com/

Model	Precision	Recall	F1 Score
BERTurk	**0.9738**	0.9456	**0.9562 (± 0.0077)**
TurkRadBERT-task v1	0.9736	**0.9462**	0.9556 (± 0.0057)
BioBERTurk	0.9731	0.9440	0.9535 (± 0.0068)
TurkRadBERT-task v2	0.9643	0.9352	0.9470 (± 0.0068)
TurkRadBERT-sim v1	0.8613	0.7969	0.8149 (± 0.0214)
TurkRadBERT-sim v2	0.8170	0.7863	0.7879 (± 0.0135)

Table 2: Average Precision, recall, and F1 Score for each model. We performed ten separate runs with different random seeds and present both the average and standard deviation.

Category	BERTurk	TurkRadBERT-task v1
Intraventricular	**0.4815 (± 0.4475)**	0.4000 (± 0.3266)
Gliosis	**0.8580 (± 0.0577)**	0.8155 (± 0.1024)
Epidural	**0.9012 (± 0.0349)**	0.9000 (± 0.0333)
Hydrocephalus	0.9458 (± 0.0327)	**0.9673 (± 0.0459)**
Encephalomalacia	0.9622 (± 0.0173)	**0.9633 (± 0.0081)**
Chronic ischemic changes	0.9918 (± 0.0044)	**0.9921 (± 0.0026)**
Lacuna	**0.9655 (± 0.0000)**	**0.9655 (± 0.0000)**
Leukoaraiosis	**0.8995 (± 0.1063)**	0.8762 (± 0.1227)
Mega cisterna magna	**0.6000 (± 0.1500)**	0.4500 (± 0.0577)
Meningioma	**1.0000 (± 0.0000)**	**1.0000 (± 0.0000)**
Subarachnoid Bleeding	0.9281 (± 0.0183)	**0.9544 (± 0.0118)**
Subdural	0.9666 (± 0.0119)	**0.9757 (± 0.0081)**
No Findings	**0.9455 (± 0.0145)**	0.9311 (± 0.0167)

Table 3: Average F1 scores for each label in the TurkRadBERT-task v1 and BERTurk models. In each experiment, we carried out ten distinct runs using different random seeds, from which we determine and report the average and standard deviation.

the annotators The annotated datasets were subsequently divided randomly into test (10%), validation (10%), and training (80%) sets for fine-tuning. The class distributions, as illustrated in Appendix A, demonstrate the varying prevalence of different categories in the datasets. The datasets exhibit an imbalanced distribution, which is a typical characteristic of text processing in the radiology domain (Qu et al., 2020).

4.2 Fine-tuning Setup

The fine-tuning of all pretrained models was conducted independently utilizing identical architecture and optimization methods as previously employed in the study (Devlin et al., 2018). In the process of fine-tuning, the objective is not to surpass the current state-of-the-art performance on the downstream tasks, but rather to assess and compare pretraining techniques for developing Turkish clinical language models. So, an exhaustive exploration of hyperparameters was not utilized. Consequently, the optimal parameters identified from a limited hyperparameter search are employed, working under the assumption that the fairness of model evaluations and comparisons isn't compromised by the potential presence of more optimal hyperparameters. Hyperparameter searches were conducted for each model, examining learning rate values ϵ from the set {2e-4, 3e-5, 5e-5}, maximum sequence lengths ϵ from the set {128, 256, 512}, batch sizes ϵ from the set {16, 32}, and the number of training epochs ϵ from the set {15, 20}. Due to memory constraints, a batch size of 64 was not considered. The configurations employed for the TurkRadBERT-sim and TurkRadBERT-task models are displayed in Table 6 and Table 7 in Appendix B respectively.

The effectiveness of distinct pre-trained BERT models on the clinical multilabel classification task was evaluated by computing average precision, recall, and F1 score across ten runs, utilizing the most suitable hyperparameter settings.

5 Experimental Results

In this study, we evaluated the performance of five different models, including BERTurk, TurkRadBERT-task v1, TurkRadBERT-task v2, TurkRadBERT-sim v1, and TurkRadBERT-sim v2, for Turkish clinical multi-label classification. We compared their performance over ten runs in terms of average precision, recall, and F1 score. Additionally, we analyzed the performance of wining two model (BERTurk, TurkRadBERT-task v1) on individual categories using their respective F1 scores. The results are presented in Tables 2 and 3.

Table 2 shows that BERTurk achieves an F1 score of 0.9562, with a precision of 0.9738 and recall of 0.9456. TurkRadBERT-task v1 has a slightly lower F1 score of 0.9556 but with comparable precision (0.9736) and recall (0.9462). Both models demonstrate strong performance on the classification task, with BERTurk slightly outperforming TurkRadBERT-task v1 in terms of the overall F1 score. While BERTurk performed better than TurkRadBERT-task v1, there are no statistical differences between these models (P value 0,255). Additionally, BERTurk has also outperformed BioBERTurk. Other models, such as TurkRadBERT-task v2, TurkRadBERT-sim v1, and TurkRadBERT-sim v2, show lower overall performance compared to BERTurk, TurkRadBERT-task v1 and BioBERTurk.

However, it is essential to evaluate the models' performance for each label, as this offers a deeper understanding of their strengths and weaknesses. Table 3 presents the F1 scores for each category for BERTurk and TurkRadBERT-task v1. The results reveal that the performance of the models varies across categories, with some labels showing a noticeable difference in F1 scores between the two models. BERTurk performs better than TurkRadBERT-task v1 in the following categories: Intraventricular, Gliosis, Epidural, Leukoaraiosis, Mega cisterna magna, and No Findings. In contrast, TurkRadBERT-task v1 outperforms BERTurk in the categories of Hydrocephalus, Encephalomalacia, Chronic ischemic changes, Subarachnoid Bleeding, and Subdural. The F1 scores for Lacuna and Meningioma are identical for both models.

6 Discussion

By assessing the experiments as a whole, we derived the following conclusions. When comparing simultaneous pre-training and task-adaptive pre-training, it is observed that, owing to the size difference between the task data and the general data, the limited domain-specific data may be overshadowed by the large general-domain data. This causes the model to focus more on learning general rather than task-specific features. This phenomenon highlights the importance of carefully balancing general and domain-specific data during the pretraining process to ensure that the model effectively captures the nuances of the specialized domain.

The performances of the BERTurk and TurkRadBERT-task v1 models are quite similar because both models leverage the knowledge gained from the large general-domain corpus during pre-training. BERTurk is directly pre-trained on this large corpus, while TurkRadBERT-task v1 is initialized with BERTurk's weights and then fine-tuned using task-adaptive pre-training on a smaller clinical corpus. This fine-tuning enables TurkRadBERT-task v1 to capture domain-specific patterns, structures, and terminologies absent in the general-domain corpus.

However, the small task-specific corpus used in task-adaptive pretraining may limit the model's learning of domain-specific knowledge. Consequently, despite the benefits of task-adaptive pre-training, TurkRadBERT-task v1, which utilized this approach, had a slightly lower performance than BERTurk. In limited data scenarios, the task-adaptive pre-training approach may be prone to overfitting, especially when pre-trained on a small task-specific corpus. The model may become overly specialized in training data and fail to generalize well to unseen examples (Zhang et al., 2022).

In terms of performance, TurkRadBERT-task v1 has a slightly higher F1 score (0.9556) than BioBERTurk (0.9535) and TurkRadBERT-task v2 (0.9470). This suggests that despite the more specialized biomedical knowledge in BioBERTurk, the general-domain BERTurk model provides a more robust foundation for task-adaptive pre-training in this specific clinical task.

Another conclusion reached in this study is that comparison between TurkRadBERT-sim v1 and v2 offers insights into the impact of domain-specific vocabulary on model performance. TurkRadBERT-sim v1, which used an amplified vocabulary built from the generated corpus, outperformed TurkRadBERT-sim v2 that employed the general domain vocabulary. This finding indicates that using a domain-specific vocabulary during pre-training can enhance the ability of the model to capture and understand domain-specific language patterns, ultimately leading to improved performance on clinical NLP tasks.

Examining the F1 scores for each label in Table 3 provides a more detailed perspective of the performance of the two most successful models. First, the optimal performance on specific labels, such as meningioma and chronic ischemic changes, might be attributed to the use of precise, standard reporting terminology to define these pathologies, a factor that likely provides high results, regardless of the classifier employed. BERTurk outperforms TurkRadBERT-task v1 in certain labels, such as Intraventricular, Gliosis, Epidural, Leukoaraiosis, Mega cisterna magna, with No Findings. The higher performance of BERTurk on certain labels could be attributed to the general domain knowledge acquired during direct pre-training (different from other pre-training methods), which may provide better coverage for specific categories, particularly those with a lower frequency in the task-specific corpus. BERTurk's broader pre-training data exposure could potentially give it an advantage over models like TurkRadBERT-task v1 when dealing with specific labels that have lower representation in the task-specific corpus, even though TurkRadBERT-task v1 is initialized with BERTurk. This suggests that a combination of general domain knowledge and task-specific fine-tuning may be critical for optimal performance across diverse categories. On the other hand, TurkRadBERT-task v1 exhibits superior performance for labels like Hydrocephalus, Encephalomalacia, Subarachnoid Bleeding, and Subdural. This suggests that task-adaptive pre-training can offer a performance boost in some instances by fine-tuning the model based on domain-specific information. However, it is worth noting that the overall performance differences between the two models are relatively small, highlighting the importance of leveraging both general-domain and task-specific knowledge in these models.

7 Conclusion

This study provides a comprehensive comparison of the performance of various models, including BERTurk, TurkRadBERT-task v1, TurkRadBERT-task v2, TurkRadBERT-sim v1, and TurkRadBERT-sim v2, on a radiology report classification task. Our findings demonstrate that the BERTurk model achieved the best overall performance, closely followed by the TurkRadBERT-task v1 model. This highlights the importance of leveraging both general domain knowledge acquired during pre-training and task-specific knowledge through fine-tuning to achieve optimal performance on complex tasks.

We also observed that the performance of these models varies across different labels, with BERTurk performing better on certain categories, particularly those with lower representation in the task-specific corpus. This finding suggests that a combination of general domain knowledge and task-specific fine-tuning may be critical for achieving optimal performance across diverse categories. Additionally, it is essential to consider label frequencies when interpreting results because performance on rare labels may be more susceptible to noise and overfitting.

The simultaneous pre-training models TurkRadBERT-sim v1 and v2 exhibit lower performance compared to their task-adaptive counterparts, indicating that task-adaptive pre-training is more effective in capturing domain-specific knowledge. Nevertheless, further investigation of alternative pre-training and fine-tuning strategies could help enhance the performance of these models.

Future research could focus on expanding the task-specific corpus to improve domain-specific knowledge and performance on rare labels as well as explore alternative pre-training and fine-tuning strategies to further enhance model performance. Moreover, investigating the factors contributing to the performance differences between the models for each label could provide valuable insights for developing more effective models in the field of medical natural language processing.

Acknowledgements

The study was approved by the Ege University Ethical Committee under study number UH150040389 and conducted in accordance with the Declaration of Helsinki. We would also like to express our gratitude to the TPU Research Cloud program (TRC) [4] and Google's CURe program for granting us access to TPUv3 units and GCP credits, respectively.

[4] https://sites.research.google/trc/about/

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Keno K Bressem, Lisa C Adams, Robert A Gaudin, Daniel Tröltzsch, Bernd Hamm, Marcus R Makowski, Chan-Yong Schüle, Janis L Vahldiek, and Stefan M Niehues. 2020. Highly accurate classification of chest radiographic reports using a deep learning natural language model pre-trained on 3.8 million text reports. *Bioinformatics*, 36(21):5255–5261.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 1802. Deep contextualized word representations. corr abs/1802.05365 (2018). *arXiv preprint arXiv:1802.05365*.

Wendi Qu, Indranil Balki, Mauro Mendez, John Valen, Jacob Levman, and Pascal N Tyrrell. 2020. Assessing and mitigating the effects of class imbalance in machine learning with application to x-ray imaging. *International journal of computer assisted radiology and surgery*, 15:2041–2048.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Dario Amodei, Daniela Amodei, Jack Clark, Miles Brundage, and Ilya Sutskever. 2019a. Better language models and their implications. *OpenAI Blog https://openai.com/blog/better-language-models*, 1(2).

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019b. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Elisa Terumi Rubel Schneider, Joao Vitor Andrioli de Souza, Julien Knafou, Lucas Emanuel Silva e Oliveira, Jenny Copara, Yohan Bonescki Gumiel, Lucas Ferro Antunes de Oliveira, Emerson Cabrera Paraiso, Douglas Teodoro, and Cláudia Maria Cabral Moro Barra. 2020. Biobertpt-a portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72.

Stefan Schweter. 2020. Berturk - bert models for turkish.

Wilson L Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

Cagri Toraman, Eyup Halit Yilmaz, Furkan Şahinuç, and Oguzhan Ozcelik. 2023. Impact of tokenization on language models: An analysis for turkish. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–21.

Hazal Turkmen, Oguz Dikenelli, Cenk Eraslan, and Mehmet Cem Callı. 2022. Bioberturk: Exploring turkish biomedical language model development strategies in low resource setting.

Hazal Türkmen, Oğuz Dikenelli, Cenk Eraslan, Mehmet Cem Çalli, and Suha Sureyya Ozbek. 2022. Developing pretrained language models for turkish biomedical domain. In *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)*, pages 597–598. IEEE.

Shoya Wada, Toshihiro Takeda, Shiro Manabe, Shozo Konishi, Jun Kamohara, and Yasushi Matsumura. 2020. Pre-training technique to localize medical bert and enhance biomedical bert. *arXiv preprint arXiv:2005.07202*.

An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y Chang, Amilcare Gentili, and Chun-Nan Hsu. 2022. Radbert: Adapting transformer-based language models to radiology. *Radiology: Artificial Intelligence*, 4(4):e210258.

Angela Zhang, Lei Xing, James Zou, and Joseph C Wu. 2022. Shifting machine learning for healthcare from development to deployment and from models to data. *Nature Biomedical Engineering*, pages 1–16.

A Additional dataset information

Category	Positive	Negative
Intraventricular	22 (%1.1)	1978 (%98.9)
Gliosis	54 (%2.7)	1946 (%97.3)
Epidural	51 (%2.55)	1949 (%97.45)
Hydrocephalus	70 (%3.5)	1930 (%96.5)
Encephalomalacia	177 (%8.85)	1823 (%91.15)
Chronic ischemic changes	951 (%47.55)	1049 (%52.45)
Lacuna	138 (%6.9)	1862 (%93.1)
Leukoaraiosis	49 (%2.45)	1951 (%97.55)
Mega cisterna magna	15 (%0.75)	1985 (%99.25)
Meningioma	39 (%1.95)	1961 (%98.05)
Subarachnoid Bleeding	209 (%10.45)	1791 (%89.55)
Subdural	227 (%11.35)	1773 (%88.65)
No Findings	299 (%14.95)	1701 (%85.05)

Table 4: Distribution of frequencies for each label's positive and negative radiology documents in the dataset.

B Pre-training and fine-tuning hyperparameters

Hyperparameters	Values
Learning rate	1e-4
Batch size	256
Optimizer	Adam
β_1	0.9
β_2	0.999
Warmp up steps	10000
Max sequence length	512
Max prediction per seq	76
Masked MLM probability	0.15
epoch	1000000

Table 5: Pre-training configuration for BERT models.

Parameters	Value
Learning rate	5e-5
Batch size	32
Optimizer	Adam
Max sequence length	512
epoch	20

Table 6: Best fine-tuning configuration for TurkRadBERT-sim family

Parameters	Value
Learning rate	3e-5
Batch size	32
Optimizer	Adam
Max sequence length	512
epoch	15

Table 7: Best fine-tuning configuration for BERTurk, BioBERTurk and TurkRadBERT-task family

A Meta-dataset of German Medical Corpora: Harmonization of Annotations and Cross-corpus NER Evaluation

Ignacio Llorca[1], **Florian Borchert**[2], **Matthieu-P. Schapranow**[2]
Hasso Plattner Institute, University of Potsdam, Germany
Prof.-Dr.-Helmert-Str. 2-3, 14482 Potsdam
[1] llorcarodriguez@uni-potsdam.de
[2] {firstname.lastname}@hpi.de

Abstract

Over the last years, an increasing number of publicly available, semantically annotated medical corpora have been released for the German language. While their annotations cover comparable semantic classes, the synergies of such efforts have not been explored, yet. This is due to substantial differences in the data schemas (syntax) and annotated entities (semantics), which hinder the creation of common meta-datasets. For instance, it is unclear whether named entity recognition (NER) taggers trained on one or more of such datasets are useful to detect entities in any of the other datasets. In this work, we create harmonized versions of German medical corpora using the BigBIO framework, and make them available to the community. Using these as a meta-dataset, we perform a series of cross-corpus evaluation experiments on two settings of aligned labels. These consist in fine-tuning various pre-trained Transformers on different combinations of training sets, and testing them against each dataset separately. We find that a) trained NER models generalize poorly, with F_1 scores dropping approx. 20 pp. on unseen test data, and b) current pre-trained Transformer models for the German language do not systematically alleviate this issue. However, our results suggest that models benefit from additional training corpora in most cases, even if these belong to different medical fields or text genres.

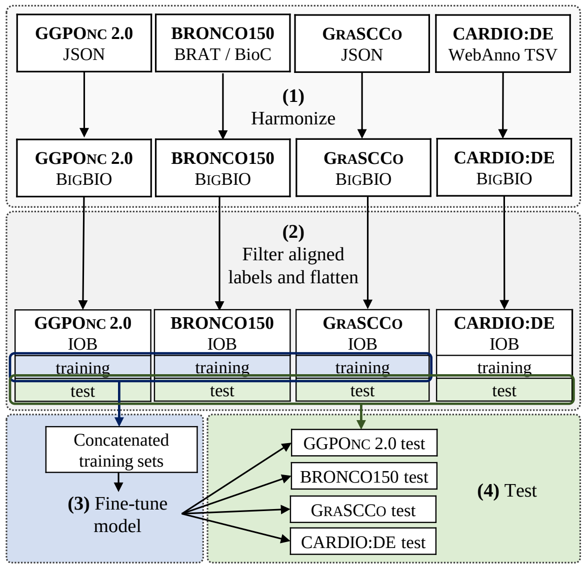


Figure 1: Overview of our experimental design. (1) Each corpus is harmonized from its source format to BigBIO via custom schema parsers. (2) Equivalent entity classes and spans are aligned, the text is tokenized and transformed into IOB format. (3) Training splits of different corpora are concatenated (in the example GGPONC 2.0, BRONCO150 and GRASCCO) and used to train a Transformer-based NER model. (4) The model is evaluated on the test splits of all individual datasets.

1 Introduction

Recently, an increasing amount of medical text datasets for the German language with semantic annotations has been released to the public (Zesch and Bewersdorff, 2022). These corpora come in unequal data formats and with widely varying definitions of annotated entities, e.g., based on ontologies like the UMLS (Bodenreider, 2004), top level hierarchies in SNOMED CT (Donnelly, 2006), or other medical terminologies such as ICD-10. The employed annotation guidelines have usually been created ad-hoc and are hardly reused across annotation projects. Corpora are difficult to compare due to these semantic and syntactic differences. Although various NER models have been trained and tested on individual datasets, their performance across medical fields and text genres has not been investigated. Our work integrates the following distributable, annotated German corpora: BRONCO150 (Kittner et al., 2021), GGPONC 2.0 (Borchert et al., 2022), CARDIO:DE (Richter-Pechanski et al., 2023), and GRASCCO (Modersohn et al., 2022). While the latter did not contain human annotations upon release, it was recently annotated according to the GGPONC 2.0 guidelines (Bressem et al., 2023).

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 171–181
July 14, 2023 ©2023 Association for Computational Linguistics

At the same time, adapted German versions of widely used Transformer models have become publicly available (Chan et al., 2020; Scheible et al., 2020), more recently also specific to biomedical texts (Lentzen et al., 2022; Bressem et al., 2023). While these models have been evaluated on many individual datasets, their performance on truly unseen data remains unknown.

To enable cross-corpus evaluations, we create a *meta-dataset* of German medical corpora by harmonizing them under the schema proposed in the BIGBIO framework (Fries et al., 2022), as illustrated in Fig. 1. This way, we also make the datasets easily available to the community as data loaders in the Hugging Face library (Llorca, 2023). An example for this schema is depicted in Fig. 2. While schema harmonization addresses the issue of syntactic interoperability, the semantics of annotated entities may still differ, as definitions of entity classes have been derived from different medical ontologies. Therefore, we propose two possible alignments of labels across the four German medical corpora and conduct a series of experiments, evaluating different combinations of training corpora and pre-trained Transformers. In this work, we focus on the entity annotations related to *medications*, as these are the only ones that can be aligned consistently across all corpora. Their definitions are still partly extracted from different medical ontologies.

Our goal is to determine whether multiple datasets with similar annotated entities contribute to creating models that can be used outside the domain they were trained in. Here we refer to the domain of a corpus broadly as the set of characteristics conforming it, mainly text genre, medical field, and annotation policy. Such experiments have been successfully conducted for the English language and resulted in robust NER taggers, albeit for entity classes different from the more clinically motivated German-language corpora (Weber et al., 2021). To the best of our knowledge, no such harmonization has been conducted for the German language and clinical entity classes.

The remainder of this work is organized as follows: in Section 2, we review literature on German clinical corpora and biomedical data harmonization. In Section 3, we describe our methods, i.e. the corpora used, data harmonization steps, and performed cross-corpus evaluation. We present our results in Section 4 and discuss them in Section 5. Our work concludes with an outlook in Section 6.

```
{'id': '0',
 'document_id': '00_mundhoehlenkarzinom_0000',
 'passages': [
   {'id': '0-0',
    'type': 'sentence',
    'text': ['Tabakkonsum ist ein wesentlicher...'],
    'offsets': [[0, 90]]}
    ],
 'entities': [
   {'id': '0-0',
    'type': 'Other_Finding',
    'text': ['Tabakkonsum'],
    'offsets': [[0, 11]],
    'normalized': []},
   {'id': '0-1',
    'type': 'Other_Finding',
    'text': ['Risikofaktor für die Entwicklung...'],
    'offsets': [[33, 89]],
    'normalized': []}
    ],
 'events': [],
 'coreferences': [],
 'relations': []}
```

Figure 2: Sample of the target schema for knowledge base construction tasks like NER from BIGBIO.

2 Related Work

In the following, we set our contribution in the context of related work.

2.1 German-language Medical Corpora

In the past, German medical text datasets have been created in closed research environments without the chance of being shared with other researchers. Notable examples include the work of Roller et al. (2016) using clinical notes from nephrology, Hahn et al. (2018) using discharge summaries of internistic or ICU units stays, and König et al. (2019) using discharge letters from the Berlin Aging Study II. Distributable corpora became available just recently, the JSYNCC corpus (Lohr et al., 2018) being a first successful example, although without semantic annotations. The BRONCO150 (Kittner et al., 2021) and CARDIO:DE (Richter-Pechanski et al., 2023) corpora are currently the only instances of annotated, distributable corpora of anonymized patient-level clinical texts. Other open corpora are based on information unrelated to individual patients, e.g. clinical guidelines in Borchert et al. (2020), or are translated versions of a public English dataset, e.g. Frei and Kramer (2022).

To the best of our knowledge, no major effort has been made in cross-corpus evaluation to assess the robustness of German biomedical NER taggers. The baseline models presented by Borchert et al. (2022) or Kittner et al. (2021) are constrained to in-domain, i.e., internal validation. Roller et al. (2022)

conducted external validation of their model on a small subset from GGPONC, but using their original annotation policy. Only Frei and Kramer (2023) and Richter-Pechanski et al. (2023) briefly report on evaluating the baseline NER model from GGPONC 2.0 on aligned medication classes from their respective datasets, with mixed results. Extending this line of research, we consider multiple possible label alignments, analyze span-wise metrics, and explore several combinations of training corpora and pre-trained Transformer models.

2.2 Data Harmonization in Clinical NLP

Several prior works have considered cross-corpus evaluations through the alignment of semantic classes across datasets for different machine learning tasks. For instance, some papers have been released on acoustic emotion recognition (Schuller et al., 2010; Zhang et al., 2011) or general NER (Nothman et al., 2009). In clinical NLP, curated dataset collections with common schemas are not frequent. Efforts like HunFlair include 31 corpora, but limited to the English language (Weber et al., 2021). There are extensive cross-corpus studies of biomedical NER models with similar entity classes and corpora, but also only for the English language (Kaewphan et al., 2016; Giorgi and Bader, 2019; Galea et al., 2018).

In relation to data standards and schemas, many annotated corpora are simply distributed in the raw format of the respective annotation tool, e.g., BRAT (Stenetorp et al., 2012) or WebAnno (Yimam et al., 2013). While formats like BioC (Comeau et al., 2013) present an attempt to standardize annotations and other metadata for biomedical text datasets, the semantics of entity annotations are not fully defined inside the standard. This is counterproductive for cross-corpus integration, as pre-processing efforts are still needed to homogenize the data.

To alleviate these problems, Fries et al. (2022) propose the BIGBIO framework, introducing fixed data schemas for different NLP tasks. BIGBIO makes minimal assumptions on pre-processing decisions to suit different sorts of datasets. In addition, it provides parsers to harmonize more than 126 corpora within this schema and allows easy access to them through the widely used Hugging Face datasets library. However, parsers for the German corpora used in this work were previously not available. Therefore, we have contributed such implementations as part of this work (Llorca, 2023).

3 Materials and Methods

In the following, we present the characteristics of each corpus and an overview of the harmonization and annotation alignment processes. We provide a description of the experimental setup and the evaluation methods used to analyze the results.

3.1 Datasets

An overview of the key details of the corpora used in our cross-corpus experiments is given in Table 1. All considered corpora have been manually annotated by medically trained personnel. Further insights on annotation policies and Inter Annotator Agreement (IAA) are given below:

- **BRONCO150**: De-identified discharge summaries annotated in two groups (A and B) of medical experts and students. IAA as micro-averaged phrase-level F_1 score ranges across entities from 0.81 to 0.94 for group A and from 0.66 to 0.87 for group B. Each semantic class is based on a different medical terminology, which are also used for grounding.
- **GGPONC 2.0**: Clinical guidelines annotated by seven medical students and curated by a medical doctor. Mean IAA, measured through the γ-method (Mathet et al., 2015), is 0.94 across all entity classes on a set of seed documents after iterative annotation guide refinement. Semantic classes are based on SNOMED CT top-level hierarchies.
- **GRASCCO**: Synthetic case reports, originally without annotations. For the benchmarks introduced by Bressem et al. (2023), it was annotated by a single medical student from the GGPONC 2.0 annotation team, following the same guidelines. Thus, the labeled entities and annotation policy are the same for both corpora. However, there is no data on annotation quality and IAA.
- **CARDIO:DE**: De-identified discharge summaries annotated by four medical informatics and two advanced medical students. Fine-grained medication information are annotated following the policy proposed by Uzuner et al. (2010). IAA is reported using token-level median F_1 scores, ranging from 0.33 to 0.98 across classes on seed documents after iterative annotation guide refinement. The lowest IAA for entity classes that we use in this work is 0.76 (*active ingredient*).

Corpus	Med. Field	Text Genre	Tokens	Format	Entities
BRONCO150 Kittner et al. (2021)	Oncology	Discharge Summaries	71K	BioC / BRAT	Diagnosis (ICD-10) Treatment (OPS) Medication (ATC)
GGPONC 2.0 Borchert et al. (2022)	Oncology	Clinical Guidelines	1,877K	JSON	Finding Substance Procedure
GRASCCO Modersohn et al. (2022)	Various	Synthetic Case Reports	43K	JSON	see GGPONC 2.0
CARDIO:DE Richter-Pechanski et al. (2023)	Cardiology	Discharge Summaries	800K	WebAnno TSV	Medications (as in Uzuner et al. (2010))

Table 1: Overview of the used corpora with their annotated entities, medical fields, text genres, size and data formats. A full list of fine-grained entity classes for each corpus can be found in Table 2.

3.2 Harmonization and Label Alignment

For each corpus, we implement a parser within the BIGBIO framework to derive a common notion of documents, passages and entity spans as outlined in Fig. 2. In order to preserve the source integrity, we consider individual sentences as the main units for our experiments, since the definitions of documents and passages differ across corpora.

To obtain semantically equivalent entity classes, we also need to align entity definitions inspired by different medical ontologies across corpora. Our attempt to do this is shown in Table 2. For GGPONC 2.0 and GRASCCO, we consider their fine-grained configuration of entity classes. In some cases, there is no exact equivalence, e.g. it is not immediately clear if *Diagnostic Procedure* in GGPONC 2.0 corresponds to *Treatment* in BRONCO150. Inspection of the annotations shows that these two do not overlap fully, unlike *Therapeutic Procedure* and *Treatment*. Therefore, *Diagnostic Procedure* is left unmapped.

Medications are the only entity class that can be consistently found across all corpora, although its definition is not identical. In fact, CARDIO:DE contains only medication annotations, but much more fine-grained than in the other corpora. BRONCO150 annotations leave out the dosage information of a medication, while CARDIO:DE annotations consider it with dedicated labels. GGPONC 2.0 (and GRASSCO) offer two span length configurations: the short configuration matches the BRONCO150 definition, while the long one covers the *Strength* and *Frequency* annotations from CARDIO:DE as well. Therefore, we can align annotated spans across all corpora as shown in Table 3.

Cases where several medication annotations are either nested or overlap are not possible in some corpora and very seldom in others. Thus, the loss of information when flattening the datasets into IOB format is minimal. Non-contiguous annotations are treated as separate entities, following the same principle used for the NER models in the papers from BRONCO150 and GGPONC 2.0.

3.3 Cross-Corpus Evaluation Experiments

As a result of the above assumptions, we only consider annotations of medication entities for our cross-corpus NER evaluation. We use the following configurations of label alignments:

- **Short-span**: Short-span version of *Clinical Drug* entities from GGPONC 2.0 and GRASCCO, the *Medication* annotations from BRONCO150 and *Drug / Active Ingredients* from CARDIO:DE (discarding linked *Strength* and *Frequency* annotations), resulting in 15 combinations of training corpora.
- **Long-span**: Long-span version of GGPONC 2.0 and GRASCCO, discarding BRONCO150, and merging *Drugs*, *Strength*, and *Frequency* annotations from CARDIO:DE that are linked to each other, as in Richter-Pechanski et al. (2023), resulting in seven combinations.

Afterward, we perform two sets of experiments:

(i) In a larger set of experiments, we fine-tune a Transformer model with a token classification head on all combinations of training data, and evaluate it separately against the test split of each corpus. For these experiments, we use

GGPONC 2.0, GRASCCO	BRONCO150	CARDIO:DE
Diagnosis / Pathology	Diagnosis	–
Clinical Drug	Medication	Active Ingredient, Drug
Therapeutic Procedure	Treatment	–
Other Finding, Diagnostic Procedure, Nutrient / Body Subst., External Subst.	–	–
–	–	Dosage, Route, Form, Reason, Duration, Strength, Frequency

Table 2: Mapping of annotated semantic classes for named entities across datasets. (–) indicates that there are no entities in a certain corpus equivalent to the entity of other dataset. Only the semantic classes for medications (Clinical Drug, Active Ingredient, Medication, Drug) can be mapped across all four corpora.

Example	`Metroprolol`	`95 mg`	`1-0-1`
GGPONC (L)	Clinical Drug		
GGPONC (S)	Clinical Drug	O	
BRONCO150	Medication	O	
CARDIO:DE	Active Ing.	Strength	Freq.

Table 3: Example of how the annotation policies for medications vary in each corpus and how they can be aligned. For GGPONC, L and S refer to the long and short configurations. These apply equivalently for GRASCCO.

the recent BioGottBERT (Lentzen et al., 2022) as the pre-trained Transformer.

(ii) In a second set of experiments, we compare the impact of different Transformer checkpoints on the out-of-domain robustness of trained NER models. For this purpose, we consider only the long-span combinations with two training datasets and an unseen test dataset. The models we compare are GBERT and GELECTRA (Chan et al., 2020), BioGottBERT (Lentzen et al., 2022), and medBERT.de (Bressem et al., 2023).

Despite BRONCO150 having five splits for cross-validation, incorporating this would greatly increase the complexity and number of experiments. Instead, we separate one random split for testing. Similarly, CARDIO:DE does not have pre-defined splits. Thus, we randomly sample a validation and test set containing 12.5 % of all documents, fixed for all experiments.

As hyperparameters, we use a learning rate of 5×10^{-5}, with linear decrease and no weight decay, warmup or label smoothing. All models are trained for 50 epochs on a single NVIDIA A40 GPU with a batch size of 32.

3.4 Evaluation Metrics

We make use of two evaluation methods: seqeval and FairEval. Seqeval is widely used in the field for sequence labeling evaluation and provides a traditional F_1 score implementation (Nakayama, 2018). FairEval is a novel approach to subdue the double-penalties that occur in traditional evaluation when a prediction misses the boundaries of an annotation (Ortmann, 2022). It also provides more fine-grained metrics for error analysis, as it outputs *true positives* (TP) and separates *boundary errors* (BE) from *false positives* (FP) and *false negatives* (FN). In order to ease its usability for the community, we implemented FairEval as a publicly available Hugging Face evaluation module (Llorca, 2022). For the aggregation of scores across test sets, we follow the conclusions of Forman and Scholz (2010) and give greater importance to the micro-averaged results. Macro scores are still reported, accounting for the large size imbalance among the datasets.

4 Results

The seqeval (traditional) F_1 scores of the first set of experiments are shown in Table 4 and 5 for the short and long-span setting, respectively. We omit FairEval scores for this set of experiments for brevity, as the directionality of results is the same.

We use abbreviations with the first three letters to refer to the datasets, i.e. GRASCCO is GRA. We recall the experiments by the row number in Table 4 and 5 or by the following notation: BRO+GGP→BRO corresponds to the model trained on BRONCO150 and GGPONC 2.0, and tested on BRONCO150, i.e. the first cell in Table 4, row 7 (0.925).

4.1 Out-of-domain Generalization of Clinical NER Models

In general, models perform considerably worse when evaluated outside their training domain, i.e., when the training split from the target corpus is not included in the joint training set. On average, F_1 scores are approx. 20 pp. points lower on unseen target corpora for both long-span and short-span experiments. The differences are especially large for the target corpus GGPONC 2.0 test set, having reductions in F_1 score of around 30 pp.

Differences are smaller when the target corpus is GRASCCO, having a decrease of around 10 pp. Notably, the short-span model trained only on GRASCCO obtains an F_1 score of 0.821 on its own test split (Table 4, row 15), while the model trained on GGPONC 2.0 alone performs just 3 pp. worse, achieving an F_1 score of 0.788 (row 14).

4.2 Effect of Adding More Training Data

In general, models are not adversely affected or mislead by adding datasets from different medical fields or text genres other than the training corpus. There are many cases where adding data from a different domain slightly improves the performance. For instance, CAR $\rightarrow$ CAR achieves an F_1 score of 0.876 (Table 4, row 13), while CAR+GGP $\rightarrow$ CAR (row 9) scores slightly higher with 0.880. The same holds true for the long-span setting: CAR+GGP $\rightarrow$ CAR outperforms CAR $\rightarrow$ CAR (Table 5, rows 17/20) by a small margin.

Cases where adding more data is only slightly detrimental are consistent across all experiments. Considering the short-span experiments with GGPONC 2.0 as the target corpus, we see how training just on itself achieves 0.910 F_1 score (Table 4, row 14) and adding more corpora decreases performance slightly up to 0.905 (for all four datasets, row 1). This finding can be observed across all experimental settings.

Such marginal loss of performance trades off positively with the robustness of models across multiple corpora. The results of the model trained on all corpora (BRO+CAR+GGP+GRA in Table 4, row 1) are slightly below those obtained by models trained on each corpus separately (shaded diagonal in Table 4, rows 12-15), while increasing the micro F_1 by a wide margin of 19 pp. on average. The same holds true for the long-span setting, with an average increase in micro F_1 of 18 pp.

4.3 Performance of Different Transformer Checkpoints

Results from the second set of experiments to investigate the impact of different pre-trained Transformer checkpoints on the out-of-domain robustness of NER taggers are presented in Table 6. This time, FairEval F_1 scores are shown together with the seqeval (traditional) scores, to gain more insights into the actual magnitude of the performance drop compared to the in-domain baseline.

For the setting tested on GGPONC 2.0, the best Transformer checkpoint varies when boundary errors are counted once instead of twice: medBERT.de obtains a higher FairEval score than BioGottBERT, whilst achieving a lower seqeval score.

There is no clear pattern with regard to the generalization capabilities of different pre-trained Transformers. GELECTRA performs best in two out of three scenarios, but falls in third place for the remaining case, where GGPONC 2.0 is the unseen target. Additionally, BioGottBERT is always the second-best checkpoint whenever GELECTRA gets the first place. The best performing Transformer for the settings tested on GGPONC 2.0 and CARDIO:DE are still far from a baseline where the model has seen the training split of the target corpus in training. In contrast, for the setting tested on GRASCCO, GELECTRA obtains a traditional F_1 score just 1 pp. below the baseline result from BioGottBERT on GRA $\rightarrow$ GRA.

5 Discussion

In this section, we discuss our findings and perform a fine-grained error analysis.

5.1 Cross-Corpus Evaluation

In general, all models perform poorly on truly unseen data, no matter if the datasets belong to the same medical field (BRONCO150 and GGPONC 2.0 concern oncology), if the annotation procedure and source format are the same (for GGPONC 2.0 and GRASCCO) or if the text genre is similar (BRONCO150 and CARDIO:DE contain discharge summaries).

When the model has not seen the target corpus during training, it performs significantly below par, which we attribute to the widely different entity definitions and annotation policies. This is the case even for a seemingly well-defined semantic class like medications. Although the pattern is less evident for the short-span configuration, this is likely

		Test set				F_1 (Average)	
		BRO	**CAR**	**GGP**	**GRA**	**Micro**	**Macro**
1	BRO CAR GGP GRA	0.928	0.876	0.905	0.783	**0.898**	**0.874**
2	BRO CAR GGP	0.930	0.873	0.906	0.694	<u>0.897</u>	0.852
3	BRO CAR GRA	0.916	0.876	0.593	0.776	0.694	0.792
4	BRO GGP GRA	0.937	0.708	0.906	**0.854**	0.850	<u>0.855</u>
5	CAR GGP GRA	0.820	0.879	0.906	0.717	0.890	0.832
6	BRO CAR	0.932	**0.883**	0.549	0.719	0.672	0.774
7	BRO GGP	0.925	0.728	<u>0.907</u>	0.745	0.855	0.829
8	BRO GRA	<u>0.946</u>	0.713	0.631	0.796	0.680	0.781
9	CAR GGP	0.812	<u>0.880</u>	0.901	0.694	0.885	0.823
10	CAR GRA	0.779	0.879	0.588	0.778	0.696	0.767
11	GGP GRA	0.798	0.724	<u>0.907</u>	<u>0.846</u>	0.846	0.823
12	BRO	**0.956**	0.740	0.562	0.681	0.647	0.745
13	CAR	0.754	0.876	0.489	0.687	0.628	0.708
14	GGP	0.758	0.684	**0.910**	0.788	0.834	0.786
15	GRA	0.774	0.812	0.669	0.821	0.718	0.773
	Mean on seen data	0.934	0.878	0.906	0.796		
	Mean on unseen data	0.785	0.730	0.583	0.715		

Table 4: F_1 scores (short-span setting) resulting from tuning BioGottBERT on each combination of training sets against each separate target corpus and their micro and macro aggregation. The example from Fig. 1 would correspond with row number 4. We highlight in **bold** and <u>underlined</u> the highest and second-highest scores for each test set. The shaded cells denote experiments where the training portion of the test corpus is seen at training. We see that (1) models generalize poorly to other domains (unshaded cells are consistently lower scores than shaded ones) and (2) models generally benefit from adding more corpora at training to the target corpus.

		Test set			F_1 (Average)	
		CAR	**GGP**	**GRA**	**Micro**	**Macro**
16	CAR GGP GRA	0.796	0.788	0.549	<u>0.769</u>	**0.716**
17	CAR GGP	**0.807**	0.788	0.485	**0.774**	<u>0.698</u>
18	CAR GRA	0.801	0.480	0.409	0.577	0.569
19	GGP GRA	0.579	**0.794**	**0.625**	0.710	0.676
20	CAR	<u>0.804</u>	0.424	0.258	0.543	0.504
21	GGP	0.560	<u>0.793</u>	0.547	0.703	0.639
22	GRA	0.593	0.496	<u>0.606</u>	0.532	0.599
	Mean on seen data	0.802	0.791	0.547		
	Mean on unseen data	0.577	0.467	0.430		

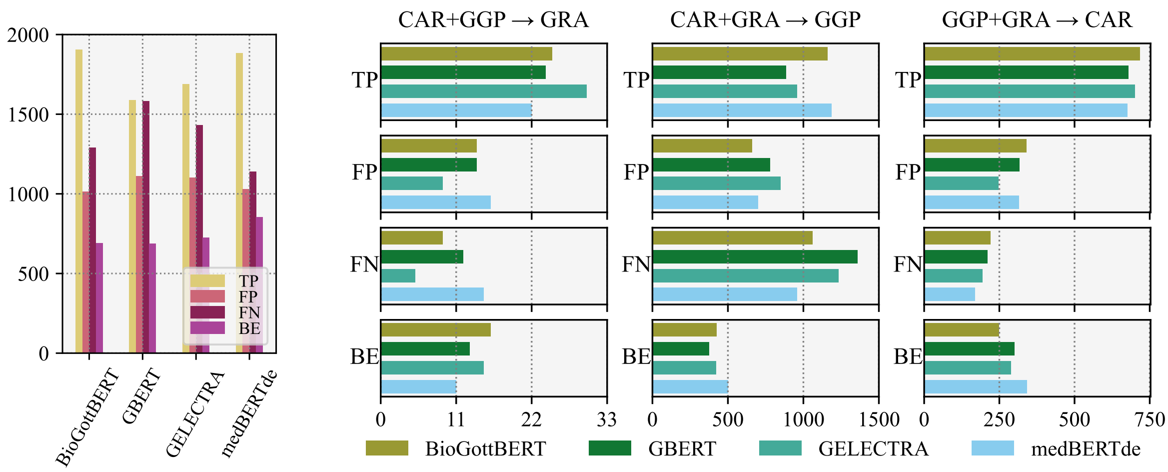
Table 5: F_1 scores (long-span setting) resulting from tuning BioGottBERT on each combination of training sets against each separate target corpus and their micro and macro aggregation. Values highlighted as in Table 4. The findings drawn in Table 4 are even more notable in this setting.

just because the task at hand is easier (i.e. shorter spans are easier to identify) – without seeing the test corpus at training, the scores remain on levels that may be deemed acceptable, but are still considerably worse.

Having models achieve higher micro and macro F_1 scores across all target sets when they have seen more corpora at training is consistent with our assumptions. The fact that adding corpora to the training split of the target corpus does not significantly reduce performance has promising implications: if there were enough open datasets, current day neural network architectures are indeed enough to obtain robust NER taggers through their combination. Concerning the cases where adding more corpora to the target corpus at training time increases model performance on it, GGPONC 2.0 seems to be the most contributing dataset: the results of CAR →

	CAR+GGP→GRA		CAR+GRA→GGP		GGP+GRA→CAR	
	seqeval	**FairEval**	**seqeval**	**FairEval**	**seqeval**	**FairEval**
BioGottBERT	0.485	0.562	**0.480**	0.520	0.579	0.640
GBERT	0.475	0.552	0.384	0.413	0.554	0.621
GELECTRA	**0.594**	**0.674**	0.398	0.434	**0.581**	**0.658**
medBERT.de	0.458	0.512	0.456	**0.524**	0.550	0.620
Baseline	0.606	0.723	0.793	0.839	0.807	0.846

Table 6: Out-of-domain evaluation of different Transformer checkpoints. We consider the experiments of the long-span configuration that included two corpora for training and were tested on the unseen dataset. We report traditional (seqeval) and FairEval F_1 scores to account for the effect that double penalties on close-to-target predictions have in model selection. For reference, we include the single-corpus, in-domain results achieved by BioGottBERT as a baseline (GRA → GRA, GGP → GGP, CAR → CAR).



(a) Aggregated error counts

(b) Error counts grouped by experiment and error type

Figure 3: Error counts (**T**rue **P**ositives, **F**alse **P**ositives, **F**alse **N**egatives and **B**oundary **E**rrors) per Transformer checkpoint for long-span experiments using two corpora for training and evaluated on the remaining, unseen corpus.

CAR and GRA → GRA improve when adding GGP to the training sets in both span-length configurations. Suspected reasons could be its large size, thematic diversity, or relatively high IAA. It also suggests that non-patient-related data (like clinical guidelines) can be useful to robust models when evaluated on patient-related data such as discharge summaries.

5.2 Error Analysis

The comparison of different checkpoints is initially favorable to GELECTRA, performing best in two out of three settings. It should also be noted that the best models in the last case (BioGottBERT and medBERT.de) included unlabelled texts from GGPONC in their pre-training phase

A more detailed error analysis shows that BioGottBERT and medBERT.de obtain more TPs, while producing fewer FNs and FPs aggregated through all three experiments than GELECTRA (see Fig. 3a). Furthermore, BioGottBERT also produces less boundary errors, making a case for the current most robust model on unseen data. However, the averaged trend is not consistently reflected across individual experiments (see Fig. 3b).

It is also noteworthy that general-domain models are more prone to FNs, i.e., completely missing some entities. We suppose that the reason for this is that biomedical-tuned models are more familiar with the medical terminology in the datasets. In contrast, the number of FPs is closer for all Transformer models.

6 Conclusion and Outlook

Reaching agreements and establishing standards for clinical entity annotation is vital to facilitate inter-corpus operability. So it is adhering to similar formats and schemas to structure the information. This can help to combine the sporadically released open medical text datasets for non-English languages to build more robust models. Our work aimed to support this goal by harmonizing all currently available, semantically annotated German medical corpora within the BIGBIO framework and make the implemented data loaders available to the community.

Our experiments show that the currently available corpora in German serve for poorly generalizable models. Our results also suggest that mixing multiple corpora in training is beneficial on single test splits and widely improves robustness, thus highlighting the importance of easing cross-corpus integration. The comparison of different pre-trained Transformers does not shed conclusive results. Both general-purpose and biomedical-specific instances seem to perform similarly on unseen data.

The results presented in this paper correspond to a single training iteration on all combinations of the pre-defined train-test splits of the data. For future work, we consider performing proper cross-validation experiments by dividing each corpus into folds and using all resulting combinations in order to obtain more stable results and confidence intervals. However, this approach increases the number of experiments from 34 trained models to 340. We have obtained preliminary results of such evaluation, and the findings are consistent with the ones presented in this work. Other options for further research include extending our comparison of multiple Transformers, or even considering generative approaches to NER.

Investigating whether other label alignments are meaningful once more comparable datasets become available would help to reinforce our results outside of medication annotations. Given the currently available corpora, the only other entities that might be comparable are the short version of *Diagnosis/Pathology* and *Therapeutic Procedure* from GGPONC 2.0 with *Diagnosis* and *Treatment* from BRONCO150, respectively. However, here the differences in semantics are even more pronounced than in the case of medication. Annotation campaigns using unpublished corpora, which concern other medical fields and text genres, suggest that we might be able to harmonize other semantic classes in the future. For instance, the *Condition* category in the fine-grained annotation scheme proposed by Roller et al. (2016) for clinical notes in nephrology roughly corresponds to the *Findings* class in GGPONC 2.0.

To conclude, our study calls for more representative large German clinical corpora to generate robust NER taggers that can be used for real-world scenarios, together with a consensus on the semantics and annotation guidelines to equate labeled entities through the datasets.

Limitations

Our findings are limited to medication entities, the only semantic class that is annotated in all available corpora. Moreover, we had to exclude BRONCO150 for long-span experiments due to a mismatch of entity definitions. Although the label alignment decisions are somewhat subjective, they are made based on a thorough inspection of definitions and samples.

The differences in annotation quality and biases may be playing an uncertain role in the models. However, making statements on the impact of the annotation quality is challenging, since each work followed a different annotation protocol and reports different measures of annotator agreement. This is another area where harmonization efforts might be warranted for future research. Furthermore, exploring different hyperparameter configurations lied out of scope for our work, but could have a substantial impact. Mainly, the results from the Transformers comparison (Table 6) could shed different conclusions if the hyperparameters were optimized for each model.

Acknowledgements

Parts of this work were generously supported by grants of the German Federal Ministry of Research and Education (01ZZ1802H, 01ZZ2314N) and the German Federal Ministry of Economic Affairs and Climate Action (01MJ21002A).

References

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32:D267–D270.

Florian Borchert, Christina Lohr, Luise Modersohn, Thomas Langer, Markus Follmann, Jan Philipp Sachs, Udo Hahn, and Matthieu-P. Schapranow. 2020. GGPONC: A corpus of German medical text with rich metadata based on clinical practice guidelines. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages

38–48, Online. Association for Computational Linguistics.

Florian Borchert, Christina Lohr, Luise Modersohn, Jonas Witt, Thomas Langer, Markus Follmann, Matthias Gietzelt, Bert Arnrich, Udo Hahn, and Matthieu-P. Schapranow. 2022. GGPONC 2.0 - the German clinical guideline corpus for oncology: Curation workflow, annotation policy, baseline NER taggers. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3650–3660, Marseille, France. European Language Resources Association.

Keno K Bressem, Jens-Michalis Papaioannou, Paul Grundmann, Florian Borchert, Lisa C Adams, Leonhard Liu, Felix Busch, Lina Xu, Jan P Loyen, Stefan M Niehues, et al. 2023. MEDBERT.de: A comprehensive German BERT model for the medical domain. *arXiv preprint arXiv:2303.08179*.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Donald C Comeau, Rezarta Islamaj Doğan, Paolo Ciccarese, Kevin Bretonnel Cohen, Martin Krallinger, Florian Leitner, Zhiyong Lu, Yifan Peng, Fabio Rinaldi, Manabu Torii, et al. 2013. BioC: a minimalist approach to interoperability for biomedical text processing. *Database*, 2013.

Kevin Donnelly. 2006. SNOMED-CT: the advanced terminology and coding system for eHealth. In *Medical and Care Compunetics 3*, number 121 in Studies in Health Technology and Informatics, pages 279–290, Amsterdam etc. IOS Press.

George Forman and Martin Scholz. 2010. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *ACM Sigkdd Explorations Newsletter*, 12(1):49–57.

Johann Frei and Frank Kramer. 2022. GERNERMED: An open German medical NER model. *Software Impacts*, 11:100212.

Johann Frei and Frank Kramer. 2023. German medical named entity recognition model and data set creation using machine translation and word alignment: Algorithm development and validation. *JMIR Form Res*, 7:e39077.

Jason Fries, Leon Weber, Natasha Seelam, Gabriel Altay, Debajyoti Datta, Samuele Garda, Sunny Kang, Rosaline Su, Wojciech Kusa, Samuel Cahyawijaya, Fabio Barth, Simon Ott, Matthias Samwald, Stephen Bach, Stella Biderman, Mario Sänger, Bo Wang, Alison Callahan, Daniel León Periñán, Théo Gigant, Patrick Haller, Jenny Chim, Jose Posada, John Giorgi, Karthik Rangasai Sivaraman, Marc Pàmies, Marianna Nezhurina, Robert Martin, Michael Cullan, Moritz Freidank, Nathan Dahlberg, Shubhanshu Mishra, Shamik Bose, Nicholas Broad, Yanis Labrak, Shlok Deshmukh, Sid Kiblawi, Ayush Singh, Minh Chien Vu, Trishala Neeraj, Jonas Golde, Albert Villanova del Moral, and Benjamin Beilharz. 2022. BigBio: A framework for data-centric biomedical natural language processing. In *Advances in Neural Information Processing Systems*, volume 35, pages 25792–25806. Curran Associates, Inc.

Dieter Galea, Ivan Laponogov, and Kirill Veselkov. 2018. Exploiting and assessing multi-source data for supervised biomedical named entity recognition. *Bioinformatics*, 34(14):2474–2482.

John M Giorgi and Gary D Bader. 2019. Towards reliable named entity recognition in the biomedical domain. *Bioinformatics*, 36(1):280–286.

Udo Hahn, Franz Matthies, Christina Lohr, and Markus Löffler. 2018. 3000pa-towards a national reference corpus of german clinical language. In *MIE*, pages 26–30.

Suwisa Kaewphan, Sofie Van Landeghem, Tomoko Ohta, Yves Van de Peer, Filip Ginter, and Sampo Pyysalo. 2016. Cell line name recognition in support of the identification of synthetic lethality in cancer from text. *Bioinformatics*, 32(2):276–282.

Madeleine Kittner, Mario Lamping, Damian T Rieke, Julian Götze, Bariya Bajwa, Ivan Jelas, Gina Rüter, Hanjo Hautow, Mario Sänger, Maryam Habibi, et al. 2021. Annotation and initial evaluation of a large annotated german oncological corpus. *JAMIA open*, 4(2):ooab025.

Maximilian König, André Sander, Ilja Demuth, Daniel Diekmann, and Elisabeth Steinhagen-Thiessen. 2019. Knowledge-based best of breed approach for automated detection of clinical events based on German free text digital hospital discharge letters. *PloS one*, 14(11):e0224916.

Manuel Lentzen, Sumit Madan, Vanessa Lage-Rupprecht, Lisa Kühnel, Juliane Fluck, Marc Jacobs, Mirja Mittermaier, Martin Witzenrath, Peter Brunecker, Martin Hofmann-Apitius, et al. 2022. Critical assessment of transformer-based ai models for german clinical notes. *JAMIA open*, 5(4):ooac087.

Ignacio Llorca. 2022. Programmatic access to FairEval as a HuggingFace evaluation module. `https://huggingface.co/spaces/hpi-dhc/FairEval`. (Last accessed: April 26th, 2023).

Ignacio Llorca. 2023. BIGBIO loaders for German clinical corpora (GGPONC 2.0, CARDIO:DE, BRONCO150) in the HuggingFace Hub. `https://huggingface.co/datasets/bigbio/`{ggponc2, cardiode, bronco}. (Last accessed: April 26th, 2023).

Christina Lohr, Sven Buechel, and Udo Hahn. 2018. Sharing copies of synthetic clinical corpora without physical distribution — a case study to get around

IPRs and privacy constraints featuring the German JSYNCC corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3):437–479.

Luise Modersohn, Stefan Schulz, Christina Lohr, and Udo Hahn. 2022. GRASCCO - the first publicly shareable, multiply-alienated german clinical text corpus. *Studies in health technology and informatics*, 296:66—72.

Hiroki Nakayama. 2018. seqeval: A Python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Joel Nothman, Tara Murphy, and James R. Curran. 2009. Analysing Wikipedia and gold-standard corpora for NER training. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 612–620, Athens, Greece. Association for Computational Linguistics.

Katrin Ortmann. 2022. Fine-grained error analysis and fair evaluation of labeled spans. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1400–1407, Marseille, France. European Language Resources Association.

Phillip Richter-Pechanski, Philipp Wiesenbach, Dominic M. Schwab, Christina Kiriakou, Mingyang He, Michael M. Allers, Anna S. Tiefenbacher, Nicola Kunz, Anna Martynova, Noemie Spiller, Julian Mierisch, Florian Borchert, Charlotte Schwind, Norbert Frey, Christoph Dieterich, and Nicolas A. Geis. 2023. A distributable German clinical corpus containing cardiovascular clinical routine doctor's letters. *Nature Scientific Data*, 10:207.

Roland Roller, Laura Seiffe, Ammer Ayach, Sebastian Möller, Oliver Marten, Michael Mikhailov, Christoph Alt, Danilo Schmidt, Fabian Halleck, Marcel Naik, et al. 2022. A medical information extraction workbench to process german clinical text. *arXiv preprint arXiv:2207.03885*.

Roland Roller, Hans Uszkoreit, Feiyu Xu, Laura Seiffe, Michael Mikhailov, Oliver Staeck, Klemens Budde, Fabian Halleck, and Danilo Schmidt. 2016. A fine-grained corpus annotation schema of German nephrology records. In *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, pages 69–77, Osaka, Japan. The COLING 2016 Organizing Committee.

Raphael Scheible, Fabian Thomczyk, Patric Tippmann, Victor Jaravine, and Martin Boeker. 2020. GottBERT: a pure German language model. *arXiv preprint arXiv:2012.02110*.

Bjorn Schuller, Bogdan Vlasenko, Florian Eyben, Martin Wöllmer, Andre Stuhlsatz, Andreas Wendemuth, and Gerhard Rigoll. 2010. Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Transactions on Affective Computing*, 1(2):119–131.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Özlem Uzuner, Imre Solti, and Eithon Cadag. 2010. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518.

Leon Weber, Mario Sänger, Jannes Münchmeyer, Maryam Habibi, Ulf Leser, and Alan Akbik. 2021. Hunflair: an easy-to-use tool for state-of-the-art biomedical named entity recognition. *Bioinformatics*, 37(17):2792–2794.

Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. Webanno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 1–6.

Torsten Zesch and Jeanette Bewersdorff. 2022. German medical natural language processing–a data-centric survey. In *The Upper-Rhine Artificial Intelligence Symposium UR-AI 2022 : AI Applications in Medicine and Manufacturing, 19 October 2022, Villingen-Schwenningen, Germany*, pages 137–145. Furtwangen University.

Zixing Zhang, Felix Weninger, Martin Wöllmer, and Björn Schuller. 2011. Unsupervised learning in cross-corpus acoustic emotion recognition. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 523–528.

Uncovering the Potential for a Weakly Supervised End-to-End Model in Recognising Speech from Patient with Post-Stroke Aphasia

Giulia Sanguedolce
Department of Computing,
Department of Electrical &
Electronic Engineering,
Department of Brain Sciences,
Imperial College London

Patrick A. Naylor
Department of Electrical &
Electronic Engineering,
Imperial College London

Fatemeh Geranmayeh
Department of Brain Sciences,
Imperial College London

Abstract

Post-stroke speech and language deficits (aphasia) significantly impact patients' quality of life. Many with mild symptoms remain undiagnosed, and the majority do not receive the intensive doses of therapy recommended, due to healthcare costs and/or inadequate services. Automatic Speech Recognition (ASR) may help overcome these difficulties by improving diagnostic rates and providing feedback during tailored therapy. However, its performance is often unsatisfactory due to the high variability in speech errors and scarcity of training datasets. This study assessed the performance of Whisper, a recently released end-to-end model, in patients with post-stroke aphasia (PWA). We tuned its hyperparameters to achieve the lowest word error rate (WER) on aphasic speech. WER was significantly higher in PWA compared to age-matched controls (10.3% vs 38.5%, $p < 0.001$). We demonstrated that worse WER was related to the more severe aphasia as measured by expressive (overt naming, and spontaneous speech production) and receptive (written and spoken comprehension) language assessments. Stroke lesion size did not affect the performance of Whisper. Linear mixed models accounting for demographic factors, therapy duration, and time since stroke, confirmed worse Whisper performance with left hemispheric frontal lesions. We discuss the implications of these findings for how future ASR can be improved in PWA.

1 Introduction

Aphasia is a language impairment that causes difficulties in speaking, understanding and/or writing coherent and meaningful sentences. This deficit negatively impacts numerous daily activities, such as working, shopping or participating in community and leisure experiences. As a consequence, patients with aphasia report high levels of depression, passiveness, social exclusion and a general decline in their quality of life (Spaccavento et al., 2014). Overall, there are at least 2 000 000 people in the USA (National Aphasia Association) and more than 350 000 people in the UK with aphasia (Stroke Association). Roughly 45% of aphasic disorders arises following a stroke (Ali et al., 2015). Stroke cases, mortality and morbidity have increased substantially over the last two decades, with 70% increase in incident strokes, 43% deaths from stroke, and 143% DALYs Feigin et al. (2022). Consequently, the incidence of aphasia has also increased. Importantly, the presence of aphasia per se worsens the overall stroke outcomes (Lazar and Boehme, 2017; Geranmayeh et al., 2016). Therefore, due to the psycho-social burden and the current increase in stroke cases, early diagnosis and treatment of aphasia need to be addressed.

The mainstay treatment of aphasia is speech and language therapy; it entails practices with language exercises for improving language ability, as well as adjusting to new ways of communicating (Palmer et al., 2018). According to the results of different meta-analyses, higher intensity speech therapy treatment is strongly associated with greater treatment efficacy (Robey 1998; Bhogal et al. 2003; Kelly et al. 2010; Breitenstein et al. 2017). Providing ongoing efficient treatment, however, can be challenging due to limited resources, which can make face-to-face speech therapy costly and difficult to achieve for every patient need (Palmer et al., 2012; Le et al., 2018). The situation became worse especially after the COVID-19 pandemic crisis, that led to the suspension or the slowdown of non-urgent care, including speech and language therapies (Chadd et al., 2021).

A solution for these issues might be the use of speech recognition models, able to remotely and automatically transcribe long pieces of conversation to easily analyse patients language profiles and to give tailored treatments. Nevertheless, even though Automatic Speech Recognition (ASR) tools have been already explored in research, until now these

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 182–190

have been slow to catch up with the performance obtained in healthy speech (Abad et al., 2013; Le et al., 2016; Jamal et al., 2017; Le et al., 2018). Indeed, the models trained on healthy data struggle to achieve high accuracy in metrics like *Word Error Rate* (WER) or *Phoneme Error Rate* (PER), mostly due to the features of aphasic speech.

Speech from PWA is largely thought to have semantic (meaning) and phonological (speech sound) errors, as well as dysfluencies, each with independent recovery trajectories (Stefaniak et al., 2022). Furthermore, aphasic speech has characteristics that might include: slow and hesitant elocution with episodes of agrammatism (e.g. absence or improper use of function words and verbs - Damico et al. 2010), word-finding problems that affect mostly nouns and picturable action words, frequent stammer, as well as an overall flow of speech that is often fragmented, choppy, unintelligible and/or awkwardly articulated (Abad et al., 2013). These aspects can be influenced also by motor control problems like apraxia and dysarthria, frequently present in aphasia, which may also produce articulation distortion and aberrant prosody (Le et al., 2016). Hence, the challenges that these models need to address include the high variability of speech errors, both between and within aphasic individuals, as well as the lack of satisfactory training datasets.

We tested the performance of a state-of-art sequence-to-sequence ASR transformer model that, to our knowledge, has not been used yet on clinical data. This model, released by Open AI (Radford et al., 2022), is named *Whisper*[1] and it is known for its superior performance in healthy speech when compared with other notable commercial and open-source ASR systems. The feasibility of this model in clinical practice is supported by the low WER in healthy speakers, and the powerful large multilingual weakly supervised dataset on which it was trained. Moreover, the ability to run Whisper locally, will help to preserve the privacy of patients' sensitive data and allow testing in compliance with local and continental regulations. For the purpose of this study, the Whisper testing is done on a novel database of speech of PWA that we have created.

We fine-tuned Whisper parameters relevant for aphasic speech, detailed in Section 3.2. This led us to retrieve the best model according to the lowest WER to test on speech audio. We then compared the patients' WER to an aged-matched control group that performed the same speech production task. After correlation analyses, we created linear mixed-effects models and observed interesting and significant relations with the average performance of the ASR (see Section 4). According to these results, our analysis offers useful insights to consider for our next steps, from which other researchers can also benefit. We expect that our study will advance the work of ASR for PWA, enriching and inspiring the research of the natural language processing community applied in the healthcare framework.

2 Related Work

Since the introduction of ASR technology in clinical studies, algorithms have had to deal with several challenges. The variability and complexity of disordered speech, sometimes unintelligible, has led researchers to move forward with the creation of novel ASR trained with pathological speech data. Nevertheless, an additional difficulty they have to face is the scarcity of datasets of such disordered speech, limiting the accuracy and/or the generalisability of the results. An example of this is the work of Peintner and colleagues (2008), which extracted language features from their corpus for distinguishing different frontotemporal lobar degeneration, one of which includes progressive non-fluent aphasia. Although the study demonstrated encouraging outcomes, it was conducted on a comparatively limited dataset, and no examination was performed regarding the reliability of the features extracted using ASR.

Similarly, Fraser et al. (2013) attempted to differentiate and diagnose primary progressive aphasia (PPA) and two of its sub-types, semantic dementia (SD) and progressive non-fluent aphasia (PNFA) extracting 58 lexical and syntactic features. Using a reduced dataset, an optimized support vector machine (SVM) and random forests (RF) classifiers, Jin and colleagues (2022) tried to face the problem of the dataset with data augmentation on a recognition model for patients with dysarthria. They reached an overall WER of 27.8% on the *UASpeech* test set, underlining that the lowest published WER on the subset of speakers with "Very Low" unintelligibly was of 57.3%.

Differently, Kohlschein and colleagues (2017) used the speech elicited in the *Aachen Aphasia*

[1]The model name comes from the acronym of WSPSR standing for Web-scale Supervised Pretraining for Speech Recognition

Test assessment as a database to train their algorithm. They built a model that automatically analysed pathological speech to identify patients' aphasia type and severity based solely on acoustic features. *AphasiaBank*, a large database open to members, was used by Le and colleagues (2018) to successfully detect medically-relevant quantitative measures to predict aphasia with WER (Word Error Rate) of 39% in spontaneous aphasic speech. Their previous work, with a WER of 45% (Le and Provost, 2016; Le et al., 2017), established the first ASR baseline on AphasiaBank, showing that this dataset can guide the understanding of aphasic speech recognition.

In Le et al. (2017), the authors used an acoustic modelling architecture of multi-task DBLSTM-RNN (double bidirectional long short-Term memory recurrent neural network) with four hidden BLSTM with 2 diverse language models for decoding. The authors investigated features based on speech duration, the quality of pronunciation, phone edit distance, and dynamic time warping on phoneme posteriorgrams. On the other hand, Le et al. (2018), even though using a similar pipeline, chose to investigate lexical diversity and complexity, posteriorgram-based dynamic time warping, pairwise variability error, dysfluency and information density in aphasic speech. Lastly, like the work of Qin et al. (2016), a Cantonese version of AphasiaBank has been implemented by Liu et al. (2018), together with the *CUSENT* and *CanPEV* Cantonese corpora. In this case, as an evaluation metric they used a Syllable Error Rate (SER) with the AphasiaBank and a multilayer time delay neural network (MT-TDNN) with a bidirectional long short-term memory (BLSTM) model structure. This obtained 18.5% of WER for unimpaired speech and 42.4% for impaired speech.

An alternative strategy is to employ ASR that already exists as per study by Mahmoud and colleagues (2023), where the authors customised existing ASR for a specific research goal, selecting Microsoft Azure Speech-to-Text API or Google Speech-to-Text API. In this study we are adopting a similar approach: given the impressive performance of Whisper on healthy speech, largely due to its training dataset being several orders of magnitude larger than preceding ASR, we expect Whisper accuracy to be similar to aforementioned models trained on aphasic data.

Table 1: Sample Characteristics

	Control ($N = 23$)	**Patients** ($N = 23$)
	Mean (Standard Deviation)	
Age (*months*)	59.96 (11.24)	61.45 (10.98)
Gender		
Male	10	14
Female	13	9
Grammatical Complexity*	15.17 (3.59)	9.60 (4.33)
Productivity	127.83 (59.87)	91.41 (53.66)
Lexical Diversity*	60.65 (21.85)	37.01 (16.64)
Fluency***	136.81 (39.66)	68.31 (39.13)
Flawed Syntax (%)**	3.53 (8.60)	37.08 (34.16)

$^{*}: p < 0.05$; $^{**}: p < 0.01$; $^{***}: p < 0.001$

3 Methods

3.1 Dataset

For our study, we used the SONIVA (*Speech recOgNItion Validation in Aphasia*) database, a comprehensive validation database that we are creating for training automated aphasic speech recognition in the research and clinical setting. SONIVA is composed of speech recordings derived from PWA taking part in the IC3[2] study (Imperial Comprehensive Cognitive Assessment in Cerebrovascular Disease; Gruia et al. 2022), and PLORAS study (Predicting Language Outcome and Recovery After Stroke; Seghier et al. 2016). The SONIVA database aims to be a large and comprehensively annotated speech database including quantitative measures of speech and English as well as IPA transcriptions. With this dataset we are producing quantitative summary measures from the *Comprehensive Aphasia Test* (CAT; Swinburn et al. 2004). To understand the various relations with the WER, we included into statistical models the patients' CAT-derived summary measures, quantitative measures of spontaneous speech, size and location of stroke lesion, and demographic factors.

We used as input to Whisper the data of 46 participants, divided into an aged-matched controls group ($N = 23$) and PWA (Patients with aphasia; $N = 23$). For patients, audio speech was collected across multiple time-points since their stroke, resulting in a total of 38 audio files. The speech is recorded during the picture description task from the CAT assessment (Swinburn et al., 2004).

The audio was transcribed verbatim by a speech

[2]https://www.ic3study.co.uk

therapist and 3 trained postgraduate students, with excellent inter-rater reliability (73% overall word-level match). The text is in CHAT format (Codes for the Human Analysis of Transcripts; MacWhinney 2014), managed and analyzed through the CLAN software. Using CLAN, the following measures were generated: *grammatical complexity* (mean length of utterance in morphemes), *productivity* (number of total words), *lexical diversity* (number of different words), *fluency* (words per minute) and *flawed syntax* (incorrect utterances that do not have at least one verb, copula, modal, or participle). All these measures are included in the sample characteristics in table 1, together with the Mann-Whitney tests results in case of significant differences between groups.

3.2 End-to-end Transformer

With 680 000 hours of training on noisy data, of which approximately 20% is derived from non-English languages, its performance on healthy speech has been near human-level with respect to accuracy (Radford et al., 2022). In addition to Whisper's large training dataset, its superior performance is enhanced by the weakly supervised transcription. Its labels are not fully precise or complete, but rather are noisy or partial, because the authors used an ASR to create the labels, which are not perfect and prone to errors. Nevertheless, in order to improve the labels' quality, any text that seemed to be created automatically was discarded. This included the elimination transcriptions that had only upper- or lower-case letters or lacked punctuation, as these were probably generated by machines rather than people. Once they created this dataset, the original version of Whisper was trained and used to understand what was wrong with the data (through error rating metrics) for manually inspecting the low-quality parts and creating an iterative training process.

The model architecture is a sequence-to-sequence transformer, commonly used since 2017 (Vaswani et al., 2017) for its reliability. The audio chunks are initially transformed into an 80-channel, 25 ms window, 10 ms stride Mel spectrogram. The features are scaled between -1 and 1 with a mean of 0 throughout the sample. Interestingly, their multi-task training set has special tokens as task specifiers or classification targets (such as language identification or timestamp tokens). Whisper uses the same byte-level BPE text tokenizer used in GPT-2 (Sennrich et al. 2015; Radford et al. 2019) for the English-only models, as they have both English-only and multi-language models, released in different sizes (from 39M parameters for tiny model to 1.55B parameters for large model).

3.3 Hyperparameters Fine-Tuning

We conducted a grid search fine-tuning, choosing the best performing model based on the WER. Therefore, we took into account the following hyperparameters: 1) the *model size* (base, small, medium, with 74 M, 244 M and 769 M parameters respectively); 2) *'compression_ratio_threshold'* (2.0, 2.4, 2.8, 3.2) and 3) *'logprob_threshold'* (-1.5, -1.0, -0.5, -0.25). These parameters were chosen as they were close to the default values, which are '2.4' for the compression_ratio and '-1.0' for the logprob_threshold.

The 'compression_ratio_threshold' regulates the degree of audio compression on the input speech. In case of PWA speech, low pitch is very frequent so modulating this normalisation parameter may be useful. Whisper used this compression rate during decoding as a criterion for adjusting its temperature parameter, increasing it when the generated text had a compression rate higher than 2.4 (Radford et al., 2022).

On the other hand, the 'log prob_threshold' regulates the required probability to add a new token to the vocabulary of the ASR. This fine-tuning is particularly helpful when in the PWA might appear frequent neologisms (newly coined word). Lower log-probability thresholds could lead to a bigger vocabulary and more accurate compression, but may also increase computational complexity. Also here Whisper used the average log probability over generated tokens as a criterion for adjusting the temperature during decoding, increasing the temperature when the average log probability fell below -1.0. By selecting values of -1.5, -1.0, -0.5, and -0.25 for 'log prob_threshold', it is possible to evaluate how these thresholds impact the balance between exploring alternative options and maintaining reliability in the generated text.

3.4 Evaluation metrics

The evaluation of the ASR performance was done with the WER based on string edit distance, calculating the least number of steps necessary to convert one string from Whisper output to the string from the actual manual transcription. However, since the WER penalizes also innocuous differences, we

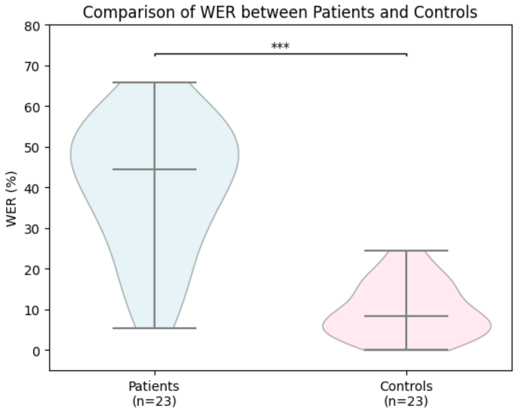


Figure 1: WER distribution density of the *Whisper* model for patients ($N = 23$) and age-matched controls ($N = 23$).
$^{***}: p < 0.001$

had to pre-process the human transcript in CHAT format, similarly to the work of Torre et al. (2021). This procedure is justified as well by the special symbols used that tag phenomena like semantic inconsistencies, repetitions, retracing or sound fragments prevalent in speech from PWA. The CHAT symbols that mark such phenomena have been removed, eliminating also all punctuation.

In the case of neologisms, if the word was not particularly clear, the transcribers would write the literal phonetic alphabet version. For the evaluation of these non-words, they were transformed into the latin alphabet leaving their phoneme sequence unchanged. It should also be noted that human transcriptions included false starts and unique symbols for filler words like "uhuh", "um", and other isolated sounds or interjections, which we decided to preserve since it is a peculiarity of speech from PWA. For the group comparison, we extracted only the participants lines for both the human and ASR transcription, deleting the assessor or carer speech.

3.5 Statistical Analysis

Before modelling the data, to understand the performance difference of Whisper across groups, we compared the WER of patients and controls. In case of repeated measurements, to derive descriptive statistics, we averaged over sessions and then over participants to obtain group characteristics. Instead, the models considered all available information without losing any variability of the data. All the summary outcomes took into account the specific observation weights (e.g. the length of the speech in each audio sample). Due to the non-normality of the distributions and the fact that the samples were independent, we used a Mann-Whitney test.

In addition, a correlation analysis was conducted to establish significant relationships between the WER and CAT scores, as well as the lesion features. Due to the continuous variables considered, we used Pearson correlation coefficients. Furthermore, to pinpoint the associations between our main variables of interest, we used linear mixed-effects models, able to take into account the characteristics of the samples such as repeated measurements and unbalanced data, as well as adjusting the results for potential confounders (Fitzmaurice et al., 2012).

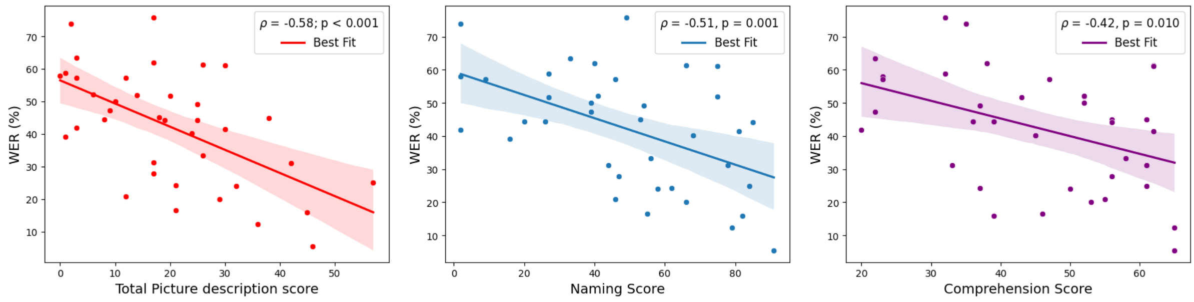
4 Results

Through the grid search optimisation, we generated a total of 48 models, obtained by the combinations of the three aforementioned parameters. The model that performed best, according to the lowest WER, was the one that used the *medium* model, with *compression_ratio_threshold* at 2.0, and *logprob_threshold* at -1.5. The WER differed significantly between controls and patients ($U = 497$, $p < 0.001$, fig.1), patients had an almost four-fold increase in WER than the control group (38.5% vs 10.3%).

Considering the correlation analysis, CAT scores and WER associations were all found to be significant and they are shown in figure 2. All the three scores of CAT showed a negative relation with the outcome, reflecting in general a worse precision of the ASR in the case of patients with more severe aphasia. As far as the stroke lesion volume is concerned, no significant correlation was found.

Since we wanted to adjust results for potential confounders and find significant and meaningful relations, we modelled the data with mixed-effect models and reported the outcomes in table **??**. In total, four main clusters of models were run to evaluate the effects of *lesioned hemisphere*, *lesion presence* considered singularly, *CAT scores*, and *lesion volume* on the abilities of the ASR to transcribe correctly the speech. All the models were adjusted for socio-demographic (age, gender and years of education) and aphasia-related information (time of test since stroke and hours of speech therapy).

Comparing patients with lesions in left and right hemispheres, the ASR performed worse in terms of WER in left *temporal* and *frontal* lobes, as well as in the left parietal lobe, although this last as-



(a) (b) (c)

Figure 2: Association between Word Error Rate of the ASR and the patients' CAT results for (a) *Total Picture description score*, (b) *Naming Score* and (c) *Comprehension Score*.

sociation was not statistically significant. Testing individual lobes confirmed that ASR performance is worse in patients with left frontal lobe lesions, linking it to the localization of expressive language. Moreover, even considering these models, the lesion volume features did not show any significant result.

Finally, CAT derived *Total Picture description score*, *Naming Score* and *Comprehension Score* all had negative relations with the WER, representing higher errors when patients performed poorly in the tests.

5 Discussion

The evaluation of the ASR using the WER metric allowed us to understand how well Whisper, a model trained on a very large healthy speech dataset, performs on PWA speech. We were able to optimise the performance of Whisper based on three hyperparameters, observing similar outcomes in terms of WER when comparing the performance of fine-tuned Whisper model with the performances of previously described ASR systems tailored for PWA.

Using two measures of overt speech production (CAT naming and CAT Total Picture description score for spontaneous speech production) and a measure of speech comprehension (CAT comprehension score), we were able to show that ASR performances is related to the severity of aphasia. These results were confirmed by the mixed-effect models when adjusting for confounding factors such as demographics, time since stroke or duration of therapy. Our findings are in keeping with the study by Torre et al. (2021) that reported 55.5% WER in severe and 22% in mild cases of aphasia.

Furthermore, we showed for the first time that stroke lesion location is related to the performance of the ASR. Speech from patients with left lateralised lesions, and more specifically in the left frontal lobe, was the hardest to recognise using Whisper. This result is consistent with the known localisation of spoken language processing in the brain. Specifically, frontal lobes, together with other parts of the language network, are thought to be primarily implicated in higher-order language functions, such as sentence comprehension, production, speech planning and overt speech production (Geranmayeh et al., 2014). Temporal lobes are essential for language processing and retrieval of semantic information during overt naming (Binder et al. 2020; Binney et al. 2010). Future studies can use information about stroke lesion or brain anatomy to improve ASR training and performance in PWA.

Qualitatively, we noted in some cases Whisper was capable of transcribing filler words (such as "hum", "umm"), frequently observed in PWA. Despite this, the WER occasionally increased as a result of the frequent usage of fillers. False starts (e.g. 'The k- kit- umm... the kitty') were rarely detected and transcribed correctly. There were cases when some words were uttered with low speech volume and were not detected at all, as well as unintelligible words that were skipped altogether by Whisper. These qualitative observations need to be validated with quantitative analysis on larger aphasia-specific datasets to identify PWA speech features that contribute to the worse performance of ASR in PWA. The 'confidence' of the ASR in detecting these aspects can accordingly be reduced, and more specific ASR training can be performed on speech encompassing these specific features.

	WER %				
	Estimate	s.e.	p	95% CI	$\sigma^2_{group}(s.e.)$
Hemisphere Lesion (Left vs Right)					
Temporal (Left)	26.72	12.98	**0.040**	[1.28, 52.16]	137.28 (11.18)
Parietal (Left)	19.79	16.70	0.236	[-12.95, 52.53]	162.68 (12.43)
Frontal (Left)	38.22	13.32	**0.004**	[12.12, 64.32]	128.92 (10.44)
Brain Lobe Lesioned (Yes vs No)					
Temporal					
Left (Yes)	6.96	8.21	0.397	[-9.14, 23.05]	178.74 (12.56)
Right (Yes)	-25.98	11.95	**0.030**	[-49.41, -2.55]	126.47 (9.37)
Parietal					
Left (Yes)	15.54	9.37	0.097	[-2.83, 33.91]	130.29 (10.58)
Right (Yes)	-16.47	15.80	0.297	[-47.44, 14.50]	165.17 (11.86)
Frontal					
Left (Yes)	28.56	10.23	**0.005**	[8.52, 48.61]	92.86 (8.37)
Right (Yes)	-25.98	11.95	**0.030**	[-49.41, -2.55]	126.47 (9.37)
Language Assessments					
Total Picture description score	-0.81	0.23	**0.000**	[-1.26, -0.37]	88.37 (6.98)
Naming Score	-0.32	0.15	**0.037**	[-0.62, -0.02]	109.48 (8.33)
Comprehension Score	-0.64	0.31	**0.041**	[-1.25, -0.03]	165.84 (11.12)
Lesion Volume					
Left Hemisphere Lesion	0.25	0.28	0.378	[-0.30, 0.80]	171 (12.1)
Right Hemisphere Lesion	-1.59	1.28	0.215	[-4.09, 0.92]	161.24 (11.52)
Total Volume	0.20	0.30	0.512	[-0.39, 0.79]	173.1 (12.38)

Table 2: Results of Linear Mixed-Effect regressions on *Hemisphere Lesioned*, the exact *location* of the lesion, *Language Assessments*, and *Lesion Volume*. The models are adjusted for socio-demographic factors (age, gender, and years of education) and aphasia-related information (time of test since stroke and hours of speech therapy).

6 Conclusion and Future Work

This study evaluated the performances of the Whisper end-to-end ASR model on speech derived from patients with post-stroke aphasia. The results highlight the importance of taking lesion location and stroke severity into account when developing speech therapy diagnostics or interventions for PWA using ASR models. Our findings require verification in larger speech databases derived from patients with post-stroke aphasia and their generalisability needs to be assessed in cases of aphasia resulting from other conditions, such as neurodegenerative dementias, which may have different characteristics.

Despite fine-tuning the in-built Whisper parameters to optimise the model performance in this clinical population, we demonstrated that even though Whisper has a competitive performance compared to existing aphasia-specific ASR, it still lacks sufficient clinical diagnostics accuracy. Furthermore, additional ASR metrics such as the confidence of the ASR transcription or the Phoneme Error Rate could be adopted in future research. A further limitation of this work is the small speech database used in this paper. We are actively building a detailed annotated large speech and language database from hundreds of patients with post-stroke aphasia, with the aim of training and developing ASR for pathological speech. We expect that such work will promote greater confidence in the use of AI and specifically NLP for healthcare intervention.

Acknowledgments

G.S. is supported by UK Research and Innovation [UKRI Centre for Doctoral Training in AI for Healthcare grant number EP/S023283/1]. F.G. is funded through MRC P79100. The authors would like to thank Sophie Brook, Joanna Friedland, Katie Stephenson and Chloe Leong for labelling the data, and the reviewers for their valuable feedback.

References

Alberto Abad, Anna Pompili, Angela Costa, Isabel Trancoso, José Fonseca, Gabriela Leal, Luisa Farrajota, and Isabel P Martins. 2013. Automatic word naming recognition for an on-line aphasia treatment system. *Computer Speech & Language*, 27(6):1235–1248.

Myzoon Ali, Patrick Lyden, and Marian Brady. 2015. Aphasia and dysarthria in acute stroke: recovery and functional outcome. *International journal of stroke*, 10(3):400–406.

Sanjit K Bhogal, Robert Teasell, and Mark Speechley. 2003. Intensity of aphasia therapy, impact on recovery. *Stroke*, 34(4):987–993.

Jeffrey R Binder, Jia-Qing Tong, Sara B Pillay, Lisa L Conant, Colin J Humphries, Manoj Raghavan, Wade M Mueller, Robyn M Busch, Linda Allen, William L Gross, et al. 2020. Temporal lobe regions essential for preserved picture naming after left temporal epilepsy surgery. *Epilepsia*, 61(9):1939–1948.

Richard J Binney, Karl V Embleton, Elizabeth Jefferies, Geoffrey JM Parker, and Matthew A Lambon Ralph. 2010. The ventral and inferolateral aspects of the anterior temporal lobe are crucial in semantic memory: evidence from a novel direct comparison of distortion-corrected fmri, rtms, and semantic dementia. *Cerebral cortex*, 20(11):2728–2738.

Caterina Breitenstein, Tanja Grewe, Agnes Flöel, Wolfram Ziegler, Luise Springer, Peter Martus, Walter Huber, Klaus Willmes, E Bernd Ringelstein, Karl Georg Haeusler, et al. 2017. Intensive speech and language therapy in patients with chronic aphasia after stroke: a randomised, open-label, blinded-endpoint, controlled trial in a health-care setting. *The Lancet*, 389(10078):1528–1538.

Katie Chadd, Kathryn Moyse, and Pam Enderby. 2021. Impact of covid-19 on the speech and language therapy profession and their patients. *Frontiers in Neurology*, 12:629190.

Jack S Damico, Nicole Müller, and Martin John Ball. 2010. *The handbook of language and speech disorders*. Wiley Online Library.

Valery L Feigin, Michael Brainin, Bo Norrving, Sheila Martins, Ralph L Sacco, Werner Hacke, Marc Fisher, Jeyaraj Pandian, and Patrice Lindsay. 2022. World stroke organization (wso): global stroke fact sheet 2022. *International Journal of Stroke*, 17(1):18–29.

Garrett M Fitzmaurice, Nan M Laird, and James H Ware. 2012. *Applied longitudinal analysis*. John Wiley & Sons.

Kathleen C Fraser, Frank Rudzicz, and Elizabeth Rochon. 2013. Using text and acoustic features to diagnose progressive aphasia and its subtypes. In *Interspeech*, pages 2177–2181.

Fatemeh Geranmayeh, Robert Leech, and Richard JS Wise. 2016. Network dysfunction predicts speech production after left hemisphere stroke. *Neurology*, 86(14):1296–1305.

Fatemeh Geranmayeh, Richard JS Wise, Amrish Mehta, and Robert Leech. 2014. Overlapping networks engaged during spoken language production and its cognitive control. *Journal of Neuroscience*, 34(26):8728–8740.

Dragos Gruia, Sabia Combrie, and Fatemeh Geranmayeh. 2022. Novel unsupervised comprehensive tool for monitoring vascular cognitive impairment following stroke. In *Alzheimer's Association International Conference*. ALZ.

Norezmi Jamal, Shahnoor Shanta, Farhanahani Mahmud, and MNAH Sha'abani. 2017. Automatic speech recognition (asr) based approach for speech therapy of aphasic patients: A review. In *AIP Conference Proceedings*, volume 1883, page 020028. AIP Publishing LLC.

Zengrui Jin, Xurong Xie, Mengzhe Geng, Tianzi Wang, Shujie Hu, Jiajun Deng, Guinan Li, and Xunying Liu. 2022. Adversarial data augmentation using vae-gan for disordered speech recognition. *arXiv preprint arXiv:2211.01646*.

Helen Kelly, Marian C Brady, and Pam Enderby. 2010. Speech and language therapy for aphasia following stroke. *Cochrane Database of Systematic Reviews*, (5).

Christian Kohlschein, Maximilian Schmitt, Björn Schüller, Sabina Jeschke, and Cornelius J Werner. 2017. A machine learning based system for the automatic evaluation of aphasia speech. In *2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom)*, pages 1–6. IEEE.

Ronald M Lazar and Amelia K Boehme. 2017. Aphasia as a predictor of stroke outcome. *Current neurology and neuroscience reports*, 17(11):1–5.

Duc Le, Keli Licata, Carol Persad, and Emily Mower Provost. 2016. Automatic assessment of speech intelligibility for individuals with aphasia. *IEEE/ACM transactions on audio, speech, and language processing*, 24(11):2187–2199.

Duc Le, Keli Licata, and Emily Mower Provost. 2017. Automatic paraphasia detection from aphasic speech: A preliminary study. In *Interspeech*, pages 294–298.

Duc Le, Keli Licata, and Emily Mower Provost. 2018. Automatic quantitative analysis of spontaneous aphasic speech. *Speech Communication*, 100:1–12.

Duc Le and Emily Mower Provost. 2016. Improving automatic recognition of aphasic speech with aphasiabank. In *Interspeech*, pages 2681–2685.

Yuanyuan Liu, Ying Qin, Siyuan Feng, Tan Lee, and PC Ching. 2018. Disordered speech assessment using kullback-leibler divergence features with multi-task acoustic modeling. In *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 61–65. IEEE.

Brian MacWhinney. 2014. *The CHILDES project: Tools for analyzing talk, Volume II: The database*. Psychology Press.

Seedahmed S Mahmoud, Raphael F Pallaud, Akshay Kumar, Serri Faisal, Yin Wang, and Qiang Fang. 2023. A comparative investigation of automatic speech recognition platforms for aphasia assessment batteries. *Sensors*, 23(2):857.

Website National Aphasia Association. 2016. Aphasia statistics - national survey on aphasia awareness. Accessed: 2022-12-2.

Rebecca Palmer, Pam Enderby, Cindy Cooper, Nick Latimer, Steven Julious, Gail Paterson, Munyaradzi Dimairo, Simon Dixon, Jane Mortley, Rose Hilton, et al. 2012. Computer therapy compared with usual care for people with long-standing aphasia poststroke: a pilot randomized controlled trial. *Stroke*, 43(7):1904–1911.

Rebecca Palmer, Helen Witts, and Timothy Chater. 2018. What speech and language therapy do community dwelling stroke survivors with aphasia receive in the uk? *PloS one*, 13(7):e0200096.

Bart Peintner, William Jarrold, Dimitra Vergyri, Colleen Richey, Maria Luisa Gorno Tempini, and Jennifer Ogar. 2008. Learning diagnostic models using speech and language measures. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4648–4651. IEEE.

Ying Qin, Tan Lee, Anthony Pak Hin Kong, and Sam Po Law. 2016. Towards automatic assessment of aphasia speech using automatic speech recognition techniques. In *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 1–4. IEEE.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Randall R Robey. 1998. A meta-analysis of clinical outcomes in the treatment of aphasia. *Journal of Speech, Language, and Hearing Research*, 41(1):172–187.

Mohamed L Seghier, Elnas Patel, Susan Prejawa, Sue Ramsden, Andre Selmer, Louise Lim, Rachel Browne, Johanna Rae, Zula Haigh, Deborah Ezekiel, et al. 2016. The ploras database: a data repository for predicting language outcome and recovery after stroke. *Neuroimage*, 124:1208–1212.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Simona Spaccavento, Angela Craca, Marina Del Prete, Rosanna Falcone, Antonia Colucci, Angela Di Palma, and Anna Loverre. 2014. Quality of life measurement and outcome in aphasia. *Neuropsychiatric disease and treatment*, 10:27.

James D Stefaniak, Fatemeh Geranmayeh, and Matthew A Lambon Ralph. 2022. The multidimensional nature of aphasia recovery post-stroke. *Brain*, 145(4):1354–1367.

Website Stroke Association. No date. Aphasia and its effects. Accessed: 2022-10-2.

Kate Swinburn, Gillian Porter, and David Howard. 2004. Comprehensive aphasia test.

Iván G Torre, Mónica Romero, and Aitor Álvarez. 2021. Improving aphasic speech recognition by using novel semi-supervised learning methods on aphasiabank for english and spanish. *Applied Sciences*, 11(19):8872.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Textual Entailment for Temporal Dependency Graph Parsing

Jiarui Yao[1] **Steven Bethard**[2] **Kristin Wright-Bettner**[3]
Eli Goldner[1] **David Harris**[1] **Guergana Savova**[1]

[1] Boston Children's Hospital and Harvard Medical School
[2] University of Arizona [3] University of Colorado at Boulder
{firstname.lastname}@childrens.harvard.edu
bethard@email.arizona.edu kristin.wrightbettner@colorado.edu

Abstract

We explore temporal dependency graph (TDG) parsing in the clinical domain. We leverage existing annotations on the THYME dataset to semi-automatically construct a TDG corpus. Then we propose a new natural language inference (NLI) approach to TDG parsing, and evaluate it both on general domain TDGs from wikinews and the newly constructed clinical TDG corpus. We achieve competitive performance on general domain TDGs with a much simpler model than prior work. On the clinical TDGs, our method establishes the first result of TDG parsing on clinical data with 0.79/0.88 micro/macro F1. Our code is available at https://github.com/Jryao/thyme_tdg.

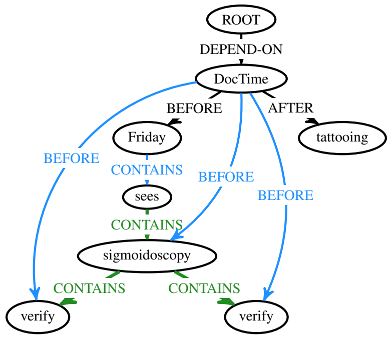


Figure 1: TDG representation for "We will have Dr. Lee perform a flexible **sigmoidoscopy** to **verify** the **tattooing** and to **verify** the location when he **sees** her on Friday." DocTime is the Document Creation Time.

1 Introduction and Background

Temporal information extraction from text is an important part of natural language understanding. Many works have framed temporal relation extraction (RE) as the task of identifying temporal relations between pairs of events, or an event and a time expression (TIMEX3) (Pustejovsky et al., 2003a,b; Cassidy et al., 2014; Styler IV et al., 2014; Ning et al., 2018; Ballesteros et al., 2020; Lin et al., 2021). This pairwise framing can make it hard to decide when to annotate a temporal relation, and the resulting timelines are usually fragmented (Kolomiyets et al., 2012) as not all events or TIMEX3s are linked to each other. Heuristics are typically applied to constrain the search space of pairwise relations, both for annotators and machine learning models. For example, many annotation efforts have constrained temporal relations to adjacent sentences: TempEval (Verhagen et al., 2007, 2010; UzZaman et al., 2013), Clinical TempEval, (Bethard et al., 2015, 2016, 2017), and TimeBank-Dense (Cassidy et al., 2014).

A more principled approach is to model the temporal information in a document as a dependency tree structure (Kolomiyets et al., 2012; Zhang and Xue, 2018b). This approach was extended by Yao et al. (2020) to *temporal dependency graph* (TDG) structure for a more comprehensive representation. An example is shown in Figure 1. With this approach, human annotators inspect each entity and find at most 2 reference times. A complete TDG can be constructed from these decisions. The automatic temporal RE task then becomes a parsing task: produce a TDG as output given a text as input. TDG datasets have been constructed for news articles (Yao et al., 2020) and contracts (Mathur et al., 2022). The current state-of-the-art (SoTA) TDG model (Mathur et al., 2022) reports 0.77 F1 score in the general domain, and 0.64 F1 on the contract dataset, showing the learnability of the TDG approach in those two domains.

In the current work, we make the following contributions:

- We bring TDGs to the clinical domain, by converting the pairwise annotations over the Mayo Clinic electronic health record (EHR) notes in the widely used THYME corpus (Styler IV et al., 2014) to TDGs.[1]

[1]Our THYME-TDG dataset will be available to the research community under the THYME data use agreement procedure.

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 191–199

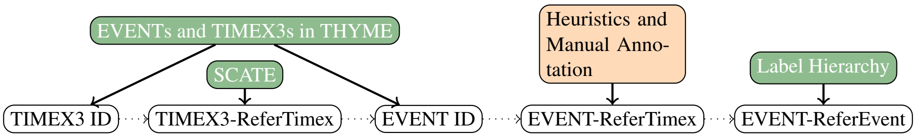


Figure 2: Overview of converting THYME pairwise annotations to TDGs. ID refers to identification, ReferTimex and ReferEvent denote reference timex and reference event respectively. Only the step of identifying reference timex for events requires manual annotation.

- We develop an natural language inference (NLI)-based TDG parser that is much simpler than prior TDG parsers, yet achieves performance competitive with the state-of-the-art in the general domain. On the newly constructed clinical TDG dataset, this parser also achieves strong performance.

Our TDG parser is inspired by works applying NLI-based methods to other information extraction tasks, including relation extraction (Sainz et al., 2021) on the TACRED data set (Zhang et al., 2017), event argument extraction (Sainz et al., 2022) on the ACE (Walker et al., 2006) and WikiEvents (Li et al., 2021) datasets, and biomedical relation extraction (Xu et al., 2022).

2 Creating a Clinical TDG Corpus

A temporal dependency graph (TDG) is defined as a 4-tuple (T, E, M, L), where T is a set of TIMEX3s, E is a set of EVENTs, M is a set of pre-defined "meta" nodes (e.g. ROOT), and L is a set of edges. The definitions and guidelines[2] of Yao et al. (2020) describe the steps to create a TDG from EVENTs and TIMEX3s: (1) For each TIMEX3 t, if t is locatable (i.e., t is not a QUANTIFIER, DURATION or SET), find its reference time expression (*reference timex*), otherwise assign no reference timex; (2) For each EVENT e, find its *reference timex*; and (3) For each EVENT e, find its *reference event* if there is one. Fig. 1 shows examples of such reference decisions, where the reference timex of "Friday" is "DocTime", the reference timex of "sigmoidoscopy" is "DocTime", and the reference event of "sigmoidoscopy" is "sees".

We semi-automate this TDG construction process by leveraging the existing annotations over the THYME corpus (Styler IV et al., 2014). Our approach is visualized in Fig 2. First, we take all the EVENTs and TIMEX3s in the THYME corpus as the building blocks of the graph. In the following steps, we include as many TLINKs (temporal links) from the THYME corpus as possible to maintain the richness and informativeness of the THYME annotations. In some cases, we reverse the TLINK label (e.g. <e_1 BEFORE e_2> becomes <e_2 AFTER e_1>) to make the final graph structure simpler and the annotation process easier (see Appendix A.1).

Temporal Operator	TLINK Labels
Last	BEFORE
Next	AFTER
Before	BEFORE
This	OVERLAP
After	AFTER

Table 1: Mapping SCATE temporal operators to THYME temporal relations.

Identifying the Reference Timex for a TIMEX3. TLINKs between two TIMEX3s are not annotated in the THYME corpus as the temporal relations between a pair of TIMEX3s can be inferred if their normalized values are available.[3] For a locatable TIMEX3, we use the gold temporal operators annotations from the Semantically Compositional Annotation of Time Expressions (SCATE; Bethard and Parker, 2016) to get a TIMEX3-TIMEX3 relation by mapping temporal operators to temporal relations as shown in Table 1.

Identifying the Reference Timex for an Event. Given an event e, we choose the reference timex of e among the TIMEX3s linked to e in the original THYME corpus via TLINKs. If there is only one TIMEX3 temporally related to e, that TIMEX3 is automatically assigned as the reference timex of e. If there are multiple TIMEX3 temporally related to e, but only one TIMEX3 CONTAINS e, that TIMEX3 is automatically selected as the reference timex of e. Otherwise, the instance's reference timex is manually annotated. If e is not TLINKed to any TIMEX3s, DocTime is selected

[2] https://github.com/Jryao/temporal_dependency_graphs_crowdsourcing

[3] https://clear.colorado.edu/compsem/documents/THYME_guidelines.pdf

	Train	Dev	Test
	THYME-TDG		
TDG	1,822	927	998
TIMEX3-TIMEX3	4,328	2,286	2,216
EVENT-TIMEX3	37,042	20,100	18,499
EVENT-EVENT	10,073	5,798	5,443
EVENT-EVENT (THYME+)	11,558	6,579	6,225
EVENT-EVENT %	87.2%	88.1%	87.4%
	general-TDG		
TDG	400	50	50
TIMEX3-TIMEX3	1,952	325	209
EVENT-TIMEX3	12,047	1,717	1,015
EVENT-EVENT	8,725	1,298	706

Table 2: THYME-TDG and general-TDG distributions. EVENT-EVENT % is the percent of EVENT-EVENT relations in THYME+ represented in THYME-TDG.

as its reference timex with the relation between an event and DocTime (DocTimeRel) as the label.

Identifying the Reference Event for an Event. To find the reference event for e among the EVENTs that are TLINKed to e, we define a label hierarchy[4] based on the specificity of different temporal relations to facilitate consistent annotation:

$$\text{BEGINS-ON, ENDS-ON} > \\ \text{CONTAINS, CONTAINS-SUBEVENT} > \\ \text{BEFORE} > \\ \text{OVERLAP, NOTED-ON}$$

If the candidates have the same level of specificity, we choose the one closest to e in textual order.

3 Corpus Statistics

Following the Clinical TempEval tasks, we use the colon cancer set of the THYME+ corpus and the same data splits as described in Wright-Bettner et al. (2020). Our statistics on the Development (Dev) set show that there are on average 155 events per note. Since 91.5% of the TLINKs occur within-section, and all the cross-section TLINKs are CONTAINS-SUBEVENT relations, we split each note into sections, drop the cross-section TLINKs, and create one TDG per section.

With our carefully designed conversion method, we automatically translated most of the pairwise annotations in the THYME+ corpus into TDGs. There were only 72, 35 and 40 events in the Train, Dev and Test sets, respectively, which could not be automatically assigned reference times. Because of this small number, one domain expert manually annotated these missing TDG edges. Statistics of the final THYME-TDG corpus are in Table 2. The EVENT-EVENT rows show that more than 87% of the original event-event TLINKs are represented in the THYME-TDG corpus. Table 2 also shows the statistics of the publicly available TDG dataset (general-TDG; Yao et al., 2020).

4 TDG Parsing

The task of temporal dependency graph parsing is to find the *parent node(s)* for each *child node*, where the child node can be an EVENT or a TIMEX3. We cast the TDG parsing task as a textual-entailment (NLI) task. Given a child node x_i and a list of candidate parent nodes $\{y_1, y_2, \ldots, y_i, \ldots, y_n\}$, our model first verbalizes the possible relations between $<x_i, y_i>$ to generate *hypotheses* using a list of pre-defined templates (shown in Appendix A.3). For example, the verbalization for the BEFORE relation is: x_i *happened before* y_i. To get the *premise*, we concatenate the sentence that contains the child node and the one that contains the parent node.[5] If y_i is not the parent of x_i, the relation label between x_i and y_i will be NO-EDGE. We also add descriptions after the premise regarding the distance between two nodes, following previous works (Zhang and Xue, 2018a; Ross et al., 2020; Yao et al., 2020).[6] Then, we run the NLI model to obtain the probability the premise entails/contradicts/is neutral to the hypothesis.

In the **training** stage, we finetune a pre-trained NLI model on the entailment data generated from the TDG training data. In a TDG, each event can have at most two parents: a reference timex, and a reference event. That is, most of the time, a candidate parent node is not the gold parent for the child, which means the relation between a child node x_i and a candidate parent y_i is NO-EDGE in most cases. To obtain a relatively balanced training set, we first divide the candidate parent nodes into two sets: A and B, where A contains the gold parent nodes for the current child node, and B contains the rest. Then, we randomly sample N_E examples

[4]While the rest of the hierarchy reflects actual specificity, we do not claim that CONTAINS is more specific than BEFORE. We gave preference to CONTAINS here to capitalize on THYME's narrative containers (Styler IV et al., 2014). This is also why we chose CONTAINS as the preferred TLINK type for selecting a reference timex.

[5]Only one sentence will be needed if the two nodes are in the same sentence.

[6]Our preliminary experiments show the additional features can increase the model performance by about 2%.

from A, and N_C examples from B. For each instance sampled from A, we generate an *entailment* example using the gold label, and randomly sample N_C incorrect labels to generate *neutral* examples. For each instance sampled from B, we generate one *contradiction* example (see Appendix A.3 for an example). In the **inference** stage, for each candidate parent node y_i of a child node x_i, we verbalize all possible relations between them and pick the candidate parent with the highest entailment probability as the final parent for x_i.

5 Experimental Setup

We evaluate our model on the two TDG data sets: the general-TDG and THYME-TDG (clinical-TDG).[7] When sampling the training data, we set N_E to 1 and N_C to 3. For general-TDG, to generate the reference timex candidates for each TIMEX3 or EVENT, we include all the TIMEX3s in the document; to generate reference event candidates for each EVENT, we include all the EVENTs from the beginning of the document to two sentences after the child node. For clinical-TDG, we include candidates in the window of 6 sentences before and 4 sentences after the child node. For both data sets, our candidate parent window setting covered > 99% of the cases.

For the general domain TDG parsing, we finetune the roberta-large-mnli (Liu et al., 2019; Williams et al., 2018) model via HuggingFace (Wolf et al., 2020) for 3 epochs. For the clinical domain TDG parsing, we finetune the PubMedBERT-mnli-snli-scinli-scitail-mednli-stsb (Deka et al., 2022) model for 3 epochs. For model initialization, we experimented with 5 random seeds: {42, 52, 62, 72, 82}. See Appendix A.2 for other hyperparameters.

We use gold EVENTs and TIMEX3s as input for the TDG parsers. Parsed <*child, relation, parent*> triples are compared against gold triples to compute F1 scores. On the general-TDG, we report the average F1 scores across all documents (macro-F) following previous practice (Zhang and Xue, 2018a; Yao et al., 2020; Ross et al., 2020; Mathur et al., 2022). On the clinical-TDG data set, we report both macro- and micro- F1 scores.

[7]The TDG dataset of 100 contracts used in Mathur et al. (2022) is not publicly available to the best of our knowledge.

	Dev	Test
BERT-Ranking (Ross et al., 2020)*	0.62	0.71
DocTime (Mathur et al., 2022)*	**0.69**	**0.77**
NLI-based TDG (best)	<u>0.67</u>	<u>0.75</u>
NLI-based TDG (average)	0.66	0.74

Table 3: TDG parsing F1 scores on the general-TDG. Best results bolded, second best underlined. * indicates results from (Mathur et al., 2022). "Best" and "average" refer to the best and average results across 5 seeds.

	Dev	Test
NLI-based TDG (best)	0.88 (0.79)	0.88 (0.79)
NLI-based TDG (average)	0.87 (0.79)	0.88 (0.79)

Table 4: TDG parsing macro F1 (micro F1) scores on the clinical-TDG data set.

6 Results and Discussion

Table 3 shows our general-TDG results. We compare our NLI-based TDG model with two existing supervised models trained for the TDG parsing task: the BERT-Ranking model (Ross et al., 2020)[8] and the DocTime model (Mathur et al., 2022). Both our NLI-based TDG (best) and (average) models outperform the BERT-Ranking model by a large margin, suggesting the advantages of our NLI approach. The NLI-based TDG (average) model and BERT-Ranking model report the average scores of 5 runs, however it is unclear whether the DocTime results are the best or the average.

Compared to the DocTime model (Mathur et al., 2022), our NLI-based TDG model (best) achieved slightly lower but competitive performance, while being much simpler. The DocTime model contains 3 graph neural networks and relies on off-the-shelf NLP tools including a co-reference resolution model, a dependency parser, a pre-trained model for sentence embeddings, and a document-level Rhetorical Structure Theory parser. Mathur et al. (2022) did not list exact tools or configurations (e.g., what is the model used for coreference resolution?) and the code is not publicly available, so it's very hard to re-implement or apply this model to other data sets currently.

We evaluated our NLI-based TDG approach on the newly created clinical-TDG data set (Table 4). This is the first result with a graph algorithm on a clinical temporal relation dataset. Thus, our re-

[8]This model was not evaluated on the TDG data set by the authors, Mathur et al. (2022) ran the experiments on general-TDG and reported the results in their publication.

sults serve as a baseline for future research. Our NLI-based parser achieved promising results on the clinical-TDG data, showing both the utility of this dataset, and the generalizability of our TDG parser.

7 Conclusion

We explore TDG representation and parsing in the clinical domain. We convert the pairwise annotations over the Mayo Clinic EHR notes in the THYME corpus to TDGs semi-automatically. We then develop a NLI-based TDG parser that is much simpler than prior TDG parsers, yet achieves performance competitive with the SoTA in the general domain. On the clinical TDG data set, our parser also achieves strong performance, which can serve as a baseline for future research on clinical TDG parsing.

Limitations

We finetuned pre-trained NLI models for TDG parsing. Both data sets we used were in English. To apply this model to other languages and to get the best results, pre-trained NLI models or NLI data sets might be required for the new language. Templates to verbalize the temporal relations in the new language are also required.

The clinical data set (i.e. THYME) we used in this work only contains EHRs from one institution: Mayo Clinic. Clinicians from different hospitals can have different writing style or use different templates when writing the notes. Future work should test the TDG representation and parsers on EHRs from other institutions, and EHRs of different patient populations.

Acknowledgments

The research was supported by NIH (R01LM010090, R01LM01348). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Library Of Medicine or the National Institutes of Health.

References

Miguel Ballesteros, Rishita Anubhai, Shuai Wang, Nima Pourdamghani, Yogarshi Vyas, Jie Ma, Parminder Bhatia, Kathleen McKeown, and Yaser Al-Onaizan. 2020. Severing the edge between before and after: Neural architectures for temporal ordering of events. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5412–5417, Online. Association for Computational Linguistics.

Steven Bethard, Leon Derczynski, Guergana Savova, James Pustejovsky, and Marc Verhagen. 2015. SemEval-2015 task 6: Clinical TempEval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 806–814, Denver, Colorado. Association for Computational Linguistics.

Steven Bethard and Jonathan Parker. 2016. A semantically compositional annotation scheme for time normalization. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3779–3786, Portorož, Slovenia. European Language Resources Association (ELRA).

Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2016. SemEval-2016 task 12: Clinical TempEval. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1052–1062, San Diego, California. Association for Computational Linguistics.

Steven Bethard, Guergana Savova, Martha Palmer, and James Pustejovsky. 2017. SemEval-2017 task 12: Clinical TempEval. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 565–572, Vancouver, Canada. Association for Computational Linguistics.

Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. 2014. An annotation framework for dense event ordering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 501–506, Baltimore, Maryland. Association for Computational Linguistics.

Pritam Deka, Anna Jurek-Loughrey, et al. 2022. Evidence extraction to validate medical claims in fake news detection. In *International Conference on Health Information Science*, pages 3–15. Springer.

Oleksandr Kolomiyets, Steven Bethard, and Marie-Francine Moens. 2012. Extracting narrative timelines as temporal dependency structures. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 88–97, Jeju Island, Korea. Association for Computational Linguistics.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 894–908, Online. Association for Computational Linguistics.

Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2021. EntityBERT: Entity-centric masking strategy for model pretraining for the clinical domain. In *Proceedings of the*

20th Workshop on Biomedical Language Processing, pages 191–201, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Puneet Mathur, Vlad Morariu, Verena Kaynig-Fittkau, Jiuxiang Gu, Franck Dernoncourt, Quan Tran, Ani Nenkova, Dinesh Manocha, and Rajiv Jain. 2022. DocTime: A document-level temporal dependency graph parser. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 993–1009, Seattle, United States. Association for Computational Linguistics.

Qiang Ning, Hao Wu, and Dan Roth. 2018. A multi-axis annotation scheme for event temporal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1318–1328, Melbourne, Australia. Association for Computational Linguistics.

James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003a. Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003b. The timebank corpus. In *Corpus linguistics*, volume 2003, page 40. Lancaster, UK.

Hayley Ross, Jonathon Cai, and Bonan Min. 2020. Exploring Contextualized Neural Language Models for Temporal Dependency Parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8548–8553, Online. Association for Computational Linguistics.

Oscar Sainz, Itziar Gonzalez-Dios, Oier Lopez de Lacalle, Bonan Min, and Eneko Agirre. 2022. Textual entailment for event argument extraction: Zero- and few-shot with multi-source learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2439–2455, Seattle, United States. Association for Computational Linguistics.

Oscar Sainz, Oier Lopez de Lacalle, Gorka Labaka, Ander Barrena, and Eneko Agirre. 2021. Label verbalization and entailment for effective zero and few-shot relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1199–1212, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

William F. Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, and James Pustejovsky. 2014. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA. Association for Computational Linguistics.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. SemEval-2007 task 15: TempEval temporal relation identification. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 75–80, Prague, Czech Republic. Association for Computational Linguistics.

Marc Verhagen, Roser Saurí, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62, Uppsala, Sweden. Association for Computational Linguistics.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *Linguistic Data Consortium*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Kristin Wright-Bettner, Chen Lin, Timothy Miller, Steven Bethard, Dmitriy Dligach, Martha Palmer, James H. Martin, and Guergana Savova. 2020. Defining and learning refined temporal relations in the clinical narrative. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 104–114, Online. Association for Computational Linguistics.

Jiashu Xu, Mingyu Derek Ma, and Muhao Chen. 2022. Can nli provide proper indirect supervision for low-resource biomedical relation extraction?

Jiarui Yao, Haoling Qiu, Bonan Min, and Nianwen Xue. 2020. Annotating Temporal Dependency Graphs via Crowdsourcing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5368–5380, Online. Association for Computational Linguistics.

Yuchen Zhang and Nianwen Xue. 2018a. Neural ranking models for temporal dependency structure parsing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3339–3349, Brussels, Belgium. Association for Computational Linguistics.

Yuchen Zhang and Nianwen Xue. 2018b. Structured interpretation of temporal relations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 35–45.

A Appendix

A.1 Annotation Details

Given a TLINK <event A, r, event B> in the THYME+ corpus, where event A is the Source (parent), event B is Target (child), and r is the label between them, we reverse the TLINK label in the following case: event B is the Target in multiple TLINKs, while event A is only TLINKed with event B.

In example 1 below, instead of looking for the reference event of **colonoscopy** between **mass** and **bleeding**, it's easier to make **colonoscopy** the reference event of **mass** and **bleeding** by reversing those two TLINKs: <mass, NOTED-ON, colonoscopy> becomes <colonoscopy, NOTED-ON-INV, mass>, and <bleeding, NOTED-ON, colonoscopy> becomes <colonoscopy, NOTED-ON-INV, bleeding>, with "INV" indicating "inverse". The subgraph representation for those 3 TLINKs is showed in Figure 3.

In example 2 below, it's not clear which event among **report, pathology, values** and **notes** should

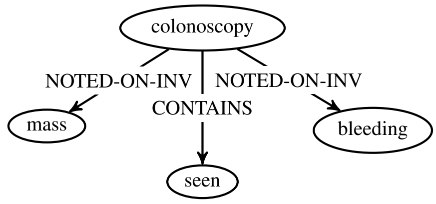


Figure 3: Final graph representation of Example 1.

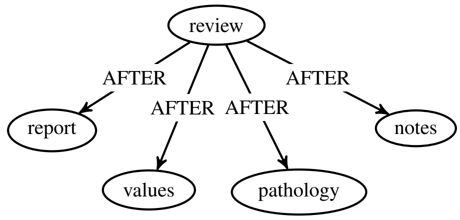


Figure 4: Final graph representation of Example 2.

be the reference event of **review** as they all have the same temporal relation with **review**. However, if we reverse those TLINKs, then **review** will become the reference event of the 4 other events, as shown in Figure 4.

1. *Review of the **colonoscopy** reports indicates that approximately 10-cm from the anal verge a 3- to 4-cm **mass** was **seen** with **bleeding**.*
    - <mass, NOTED-ON, colonoscopy>
    - <colonoscopy, CONTAINS, seen>
    - <bleeding, NOTED-ON, colonoscopy>
2. *I have had the opportunity to **review** the operative **report**, surgical **pathology**, laboratory **values**, and **notes**.*
    - <report, BEFORE, review>
    - <pathology, BEFORE, review>
    - <values, BEFORE, review>
    - <notes, BEFORE, review>

A.2 Implementation Details

For the general-TDG data set, we carried out a grid-search of training epochs in {**3**, 4, 10, 20}, batch size in {16, **32**, 64}, maximum sequence length in {64, **128**}, learning rate in {1e-5, 2e-5, **3e-5**}, and weight decay in {0.1, 0.2, **0.3**}, final parameter settings are in bold.

For the clinical-TDG data set, we experimented with training epochs in {**3**, 4}, batch size in {**16**, 32}, learning rate in {**1e-5**, 2e-5, 3e-5}, and weight decay in {**0.1**, 0.2, 0.3}, final parameter settings are in bold.

Experiments were run on an NVIDIA Titan RTX GPU cluster of 7 nodes. It took 80 - 90 minutes to run one training epoch for both data sets.

A.3 Templates

The templates we used to verbalize temporal relations are listed in Table 5 and Table 6.

We give a concrete example to show how we generate the NLI instances from our TDG data sets. Given an event e_i, let $\{e_1, e_2, e_3\}$ be its candidate reference events, e_2 be the gold reference event, and let BEFORE be the gold relation between e_2 and e_i. The following are the NLI instances we can generate for this example:

- *Entailment*: e_2 happened before e_i.
- *Neutral*: e_2 happened at around the same time as e_i.
- *Contradiction*: During e_3, e_i happened.

Both *Entailment* and *Neutral* examples are generated with the gold candidate event e_2, the difference is that the *Neutral* instance has the wrong label that is randomly sampled from the label set. The *Contradiction* example is generated by randomly sampling an incorrect reference event from the candidates with a random label.

Please note that in both the training and inference stage, entity type constraints are applied when verbalizing a temporal relation. For example, in the general-TDG data set, "included" is only used for event-timex pairs. Therefore, when verbalizing event-event relations, the "included" label will be ignored.

A.4 Features

The linguistic features we used are showed in Table 7.

Label	Template
before	{subj} happened before {obj}
after	{subj} happened after {obj}
overlap	{subj} happened at around the same time as {obj}
included	{subj} happened {obj}
Depend-on	{subj} depended on {obj}

Table 5: Templates we used to verbalize temporal relations in the general-TDG data set. {subj} and {obj} are placeholders for entities.

Label	Template
BEFORE	{subj} happened before {obj}
AFTER	{subj} happened after {obj}
OVERLAP	{subj} happened at around the same time as {obj}
CONTAINS-SUBEVENT	{obj} is a sub-event of {subj}
CONTAINS-SUBEVENT-INV	{subj} is a sub-event of {obj}
NOTED-ON	The {obj} test showed the result {subj}
NOTED-ON-INV	The {subj} test showed the result {obj}
AFTER/OVERLAP	{subj} happened after or overlap {obj}
CONTAINS	During {subj}, {obj} happened
CONTAINS-INV	During {obj}, {subj} happened
Depend-on	{obj} depended on {subj}
BEGINS-ON	{subj} begins on {obj}
ENDS-ON	{subj} ends on {obj}

Table 6: Templates we used to verbalize temporal relations in the clinical-TDG data set. {subj} and {obj} are placeholders for entities. "INV" means "inverse", for example, CONTAINS-INV is the inverse of CONTAINS.

Description
Same sentence
Parent sentence before child sentence
Parent sentence after child sentence
No reference event
Parent is Root
Parent is DCT
Parent is the immediately previous node of the child node
Parent is two nodes before the child node in textual order
Parent is the immediately succeeding node of the child node
Parent node after the child node in text order

Table 7: We describe the sentence distance and node distance between two nodes in natural language, as listed in this table.

Generating medically-accurate summaries of patient-provider dialogue: A multi-stage approach using large language models

Varun Nair
Curai Health
varun@curai.com

Elliot Schumacher
Curai Health
elliot@curai.com

Anitha Kannan
Curai Health
anitha@curai.com

Abstract

A medical provider's summary of a patient visit serves several critical purposes, including clinical decision-making, facilitating hand-offs between providers, and as a reference for the patient. An effective summary is required to be coherent and accurately capture all the medically relevant information in the dialogue, despite the complexity of patient-generated language. Even minor inaccuracies in visit summaries (for example, summarizing "patient does not have a fever" when a fever is present) can be detrimental to the outcome of care for the patient.

This paper tackles the problem of medical conversation summarization by discretizing the task into several smaller dialogue-understanding tasks that are sequentially built upon. First, we identify medical entities and their affirmations within the conversation to serve as building blocks. We study dynamically constructing few-shot prompts for tasks by conditioning on relevant patient information and use GPT-3 (Brown et al., 2020) as the backbone for our experiments. We also develop GPT-derived summarization metrics to measure performance against reference summaries quantitatively. Both our human evaluation study and metrics for medical correctness show that summaries generated using this approach are clinically accurate and outperform the baseline approach of summarizing the dialog in a zero-shot, single-prompt setting.

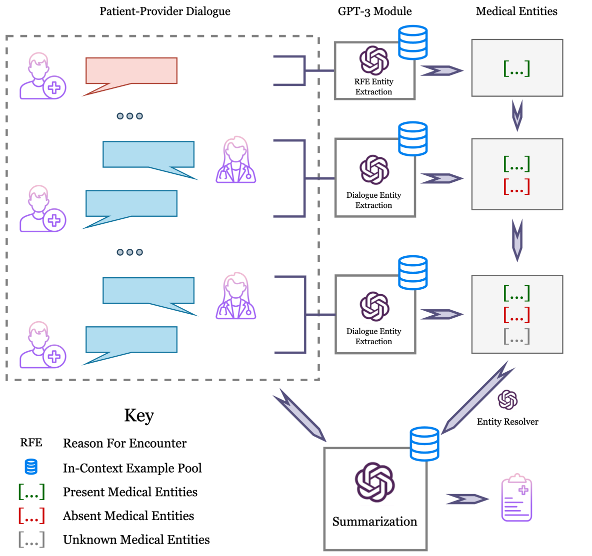


Figure 1: MEDSUM-ENT utilizes a multi-stage approach for medical dialogue summarization with GPT-3 that improves upon naive summarization. The approach utilizes intermediate model calls to extract medical concepts that inform summarization generation.

1 Introduction

A critical clinical task during a medical encounter between a patient and a physician is summarizing the conversation. This summarized note, whether created by a physician or medical assistant, contains important information about the visit and serves as a reference for future patient visits and for the patient. Physicians often spend many hours each week performing such tasks. Charting work, in general, has been identified as a contributing factor to increased rates of physician burnout (Eschenroeder et al., 2021).

Automating medical conversation summarization has been studied with limited success (Pivovarov and Elhadad, 2015; Liang et al., 2019; Gao et al., 2022; MacAvaney et al., 2019; Chintagunta et al., 2021). Some methods try to directly summarize the chat (Enarvi et al., 2020; Zhang et al., 2021) while others pair deep learning methods with information extracted from knowledge bases to produce accurate summaries (Joshi et al., 2020). As base deep learning methods have improved and pre-trained language models specific to summarization such as PEGASUS (Zhang et al., 2020), BART (Lewis et al., 2020), and GPT-3 (Brown et al., 2020) have emerged, we have seen increased fidelity of the summaries generated. However, performance is still not to a reliable standard in prac-

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 200–217
July 14, 2023 ©2023 Association for Computational Linguistics

tical settings for several reasons. First, the lack of labeled clinical data makes it hard to build high-performance fine-tuned models. This reflects lower-than-expected specificity and accuracy in faithfully capturing medical concepts and their affirmations (e.g., present, absent, unknown). Second, custom-trained models need more world knowledge to understand patient language in these conversations and how they map to medical concepts. Third, these models often require breaking conversations into smaller segments to deal with limited context windows. This in turn introduces challenges such as incorrect anaphora and coreference resolution across segmented pieces of the conversation.

The key contributions of this paper include;

- **MEDSUM-ENT**: Inspired by recent works of Chintagunta et al. (2021) and Agrawal et al. (2022), we introduce MEDSUM-ENT: a medical conversation summarization model that takes a multi-stage approach to summarization, using GPT-3 (Brown et al., 2020) as the backbone. MEDSUM-ENT (Fig. 1) grounds the task by first extracting medical entities and their affirmations. These extractions are included as additional input that informs the final summarization step through prompt chaining (Wu et al., 2022). MEDSUM-ENT also exploits few-shot prompting for medical concept extraction and summarization through in-context example selection.

  In both qualitative physician analysis of medical dialogue summaries and quantitative metrics, MEDSUM-ENT generates clinically accurate summaries and produces summaries that are preferable to a zero-shot, single prompt baseline.

- **Automated metrics**: Quantitative metrics are hard to design for generative tasks. We extend proxy metrics of Joshi et al. (2020) by leveraging GPT-3 to compare the coverage of the presence of medical entities in the generated texts. Beyond only identifying exact matches, our approach better accounts for paraphrasing those medical events within the larger text.

2 Methods

We now detail the components of our MEDSUM-ENT framework for medical dialogue summarization, represented in Figure 1.

Medical Entity Extraction To highlight clinical concepts, we extract medical entities (symptoms, diseases etc.) and their affirmation status of either **present**, **absent**, or **unknown**. These entities and their status will be used as additional inputs to the final summarization step.

We first perform entity extraction on the patient's first message of the encounter, which is often lengthy and information dense. We call this message the *reason for encounter* (RFE). Conversational turns between the medical provider and the patient follow the RFE. We also extract medical entities from the conversation, one provider and one patient turn at a time. To accommodate these two types of texts, we use two different prompts, included in Prompt 1 (for RFE entity extraction) and Prompt 2 (for dialogue entity extraction). Both prompts are populated with in-context examples (see In-Context Example Selection) along with the patient's age and sex. The final list of entities in the dialogue is obtained by collating all entities extracted across the RFE and all dialogue turns.

Additionally, we also use an entity resolver similar to those used in Agrawal et al. (2022) to resolve entities in the unknown entities list whose status may have changed during the dialogue (see Prompt 3). For instance, a dialogue turn pair may not have enough information to definitively assign a present or absent status and thus an entity is "unknown". A later dialogue turn may contain information that changes that assignment. By introducing this refinement step, we reduce mistakes in the "Pertinent Unknowns" section of the summary (see Table 1).

Summarization Given a list of medical entities, we summarize the medical dialogue using the dialogue and the entities as input. Our summaries are structured into six sections: *Demographics and Social Determinants of Health*, *Medical Intent*, *Pertinent Positives*, *Pertinent Negatives*, *Pertinent Unknowns*, and *Medical History* (see Prompt 4 for details).

In-Context Example Selection For the entity extraction and summarization modules, we compare semantic-similarity and random in-context example selection. Semantic-similarity-based selection selects labeled examples from a pool using the patient's age, sex, and the query point. Random selection randomly selects in-context examples from these pools to populate our prompts. Further implementation details are in Appendix A.1.

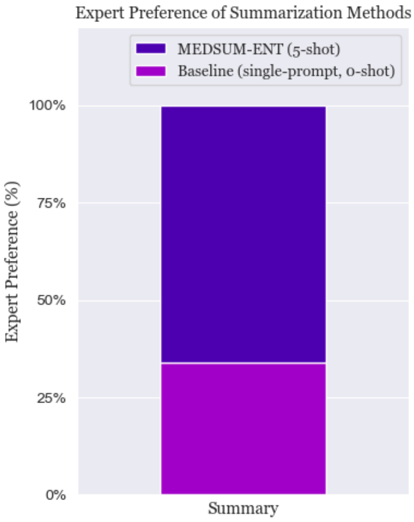


Figure 2: Results of human expert evaluations show MEDSUM-ENT (5-shot) is preferred 66% to 34% over a single-prompt, 0-shot naive summarization baseline.

3 Experiments

Dataset: We use a dataset of 100 clinical encounters of dialogue-summary pairs that occurred between a licensed physician and a patient on a telehealth platform over chat. Encounters in this dataset cover a variety of common presentations in telehealth, including urinary tract infections, back-/abdominal pains, toothaches, and others. All data was de-identified and scrubbed for protected health information prior to experimentation. Conversations contain 46 dialogue turns on average (min of 8 turns, max of 92 turns) and an average of 2342 unigram tokens per encounter. Ground truth summaries were created by using `text-davinci-002` on encounter data to generate an initial summary, which physicians then edited for correctness.

Baselines/Ablations: We compare MEDSUM-ENT to a "naive" zero-shot, single-prompt baseline (i.e. without chaining) that prompts GPT-3 to summarize the conversation (see Prompt 5). For MEDSUM-ENT, we evaluate extraction k-shot configurations (1,3,5-shot) and in-context example selection methods (semantic-similarity based, random) for entity extraction. We use RFE and dialogue entity extraction prompts in at least a 1-shot configuration for MEDSUM-ENT to ensure valid output. Our summarization prompt for baselines and MEDSUM-ENT cannot go beyond 1-shot due to token limits. All experiments are run once and use GPT-3 (davinci-003) (see Appendix A.2 for prompt settings).

3.1 Evaluation Metrics

Expert Evaluation We also asked four doctors, who serve tele-health patients, to judge between the MEDSUM-ENT and baseline-generated summaries on three points on a random set of 50 encounters. For a given encounter, we asked 1) for preference between baseline and MEDSUM-ENT summaries, 2) what amount of clinical information was captured in MEDSUM-ENT's summaries, and 3) about the presence of clinically harmful information in MEDSUM-ENT summaries (see Appendix A.3 for full instruction details).

GPT-Driven Automated Summarization Metrics: Acknowledging the challenges in automatic evaluations of summarization (Peyrard, 2019; Goyal et al., 2022), we focus on quantitatively evaluating the correctness/faithfulness of capturing medical concepts and their affirmation status.

We extend the approach to metrics in Joshi et al. (2020) to have two components, both powered by GPT-3: a medical concept extractor (Appendix Prompt 6) and a verifier (Appendix Prompt 7). The verifier checks if the concepts extracted from one piece of text are present in another and permits the same medical concept extracted or written in different ways to count towards a true positive. For example, for the "Pertinent Positives" section, the predicted value may be *"Patient has back pain and COVID-19"* and the resulting concepts ["back pain", "COVID-19"] and the ground-truth *"Patient has COVID and some pain in the backside"* with concepts ["COVID", "pain in the back"]. Prior metrics that rely on verbatim matches would fail to recognize the predicted text as correct. We define the following metrics:

GPT-Recall: We extract medical entities from both the predicted text and ground-truth text of the same summary section. We use the verifier to infer if the entities extracted from the ground-truth section are also present in the predicted text. This produces tp_{gt} and f_n, which is used to calculate GPT-Recall $= \frac{tp_{gt}}{tp_{gt}+f_n}$.

GPT-Precision: We verify concepts extracted from the predicted section are also present in the ground-truth text, either as exact matches or rephrasings. This produces tp_{pred} and f_p, which is used to calculate GPT-Precision $= \frac{tp_{pred}}{tp_{pred}+f_p}$.

GPT-F1 is the harmonic mean of GPT-Precision and GPT-Recall. Note our approach maintains the integrity of recall and precision (neither score can

Method	Extraction K-shot	Summarization K-shot	Example Selection	Entity Resolver	GPT-F1 (↑) Pertinent Positives	Pertinent Negatives	Pertinent Unknowns	Medical History	Average
Naive	-	0-shot	-	-	**72.9**	**71.7**	45.4	43.9	58.5
	-	1-shot	semantic	-	71.0	69.5	42.1	48.3	57.7
	-	1-shot	random	-	69.4	69.1	47.5	44.7	57.7
MEDSUM-ENT	1-shot	1-shot	semantic	✓	72.4	70.1	50.0	46.2	59.7
	1-shot	1-shot	random	✓	71.4	71.1	54.0	48.3	61.2
	3-shot	1-shot	semantic	✓	71.9	69.0	42.5	47.0	57.6
	3-shot	1-shot	random	-	72.1	69.4	46.4	45.8	58.4
	3-shot	1-shot	random	✓	72.2	70.9	**55.8**	**50.4**	**62.3**
	5-shot	1-shot	semantic	✓	71.8	70.2	46.6	46.3	58.7
	5-shot	1-shot	random	✓	71.9	68.3	51.9	48.2	60.0

Table 1: Results of GPT-driven metrics. Performance across “Pertinent Positives”, “Pertinent Negatives” sections are fairly consistent across methods. MEDSUM-ENT demonstrates consistently improved performance in the “Pertinent Unknowns” and “Medical History” sections. Surprisingly, we also find consistently higher performance across experiments using random in-context example selection over semantic-similarity-based selection.

take on a value > 1). We evaluate MEDSUM-ENT via the GPT-Precision and GPT-Recall metrics described in section 3.1 on all 100 clinical encounters.

4 Results

Table 1 shows quantitative metrics on summaries produced by the baselines and MEDSUM-ENT. Both generated summaries are compared to the ground truth summaries. We see that while GPT-F1 performance for “Pertinent Positives” and “Pertinent Negatives” is consistent across methods, MEDSUM-ENT’s ability to capture the “Pertinent Unknowns” and “Medical History” pushes its average consistently above that of the naive zero-shot, non-chained baseline. These sections are crucial to include correctly as they often influence clinical decision-making. Also, the Unknown Entity Resolver improves performance specifically in the “Pertinent Unknowns” section (ablated in rows 7 vs. 8 with 46.4 vs. 55.8 for with and without the resolver). The “Demographics and Social Determinants of Health” and “Medical Intent” sections have nearly identical, accurate output across all experiments, so we do not calculate metrics for them. See Appendix A.4 for example generated summaries.

We find two surprising results. First, there is no correlation between a larger k-shot and increased performance. This may demonstrate diminishing returns of GPT-3 to perform medical concept extraction. Furthermore, the use of semantic similarity to select in-context examples performs **worse** than randomly selecting examples. This follows Ye et al. (2022) which claims diversity of in-context samples is more important than similarity.

In our expert human evaluations, Figure 2 demonstrates MEDSUM-ENT (5-shot, semantic) summaries are preferred over the baseline summaries 66% to 34%. Our expert evaluators also rate MEDSUM-ENT capturing all relevant medical information in 40% of evaluated summaries, most information in 48%, some information in 12%, and zero information in 0%. This provides further qualitative evidence for MEDSUM-ENT’s ability to effectively summarize. However, our expert evaluators also rate 28% of the summaries evaluated as containing incorrect information that could harm the patient if acted on by medical providers. Often these are due to misattributed symptoms and conditions (e.g., symptoms marked as absent but were present, missed medication allergies). This is consistent with the GPT-F1 measures for pertinent positives and negatives in Table 1 and highlights the challenge involved in deploying a system such as MEDSUM-ENT. Further work is needed to trust such systems in the wild.

5 Conclusion

We introduce MEDSUM-ENT, a multi-stage framework for medical dialogue summarization that modularizes summarization into multiple stages that extract and refine medical entities from dialogue turns. Through human evaluation and quantitative metrics, we show that this method is clinically accurate and preferable to naive zero-shot summarization with GPT-3. We hope that future work can investigate refinement modules and iterative summarization further and conduct wider expert human evaluation studies to better understand challenges in bringing model-assisted summarization to medical providers in the near term.

6 Limitations

The experiments in this paper were performed using OpenAI's GPT-3 API. While running locally does not require a large amount of computational resources, the server-side service cannot be easily replicated and requires a large amount of computational resources. Additionally, given the inherently restrictive nature of medical text, we can only evaluate our approach on a small corpus of English-language dialogues taken from the dataset of a single company's medical service, which we cannot release due to privacy concerns. Finally, given summarization is a challenging task to evaluate, we rely on a small number of expert human annotators and automatic metrics. However, additional annotations may be helpful and it may also help to study and report labeler agreement when reporting human preferences.

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are zero-shot clinical information extractors. *arXiv preprint arXiv:2205.12689*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2021. Medically aware GPT-3 as a data generator for medical dialogue summarization. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 66–76, Online. Association for Computational Linguistics.

Seppo Enarvi, Marilisa Amoia, Miguel Del-Agua Teba, Brian Delaney, Frank Diehl, Stefan Hahn, Kristina Harris, Liam McGrath, Yue Pan, Joel Pinto, Luca Rubini, Miguel Ruiz, Gagandeep Singh, Fabian Stemmer, Weiyi Sun, Paul Vozila, Thomas Lin, and Ranjani Ramamurthy. 2020. Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, pages 22–30, Online. Association for Computational Linguistics.

Jr Eschenroeder, H. C, Lauren C Manzione, Julia Adler-Milstein, Connor Bice, Robert Cash, Cole Duda, Craig Joseph, John S Lee, Amy Maneker, Karl A Poterack, Sarah B Rahman, Jacob Jeppson, and Christopher Longhurst. 2021. Associations of physician burnout with organizational electronic health record support and after-hours charting. *Journal of the American Medical Informatics Association*, 28(5):960–966.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, Dongfang Xu, Matthew M. M. Churpek, and Majid Afshar. 2022. Summarizing patients' problems from hospital progress notes using pre-trained sequence-to-sequence models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2979–2991, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3755–3763, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jennifer Liang, Ching-Huei Tsou, and Ananya Poddar. 2019. A novel system for extractive clinical note summarization using EHR data. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 46–54, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Sean MacAvaney, Sajad Sotudeh, Arman Cohan, Nazli Goharian, Ish Talati, and Ross W Filice. 2019. Ontology-aware clinical abstractive summarization. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1013–1016.

Maxime Peyrard. 2019. Studying summarization evaluation metrics in the appropriate scoring range. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5093–5100, Florence, Italy. Association for Computational Linguistics.

Rimma Pivovarov and Noémie Elhadad. 2015. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association*, 22(5):938–947.

Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Xi Ye, Srini Iyer, Asli Celikyilmaz, Ves Stoyanov, Greg Durrett, and Ramakanth Pasunuru. 2022. Complementary explanations for effective in-context learning. *ArXiv*, abs/2211.13892.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Longxiang Zhang, Renato Negrinho, Arindam Ghosh, Vasudevan Jagannathan, Hamid Reza Hassanzadeh, Thomas Schaaf, and Matthew R Gormley. 2021. Leveraging pretrained models for automatic summarization of doctor-patient conversations. *arXiv preprint arXiv:2109.12174*.

A Appendix

A.1 Dynamic example selection

We create labeled in-context example pools for RFE entity extraction and dialogue entity extraction using physician labels for what medical concepts would have been extracted and created a summarization pool using physician-written dialogue summaries. The dialogue summaries for this pool were created by physicians editing the outputs of summaries created by text-davinci-002. Semantic-similarity based example selection is implemented using nearest-neighbor search with the LangChain[1] and FAISS (Johnson et al., 2019) libraries.

A.2 Experiment details

Prompt	temperature	max_tokens	top_p
RFE Medical Entity Extr.	0.1	200	1.0
Dialogue Medical Entity Extr.	0.1	200	1.0
Unknown Entity Resolver	0.1	200	1.0
Summarization	0.7	512	1.0
Metric: Medical Entity Extr.	0.0	200	1.0
Metric: Medical Entity Verif.	0.0	200	1.0

Table 2: Experimental settings for all prompts used in this work, no hyper-parameter search was run to obtain these values. We use lower temperature values for model calls where we expect lower variability in its inputs (summarization takes in dialogues and list of medical entities of varying lengths and sizes respectively, thus has a higher temperature). Running the metric concept extraction and verification prompts at a temperature of 0 ensures maximal reproducibility of metric computation. Each experiment (line in Table 1 took approximately 3 hours to run, with exponential back-off used during GPT-3 queries.)

A.3 Expert evaluation

To qualitatively evaluate our summaries, we conducted physician evaluations focused on three questions:

- Q1: *How often are summaries written using MEDSUM preferred over naively generated summaries?*
- Q2: *What fraction of relevant clinical information is captured in the summaries generated by our method? (All, Most, Some, None)*
- Q3: *Does the summary generated by our method contain incorrect information that could significantly alter the course of treatment and potentially harm the patient if*

[1] https://github.com/hwchase17/langchain

this summary was used by another medical provider?

Q1 was asked alongside some basic instructions for how choices should be made, shown below:

For this task, you will see a dialogue, RFE, and age/sex information for a patient. The task is to identify which of the 3 summary options you would prefer to use as a visit summary. You may use your own discretion in selecting which of the 3 options you prefer. Some things to note when selecting are:

- How thorough and clinically accurate is each summary?
- Is the summary missing clinically relevant information?
- Does the summary contain extraneous information that is harmful if a provider were to read and act upon information in the summary?
- Which summary is stylistically preferable and/or easier to read?

A.4 Qualitative Analysis

We provide two examples of outputs from our naive 0-shot, single-prompt baseline and MEDSUM (5-shot, semantic) below. All protected health information is anonymized using tags in all caps (eg. DATE_1, LOCATION_1, NAME). Note the green text which captures correctly summarized information that is not present in the other method's output, and red text which marks a mistake that is not present in the other method's output.

Chat A. *Reason for Encounter:* UTI

Doctor: Do you have any pain or discomfort when you urinate ?
Patient: present
Doctor: Do you experience sudden urges to urinate ?
Patient: present
Doctor: Have you been urinating more frequently than usual ?
Patient: present
Doctor: Have you noticed any discharge from the urethra ?
Patient: present
Doctor: Have you been sexually active in the past 3 months ?
Patient: present
Doctor: Do you have to urinate at night , waking up two or more times to urinate ?
Patient: present
Doctor: Do you have pain on the side of your torso just below your ribs ?
Patient: absent
Doctor: Does your abdomen or belly hurt ?
Patient: unknown
Doctor: Since when are you having these symptoms ?
Patient: 2 weeks ago
Doctor: Have you taken anything for these symptoms ?
Patient: Azo
Doctor: Has it helped ?
Patient: Somewhat not much
Doctor: Can you describe the urine ?
Patient: Lots of it Bloated Orange when I take that medicine Dark yellow the rest
Doctor: Is there any blood in the urine ?
Patient: No
Doctor: Does the urine have any foul smell ?
Patient: No
Doctor: Do you have pain below your belly button ?
Patient: present Some not much
Doctor: Have you noticed any difficulty passing urine ?
Patient: absent
Doctor: Do you have pain with sexual intercourse ?
Patient: present This is when it started
Doctor: Do you have pain in lower abdomen lower belly ?
Patient: present
Doctor: Have you noticed any vaginal discharge ?
Patient: unknown
Doctor: Do you have low back pain ?
Patient: present
Doctor: Does the outer part of your vagina itch ?
Patient: absent
Doctor: Do you have a fever ?
Patient: absent
Doctor: Do you have nausea ?
Patient: present
Doctor: What was the first day of your last menstrual period ?
Patient: I have no cycle hysterectomy
Doctor: When were you last sexually active ?
Patient: Two weeks ago
Doctor: Have you had a UTI in the past ?
Patient: Yes
Doctor: When was it [NAME] and what was done ?
Patient: Six months ago antibiotics

MEDSUM Summary for Chat A

Demographics and Social Determinants of Health:
A 46 year old female.

Patient Intent:
Patient came for UTI.

Pertinent Positives:
Patient reports pain and discomfort when urinating, sudden urges to urinate, increased urination frequency, urethral discharge, sexual activity in the past 3 months, orange urine when taking medication, dark yellow urine the rest, abdominal pain below the belly button, pain with sexual intercourse when the symptoms started, lower abdominal pain, low back pain and nausea. **Symptoms have been present for 2 weeks.** She has taken Azo to manage the symptoms which have only slightly helped.

Pertinent Unknowns:
Patient is unsure if her urine has any foul smell and if there is any vaginal discharge.

Pertinent Negatives:
Patient reports no fever, no difficulty passing urine, no vulvar pruritus, and no hematuria.

Medical History: Patient has a hysterectomy and was last sexually active two weeks ago. She had a UTI six months ago and was treated with antibiotics.

Baseline Summary for Chat A

Demographics and Social Determinants of Health:
The patient is a 46 year old female.

Patient Intent:
The patient is seeking care for UTI.

Pertinent Positives:
The patient has experienced pain or discomfort when urinating, sudden urges to urinate, increased urination frequency, discharge from the urethra, and pain with sexual intercourse in the past 3 months. She has also experienced pain on the side of her torso below her ribs, and pain in her lower abdomen. She has taken Azo, which somewhat helped her symptoms. The urine is abundant and bloated, orange when she takes the medicine, and dark yellow the rest of the time.

Pertinent Unknowns:
The patient is unsure of **any abdominal pain**, and if there is any foul smell to the urine. She is also unsure if there is any vaginal discharge.

Pertinent Negatives:
The patient does not have any blood in the urine, and the outer part of her vagina does not itch. She also does not have a fever or difficulty passing urine.

Medical History:
The patient has had a hysterectomy and was last sexually active two weeks ago. She has had a UTI in the past, which was six months ago, and she was treated with antibiotics.

Chat B. *Reason for Encounter:* I took my inhaler a few minutes and realized i am almost out of it. I was woundering if i can get a prescription from you? I have asthma and the last few days tight chest and breath with weezing issues

Doctor: Thanks for confirming Sorry to hear about the symptoms you are currently experiencing , [NAME] . May I know when you were diagnosed with asthma ?
Patient: When i was born I was dx Last asthma attack has been at least 2 years
Doctor: Thanks for letting me know about it . Which inhaler have you been prescribed on and how long have you been using it ?
Patient: Albuteraol
Doctor: Okay . Would you be able to share a picture of the inhaler with the last prescription , that clearly mentions about the dosage and frequency ?
Patient: Prn . I do not have the prescription . I got it almost 2 years ago
Doctor: Thanks for sharing , [NAME] . Have you been using it since childhood ?
Patient: I have been using this one only on prn bases . In the past i have used a steroid one spary twice a day . Do not remember the name
Doctor: Okay . How often do you generally use the inhaler and how many puffs do you use each time ?
Patient: Albuterol i use maybe best guess once a month too once every two months . When i use it two sparys . I take one wait for 5 minutes and repeat
Doctor: Okay . When was the last
Doctor:s visit ?
Patient: In January . Just lost my job so i dont have insurance to go back right now
Doctor: I hear you , [NAME] . Sorry to hear about the job loss . How long have you been experiencing these symptoms now ? . Have you noticed any trigger factors associated with them ? Anything that could have resulted in the symptoms or worsened it ?
Patient: Possible since i am obsessed 340 lbs . But the weather here has been muggy hot humidity has been off the charts . Not normal for LOCATION_1
Doctor: Thanks for letting me know about it . Im going to send you questions to make sure Im correctly collecting all of your symptoms . Please select “Yes” , “No” , or “Unsure” in the question and hit Send once youve input your response Do you get more short of breath than expected with activity ?
Patient: present Lately yes
Doctor: Do you have a dry cough ?
Patient: present Chronic
Doctor: Do you have a cough that brings up phlegm or mucus ?
Patient: present Clear
Doctor: Do you have chest pain that worsens when you take a deep breath ?
Patient: absent
Doctor: Do you wake up from sleep and find yourself out of breath ?
Patient: absent
Doctor: Do you feel short of breath while lying down ?
Patient: absent
Doctor: Do you have shortness of breath at rest or while sitting ?
Patient: absent
Doctor: Do you have difficulty falling asleep ?
Patient: absent
Doctor: Do you frequently find yourself waking up in the middle of the night ?
Patient: present
Doctor: Do you feel fatigued ?
Patient: present Possible due to the over weight that i am currently working on loosing
Doctor: Have you recently been exposed to animals ?
Patient: absent
Doctor: Have you recently been in close contact with someone diagnosed with COVID-19 ?
Patient: absent
Doctor: Do you have difficulty breathing when you lean forward ?
Patient: present
Doctor: Do you have difficulty speaking more than short phrases due to the breathing difficulty ?
Patient: absent
Doctor: How long have you been noticing all of these symptoms , [NAME] ?
Patient: Short of breath leaning forward past 30 minutes . Took two puffs starting to feel ok NAME_1... was dx with sinus issues . Always draining

Doctor: Has the breathing difficulty , wheezing , chest pressure started only today within the past hour ?

Patient: Wheezing yesterday . Chest pressure and wheezing only today

MEDSUM Summary for Chat B

Demographics and Social Determinants of Health:
A 43 year old male. He recently lost his job and currently lacks insurance. He is currently working on losing weight and the weather in LOCATION_1 has been extremely muggy and hot.

Patient Intent:
Patient came to address his concern of needing a prescription for his inhaler and to address his asthma symptoms of tight chest, shortness of breath, and wheezing.

Pertinent Positives:
Patient has asthma since birth, he has been using an Albuterol inhaler as needed and a steroid spray twice a day in the past, he has a dry cough, difficulty breathing when leaning forward, is coughing up phlegm, **has chest pressure, wakes up in the middle of the night, is fatigued,** has sinus issues, and has wheezing.

Pertinent Unknowns:
Patient is unsure of the dosage and frequency of his prescription medication and the name of his steroid medication.

Pertinent Negatives:
Patient has no chest pain, **difficulty speaking more than short phrases because of breathing difficulty**, shortness of breath while lying down, difficulty initiating sleep, mucus, **animal exposure, or contact with someone diagnosed with COVID19**.

Medical History:
Patient was diagnosed with asthma at birth, his last asthma attack was at least 2 years ago, he has been using an Albuterol inhaler as needed for the past 2 years, and he has been experiencing symptoms for the past 30 minutes.

Baseline Summary for Chat B

Demographics and Social Determinants of Health:
43 year old male patient, unemployed and obese (340 lbs). Living in a hot, humid area with high levels of humidity.

Patient Intent:
Patient is requesting a prescription for an inhaler, as the patient is almost out of it and has been experiencing tightness in their chest and wheezing issues for the past few days.

Pertinent Positives:
Patient has asthma, has been using Albuterol inhaler on a prn basis, has a chronic dry cough, and has been noticing shortness of breath and difficulty leaning forward for the past 30 minutes.

Pertinent Unknowns:
Patient is unsure if the weather or being overweight is causing the symptoms.

Pertinent Negatives:
Patient does not have chest pain, difficulty breathing when lying down, difficulty speaking, difficulty falling asleep, or **waking up in the middle of the night due to breathing difficulty.**

Medical History:
Patient was diagnosed with asthma at birth and had their last asthma attack 2 years ago. Patient was last at the doctor in January and has been experiencing the symptoms for the past few days. Patient has been using the current inhaler for 2 years, and has used a steroid inhaler twice a day in the past, but can not remember the name. Patient has been diagnosed with sinus issues and has been having clear mucus drainage.

```
----
Below is the first message from a{age_and_sex} patient seeking care:
----
Patient: {rfe}
----
Using the patient's message above, please find the medical entities (medical concepts, symptoms, or medical conditions) and each one's status (present, absent, or unknown) that would be important for a doctor to know.
If the patient states the presence of a medical concept, symptom or condition, the medical entity's status should be present.
If the patient denies the presence of a medical concept, symptom or condition, the medical entity's status should be absent.
Medical entities should have an unknown status ONLY if the patient is themselves unsure or hesitant about a medical entity (eg: an answer of "Unknown", or "I'm not sure about..."").
Do NOT add a medical entity as unknown if the uncertainty is due to a DATE_1, DATE_2, NAME, or LOCATION tag. If there is such a medical entity associated with a tag, it must be either positive or negative.
Only extract medical entities that exist in the patient's message. DO NOT EXTRACT NON-MEDICAL ENTITIES.
Each medical entity should belong to one of six categories: Demographics and Social Determinants of Health, Patient Intent, Pertinent Positives, Pertinent Negatives, Pertinent Unknowns, or Medical History.
----
```

Prompt 1: Prompt for reason for encounter (RFE) medical entity extraction.

```
----
Below is a dialogue between a doctor and a{age_and_sex} patient seeking care:
----
{dialogue}
----
Using the patient's message above, please find the medical entities (medical concepts, symptoms, or medical conditions) and each one's status (present, absent, or unknown) that would be important for a doctor to know.
If the patient states the presence of a medical concept, symptom or condition, the medical entity's status should be present.
If the patient denies the presence of a medical concept, symptom or condition, the medical entity's status should be absent.
Medical entities should have an unknown status ONLY if the patient is themselves unsure or hesitant about a medical entity (eg: an answer of "Unknown", or "I'm not sure about..."").
Do NOT add a medical entity as unknown if the uncertainty is due to a DATE_1, DATE_2, NAME, or LOCATION tag. If there is such a medical entity associated with a tag, it must be either positive or negative.
Only extract medical entities that exist in the patient-physician dialogue. DO NOT EXTRACT NON-MEDICAL ENTITIES.
Each medical entity should belong to one of six categories: Demographics and Social Determinants of Health, Patient Intent, Pertinent Positives, Pertinent Negatives, Pertinent Unknowns, or Medical History.
----
```

Prompt 2: Prompt for dialogue medical entity extraction.

```
---
Below are medical entities (concepts, symptoms, or conditions) extracted from a medical conversation between a 22 year old patient and doctor.

Your job is to clean up the "Unknown Entities" list given the patient-doctor dialogue and a list of positive and negative entities.

Remove any entities that that not are medical entities, or any entities that are unnecessary or completely irrelevant entities given the dialogue, positive entities, and negative entities.

If a similar entity is both present and unknown, or, both absent and unknown, keep it in the unknowns ONLY if this information is still unknown after the entire dialogue.

---Dialogue---
//Example dialogue
---Dialogue---

Positive Entities: cough, headache, lower back pain

Negative Entities: fever, chest pain, chest tightness

Unknown Entities: past episode of flu, age_1, covid vaccination, symptoms, cough, fever, difficulty breathing, runny nose, frequency of headache, headache
---
Cleaned Unknown Entities: past episode of flu, covid vaccination, difficulty breathing, runny nose
---
---
Below are medical entities (concepts, symptoms, or conditions) extracted from a medical conversation between a {age_and_sex} patient and doctor.

Your job is to clean up the "Unknown Entities" list given the patient-doctor dialogue and a list of positive and negative entities.

Remove any entities that that not are medical entities, or any entities that are unnecessary or completely irrelevant entities given the dialogue, positive entities, and negative entities.

If a similar entity is both present and unknown, or, both absent and unknown, keep it in the unknowns ONLY if this information is still unknown after the entire dialogue.

---Dialogue---
Patient: {rfe}
{dialogue}
---Dialogue---

Positive Entities: {positive_entities}

Negative Entities: {negative_entities}

Unknown Entities : {unknown_entities}
---
Cleaned Unknown Entities:
```

Prompt 3: Prompt for resolving unknown entities.

```
Below is a medical encounter between a {age_and_sex} patient and a doctor done over chat.
The reason for the visit is: "{rfe}".
----
Medical Encounter #2
----
Patient: {rfe}
{dialogue}
----
Below are the medical entities and their status extracted from the patient-doctor dialogue from medical encounter #2. These entities can be used to help summarize the conversation below, but must be placed in the correct section (Demographics and Social Determinants of Health, Patient Intent, Pertinent Positives, Pertinent Unknowns, Pertinent Negatives, Medical History).

Positive Entities: {positive_entities}

Negative Entities: {negative_entities}

Unsure Entities: {unknown_entities}
----
Summary Instructions
----
Provide a summary of the medical encounter #2 between the doctor and the {age_and_sex} patient in 6 sections (Demographics and Social Determinants of Health, Patient Intent, Pertinent Positives, Pertinent Unknowns, Pertinent Negatives, Medical History).

Use the extracted entities to help summarize and place them in the appropriate section. Medical entities can be appropriate for any of the 6 sections and should be presented in an organized fashion.

Add any important details from the dialogue to further explain, elaborate, or qualify a medical entity. If a medical entity is clinically inaccurate or completely irrelevant to the summary of the encounter, then do not summarize it.

The 6 sections to write the summary with are Demographics and Social Determinants of Health, Patient Intent, Pertinent Positives, Pertinent Unknowns, Pertinent Negatives, and Medical History. The definitions of each section are listed below.

Demographics and Social Determinants of Health:
//Definition of section

Patient Intent:
//Definition of section

Pertinent Positives:
//Definition of section

Pertinent Unknowns:
//Definition of section

Pertinent Negatives:
//Definition of section

Medical History:
//Definition of section

----
Summary of Medical Encounter #2
----
```

Prompt 4: Prompt for MEDSUM summarization.

```
Below is a medical encounter between a {age_and_sex} patient and a doctor
     done over chat.
The reason for the visit is: "{rfe}".
----
Medical Encounter
----
Patient: {rfe}
{dialogue}
----
Summary Instructions
----
Provide a summary of the medical encounter between the doctor and the {
    age_and_sex} patient in 6 sections (Demographics and Social
    Determinants of Health, Patient Intent, Pertinent Positives, Pertinent
     Unknowns, Pertinent Negatives, Medical History).

Use the extracted entities to help summarize and place them in the
    appropriate section. Medical entities can be appropriate for any of
    the 6 sections and should be presented in an organized fashion.

Add any important details from the dialogue to further explain, elaborate
    , or qualify a medical entity. If a medical entity is clinically
    inaccurate or completely irrelevant to the summary of the encounter,
    then do not summarize it.

The 6 sections to write the summary with are Demographics and Social
    Determinants of Health, Patient Intent, Pertinent Positives, Pertinent
     Unknowns, Pertinent Negatives, and Medical History. The definitions
    of each section are listed below.

Demographics and Social Determinants of Health:
//Definition of section

Patient Intent:
//Definition of section

Pertinent Positives:
//Definition of section

Pertinent Unknowns:
//Definition of section

Pertinent Negatives:
//Definition of section

Medical History:
//Definition of section

----
Summary of Medical Encounter
----
```

Prompt 5: Prompt for naive zero-shot single-prompt summarization.

```
Given the following snippet of a medical dialogue summary, extract the medical concepts (symptoms, diseases, conditions, allergies, lab tests, etc.) present.

The heading of the section from which the summary was extracted will also be provided.

---Example 1---
Pertinent Negatives:  Patient reports no <concept_1>, no <concept_2>, <concept_3>, and <concept_4>. Patient also reports having no trouble with <concept_5>.

Medical Concepts: [<concept_1>, <concept_2>, <concept_3>, <concept_4>, <concept_5>]
---Example 1---

---Example 2---
Pertinent Positives:  Patient ongoing <concept_1> for the past 5 days, <concept_2>, and some <concept_3>. Patient had <concept_4> done in May 2021.

Medical Concepts: [<concept_1>, <concept_2>, <concept_3>, <concept_4>]
---Example 2---

---Example 3---
Pertinent Unknowns:  Patient is unsure about <concept_1> and <concept_2>.

Medical Concepts: [<concept_1>, <concept_2>]
---Example 3---

---Example 4---
Medical History: Patient reports some <concept_1> in the past, and had last <concept_2> on DATE_1.

Medical Concepts: [<concept_1>, <concept_2>]
---Example 4---

Here is the example to extract medical concepts from:

{section_heading}: {section_value}

Medical Concepts:
```

Prompt 6: Prompt for extracting medical concepts in metric computation.

```
Given a snippet (snippet) from a medical dialogue summary and a
    corresponding list (list_a) of medical concepts extracted from that
    snippet, evaluate what medical concepts from a separate list (list_b)
    can be found in either list_a or snippet.

Note that on some occasions a medical concept from list_b may not be
    found in list_a, but can be appropriate to be present given the
    snippet. This could include rephrasings of medical concepts that are
    clinically equivalent (Ex: COVID and COVID-19).

---Example---
snippet: <snippet>
list_a: [<concept_1>, <concept_2>, <concept_3>, <concept_4>, <concept_5>,
     <concept_7>]
list_b: [<concept_0>, <concept_1>, <concept_3>, <concept_4>, <concept_5>,
     <concept_6>]

found_b: [<concept_1>, <concept_3>, <concept_4>, <concept_5>]
not_found_b: [<concept_0>, <concept_6>]

---Example---

Here is the snippet, list_a. Evaluate the medical concepts in list_b as
    above.

snippet: {snippet}
list_a: {list_a}
list_b: {list_b}

found_b:
```

Prompt 7: Prompt for verifying concepts in metric computation.

Factors Affecting the Performance of Automated Speaker Verification in Alzheimer's Disease Clinical Trials

Malikeh Ehghaghi, Marija Stanojevic, Ali Akram, Jekaterina Novikova

Winterlight Labs Inc., Toronto, Canada
{malikeh, marija, aliakram, jekaterina}@winterlightlabs.com

Abstract

Detecting duplicate patient participation in clinical trials is a major challenge because repeated patients can undermine the credibility and accuracy of the trial's findings and result in significant health and financial risks. Developing accurate automated speaker verification (ASV) models is crucial to verify the identity of enrolled individuals and remove duplicates, but the size and quality of data influence ASV performance. However, there has been limited investigation into the factors that can affect ASV capabilities in clinical environments. In this paper, we bridge the gap by conducting analysis of how participant demographic characteristics, audio quality criteria, and severity level of Alzheimer's disease (AD) impact the performance of ASV utilizing a dataset of speech recordings from 659 participants with varying levels of AD, obtained through multiple speech tasks. Our results indicate that ASV performance: 1) is slightly better on male speakers than on female speakers; 2) degrades for individuals who are above 70 years old; 3) is comparatively better for non-native English speakers than for native English speakers; 4) is negatively affected by clinician interference, noisy background, and unclear participant speech; 5) tends to decrease with an increase in the severity level of AD. Our study finds that voice biometrics raise fairness concerns as certain subgroups exhibit different ASV performances owing to their inherent voice characteristics. Moreover, the performance of ASV is influenced by the quality of speech recordings, which underscores the importance of improving the data collection settings in clinical trials.

1 Introduction

Healthcare systems are increasingly relying on automatic speaker verification (ASV) models to ensure secure and accurate identification of patients and healthcare providers, with the aim of preventing fraud, safeguarding patient privacy, and ensuring the accuracy of medical records (Upadhyay et al., 2022; Arasteh et al., 2022).

Conducting large-scale clinical trials, involving numerous patients, doctors, clinics, and even different countries can pose significant challenges in identifying instances of duplicate participation, which occurs when a single individual joins the same study more than once, either at different sites or time points, leading to skewed results and undermining the validity of study findings (Irum and Salman, 2019). Shiovitz et al. (2013) discovered that as much as 7.78% of patients involved in a clinical study were duplicated across different sites.

In some cases, individuals participate in multiple clinical trials concurrently in order to earn more money. When a trial enrolls an adequate number of substandard participants, it risks not meeting the primary endpoints and ultimately causing a multimillion-dollar study to fail. Pinho et al. (2021) examined the financial effect of duplicate participants on the pharmaceutical companies conducting a set of short-term study programs across psychiatric disorders including Schizophrenia, Major Depressive Disorder, and Bipolar Depression. Based on their results, enrolling ineligible subjects in the selected studies results in a loss of around $29,680,000 for the sponsor pharmaceutical company. In addition, duplicate participation results in higher placebo rates and compromised data integrity. These findings highlight the importance of addressing the duplicate participant problem and underscore the need for reliable and accurate ASV methods in healthcare systems to verify whether an unknown voice belongs to a known enrolled individual (Upadhyay et al., 2022; Arasteh et al., 2022).

Cognitive impairment has been linked to a decline in vocabulary richness, syntactic complexity, and speech fluency, according to previous research (Thomas et al., 2005; Roark et al., 2011; Guinn and Habash, 2012; Meilán et al., 2012). Therefore,

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 218–227
July 14, 2023 ©2023 Association for Computational Linguistics

it is important to investigate whether the abnormal speech patterns exhibited by individuals with cognitive impairment can affect ASV performance. Despite this concern, there is a paucity of research examining the relationship between cognitive impairment and ASV in the existing literature. This research gap motivated us to address this issue by examining the effect of Alzheimer's disease (AD) severity level on ASV performance.

Furthermore, external factors such as participants' demographic information (Si et al., 2021), recording environment, or data collection procedure (Woo et al., 2006; Wan, 2017) may also have an impact on ASV performance, but their impact is not well-studied in the healthcare industry. An extensive analysis of these external factors could provide valuable insights into the accuracy and reliability of ASV models and identify potential sources of bias due to differences in inherent voice characteristics among subgroups (Si et al., 2021).

The purpose of this study is to investigate the effectiveness of ASV models in identifying duplicate patient participation in large-scale clinical trials, and to explore the factors that influence ASV performance in such settings. To this end, we utilize a longitudinal clinical dataset of English speech recordings obtained through multiple speech tasks from 659 participants with varying levels of AD. We employ the TitaNet model, an end-to-end deep learning text-independent ASV model pre-trained on a large volume of speech recordings of English speakers. ASV models can be classified into two groups: text-dependent (TD) and text-independent (TI). TI ASV models allow for more flexibility in the enrollment and verification phases without constraints on the speech content. When pre-trained on extensive audio datasets, TI models demonstrate a comparable level of accuracy to TD models. We evaluate the performance of TitaNet on our dataset in a zero-shot setting, achieving a 3.1% equal error rate (EER). In addition, we analyze the impact of various external factors on ASV performance, including participant demographic characteristics (i.e., age, and gender), audio quality criteria (i.e., clinician interference, background noise, participant accent, and participant clarity), as well as AD severity level. This study aims to provide valuable insights into the factors that can affect the performance of ASV models in clinical trial environments, with the goal of improving the accuracy, fairness, and reliability.

Our findings indicate that voice biometrics may present fairness issues, as certain subgroups demonstrate differing speaker verification performances due to their inherent voice characteristics. In addition, the quality of speech recordings can impact ASV performance, highlighting the importance of monitoring and enhancing data collection and recording settings during clinical trials.

2 Related Work

Speaker verification technology has been increasingly utilized in various domains, including healthcare systems. Several studies have analyzed the feasibility and effectiveness of speaker verification models in healthcare settings (Hao and Hei; Weng et al., 2014). However, the external factors that can affect the performance of ASV models has not been extensively studied through research in the healthcare field.

Race and Gender Effect: Si et al. (2021) utilized three state-of-the-art ASV models including the Xvector-TDNN (Snyder et al., 2018), ECAPA-TDNN (Desplanques et al., 2020), and DTW (Dutta, 2008) models to explore demographic effects on speaker verification. For this purpose, they used a subset of the mPower study (Bot et al., 2016), a Parkinson's disease mobile dataset, comparing a diverse group of 300 speakers by race and gender. Their results demonstrated that the Latinx subgroup indicates the worst ASV performance among the four major races in the dataset (i.e., White, Black, Latinx, and Asian). Based on their findings, gender represents minor differences in ASV performance between male-only and female-only subgroups. We did similar gender-level and accent-level analyses on patients with Alzheimer's disease to detect the potential sources of bias in ASV due to the inherent voice characteristics of distinct genders or English accents.

Age Effect: Kelly and Harte (2011) analyzed the effect of long-term ageing on ASV performance. They utilized a conventional GMM-UBM system (Irum and Salman, 2019) on a longitudinal voice dataset of a cohort of 13 adult speakers, whose recordings were collected over a time span of 30-40 years. According to their results, short-term aging (less than 5 years) does not have a significant impact on verification performance, compared to normal inter-session variations. However, for longer periods, aging has a negative effect on verification accuracy. Moreover, the researchers found that the rate of verification score decline is more rapid for speakers aged 60 years and above. However, they only evaluated their models on small cohorts and inter-speaker differences across different age groups were not further analyzed, while in the present work, we evaluate the ASV performance across different age groups over 55 years old and incorporate a larger clinical dataset with 659 speakers in total, while we controlled for AD effect. Taylor et al. (2020) also demonstrated that some speech and vocal characteristics (e.g., the spectral center of gravity, spectral skewness, or spectral kurtosis) undergo alterations with aging, and these changes can vary between men and women. These findings suggest that age is an effective factor in speaker's voice characteristics and this underscores the importance of assessing age effect on our ASV model to ensure the fairness of the model across different age groups.

Noise Effect: Wan (2017) applied LibriSpeech corpus (Panayotov et al., 2015) of English novel reading speech with varying lengths and tested ASV performance across different types and levels of background noise (e.g., babble, car, office and airplane noise) with a great mismatch between training and testing speech. Based on their findings, performance varies across different types of noise and the number of errors grow with a decrease in the sound-to-noise-ratio value. However, other metrics of audio quality were not considered in their study and their models were only trained and tested on healthy speech recordings. In the present study, we assess the effect of other audio quality aspects, such as participant clarity, clinician interference, and background noise, on ASV performance using a dataset of speakers with varying severity levels of AD, which was collected in clinical environment.

Speech Pathology Effect: Arasteh et al. (2022) have investigated the vulnerability of pathological speech to re-identification in ASV systems. In a large-scale study, they explored the effects of different speech pathologies on ASV using a real-world pathological speech corpus of more than 2,000 test subjects with various speech and voice disorders. Their results indicated that some types of speech pathology, particularly dysphonia, regardless of speech intelligibility, are more vulnerable to a breach of privacy compared to healthy speech. They do not analyze the effect of AD on ASV performance, even though speech and language impairment are prevalent issues in moderate to severe stages of AD that may potentially affect the ASV performance. This motivates us to evaluate ASV performance across varying severity levels of AD.

3 Methods

3.1 Datasets

The Alzheimer's Disease Clinical Trial (ADCT) dataset comprises speech recordings of English-speaking patients with a clinical diagnosis of mild to moderate AD who participated in a clinical trial. This is a proprietary dataset, which was collected every 12 weeks for a 48-week treatment period. It includes recordings of participants performing a set of self-administered speech tasks, including picture description (Goodglass et al., 2001; Becker et al., 1994), phonemic verbal fluency (Borkowski et al., 1967), and semantic verbal fluency (Tombaugh et al., 1999) tasks.

3.1.1 Demographic Information

Demographic data were collected about the participants at the beginning of the study. This data includes the age, and gender of the individuals upon consenting. The data collection study was approved by the ethical committee.

3.1.2 Transcription and Quality Assessment

All the audio recordings were manually transcribed by 49 trained transcriptionists based on the CHAT protocol and annotations (MacWhinney, 2014). The transcriptionists utilized an online tool that granted them access to the recordings and enabled them to transcribe the audio content, segment the files into utterances, and perform quality assessment. In addition, the transcriptions manually rated the quality of the recordings according to different quality criteria. The values range from 0 to 3 for each quality criterion. Value higher than 0 indicates that the audio recording has minor to major issues under that quality criterion. The quality criteria consist of background noise, clinician interference, participant accent, and participant clarity. The background noise criterion indicates whether there is noise in the background from the environment. Clinician interference indicates whether the clinician (or another speaker) interferes with the speech task. The participant accent criterion indicates whether the participant is a native or near-native speaker (values of 0) or has a detectable non-native accent (value higher than 0). Participant clarity indicates whether the participant's voice is hard to hear or understand.

3.1.3 Clinical Assessment

Participants were assessed on the severity level of AD using the Mini-Mental State Examination (MMSE) (Henneges et al., 2016) rating scale. MMSE is a brief cognitive function assessment, which consists of 30 questions that can be completed in less than 10 minutes. The questions are divided into seven categories, with each subscore examining a particular aspect of cognition: Orientation in time (score range 0-5), orientation in place (score range 0-5), registration (score range 0-3), attention and concentration (score range 0-5), recall (score range 0-3), language (score range 0-8), and drawing (score range 0-1). MMSE total score ranges from 0 to 30, with higher scores indicating better cognitive function and lower scores indicating more severe cognitive impairment. In this study, the participants were categorized into four levels of AD severity based on MMSE criteria (Wimo et al., 2013): Healthy Control (HC) (MMSE score > 26 points), Mild AD (MMSE score 21-26 points), Moderate AD (MMSE score 15-20 points), and Severe AD (MMSE score < 15 points).

3.1.4 Dataset Composition

The ADCT data comprises 7084 audio recordings from 659 speakers with 10.7±7.0 samples on average per each speaker. The average duration of total audio and speech-only audio are equal to 69.31 and 37.30 seconds, respectively.

In the dataset, 43.4% of the speakers are male and 56.6% of the speakers are female. The age range of the subjects spans from 55 to 80 years old. Age distribution of the subjects is represented in Figure 1, with an average value equal to 69.7±6.7.

Figure 2 indicates the distribution of MMSE scores in the ADCT dataset, showing that the majority of the samples consist of mild to severe levels of AD with scores in the range of 15 to 26 points. The average MMSE score is equal to 17.3±4.4. It should also be noted that the severity level of AD may vary over time for some of the speakers.

3.2 Models

In this study, we utilized the TitaNet model (Koluguri et al., 2022), which is a state-of-the-art end-to-end TI ASV model from the Nvidia NeMo toolkit[1] that had been pre-trained on an

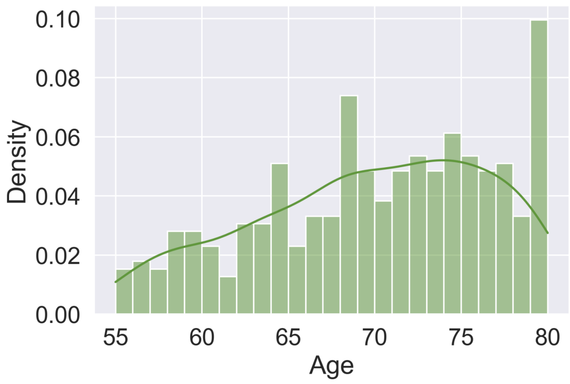


Figure 1: Age distribution of the speakers in ADCT dataset.

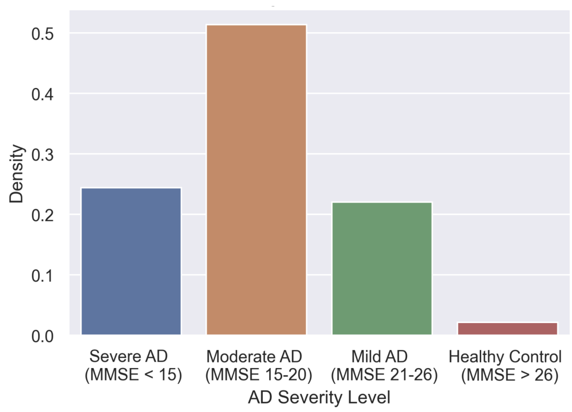


Figure 2: Distribution of AD severity levels in ADCT dataset.

extensive collection of English speech data, from various publicly-available resources. The TitaNet model is a neural network model that adopts an encoder-decoder architecture to extract speaker embeddings from speech. The model architecture is inspired by ContextNet (Han et al., 2020) model, which comprises 1D depth-wise separable convolutions followed by squeeze and excitation (SE) layers combined with channel attention pooling to convert utterances of varying lengths into a fixed-length embedding. The model contains 25.3M parameters and it is pre-trained on the VoxCeleb1 Dev (Nagrani et al., 2017), VoxCeleb2 Dev (Chung et al., 2018), Fisher (Cieri et al., 2004), Switchboard-Cellular1, Switchboard-Cellular2 (Godfrey and Holliman, 1993), and LibriSpeech (Panayotov et al., 2015) datasets.

We applied the model to the ADCT dataset in a zero-shot setting and used it for further analysis of the effect of external factors.

[1]https://github.com/NVIDIA/NeMo

3.3 Evaluation

To analyze the effect of external factors on ASV performance, we separately evaluated the performance of the TitaNet model across different subsets of ADCT data including genders, age groups, audio quality levels, and AD severity levels. Initially, we produced embeddings for all audio files within each group. Subsequently, we aggregated these embeddings to create a set of tuples comprising positive and negative pairs.

Positive tuples refer to pairs of embeddings that belong to the same speaker. The total number of positive tuples in each group is given by $\sum_{i=1}^{m} \binom{n_i}{2}$, which is calculated by summing up all pairs of n_i speech recordings for the ith speaker, where m is the total number of speakers in the same group.

Negative tuples refer to pairs of embeddings from different speakers within the same group. The total number of negative pairs is calculated as $\sum_{i=1}^{m} n_i * (N - n_i)/2$, where N is the total number of audio files in the group, and n_i is the number of speech recordings for speaker i. The sum is divided by 2 to avoid counting each pair twice.

After generating all the positive and negative pair tuples, we proceeded to compute the cosine similarity between the pairs of vector embeddings within each tuple. Subsequently, we adjusted a threshold value θ for each group through manual tuning until the true positive rate equalled the true negative rate, which enabled us to evaluate the performance of the ASV model using the equal error rate (EER) metric. If the cosine similarity value exceeded the threshold, we considered the corresponding tuple as belonging to the same speaker. Conversely, if the cosine similarity value was below the threshold, we deemed the two embeddings to represent different speakers.

4 Results and Discussion

In order to have a baseline level of ASV performance, we evaluated the TitaNet model on all the speech recordings of the ADCT dataset and obtained a 3.1% EER. To further analyze the impact of participant demographic characteristics, audio quality and AD severity level, we then recalculated EER and compared the ASV performance across different subgroups.

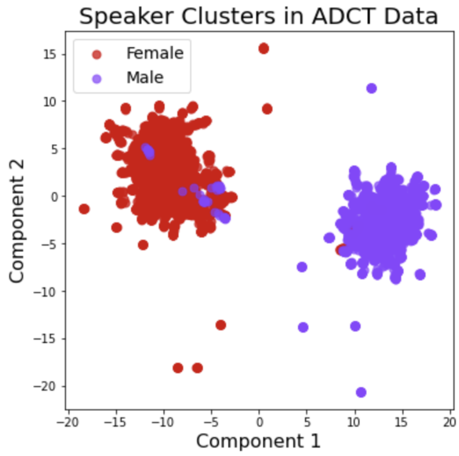


Figure 3: Visualization of speaker clusters using TitaNet embeddings of all audio recordings in ADCT dataset, created using the PaCMAP (Wang et al., 2021) dimensionality reduction method, where each color represents the gender of the speaker.

4.1 Effect of Participant Demographic Characteristics on ASV Performance

4.1.1 Gender Effect

We first analyzed the effect of speaker genders on the performance of ASV. As shown in Figure 3, two visually distinguishable speaker clusters appeared in the visualization of speech embeddings of all speakers in our dataset. Colouring the data points based on their gender indicates that each cluster is representative of a specific gender. The left cluster mostly comprises female speakers and the majority of the right cluster consists of male speakers. For further analysis of the ASV performance, we separately evaluated the model performance on male and female speech samples within each cluster. To control for the confounding factors, we randomly downsized the size of the female subgroup to the number of speakers in the male subgroup and also, matched the average age and MMSE score between the two subgroups. The results (Table 1) show that according to the EER metric, the ASV model performs better on the total dataset comprising diverse genders compared to when it is applied to the male-only or female-only speakers (Table 1). The ASV performance for the male subgroup (EER = 4.98%) is slightly better than that for the female subgroup (5.13% EER), although the difference is not substantial. This is in line with prior literature (Hanilçi and Ertaş, 2013) demonstrating that male speakers exhibit higher speaker recognition accuracy compared to female speakers regardless of the database

Gender	Tuned Thr.	EER(%)	#Spkrs	#Smpls	Avg #Smpls per Spkr	Avg Age	Avg MMSE Score
Female	0.75	5.13	170	2735	16.09±3.94	69.53±6.72	17.33±4.37
Male	0.75	4.98	170	2671	15.72±4.02	69.41±6.96	17.45±4.45
All	0.74	**3.10**	659	7084	10.70±7.00	69.55±6.75	17.32±4.44

Table 1: Benchmarking TitaNet ASV model across different genders on the ADCT English dataset. '#Spkrs' denotes the number of speakers per each gender. '#Smpls' denotes the number of audio recordings per each gender. 'Tuned Thr.' denotes tuned threshold.

Age Group	Tuned Thr.	EER(%)	#Spkrs	#Smpls	Avg #Smpls per Spkr	Gender	Avg MMSE Score
Age <= 70	0.73	3.62	197	3235	16.42±3.86	Male+Female	17.09±4.72
Age > 70	0.74	4.20	195	3022	15.50±4.07	Male+Female	17.57±4.11
All	0.74	**3.10**	659	7084	10.7±7.0	Male+Female	17.32±4.44

Table 2: Benchmarking TitaNet ASV model across different age groups on the ADCT English dataset. '#Spkrs' denotes the number of speakers per each age group. '#Smpls' denotes the number of audio recordings per each age group. 'Tuned Thr.' denotes tuned threshold.

and classifier used. The results also align with Si et al. (2021) indicating that there is little difference in gender in terms of the performance of ASV models in general.

4.1.2 Age Effect

We then evaluated the age effect on ASV performance. For this purpose, we categorized the speakers into two age subgroups with Age <= 70 and Age > 70 according to the speaker ages at the study enrollment date. The threshold was set to 70 because it is equal to the approximate median and mean value of the age distribution of the speakers in the ADCT dataset (Section 3.1.4). To control for the confounding factors, we designed our experiments to ensure that the age-based subgroups had a comparable number of speakers and average MMSE scores and included a combination of male and female speakers.

Based on the results indicated in Table 2, EER for participants under 70 is 0.58% lower than the older age group. These results can be explained by Taylor et al. (2020) revealing that specific attributes of speech and voice characteristics (e.g., fricative spectral moments, semitone standard deviation, etc.) vary according to age.

4.2 Effect of Audio Quality on ASV Performance

We examined how different audio quality factors affect the performance of speaker verification in clinical environments. We divided ADCT samples into two subgroups based on their quality rating for each criterion: 'No Issue' for samples with a rating of 0 and 'Minor to Major Issue' for those with a rating higher than 0 (maximum is 3). In order to mitigate the influence of possible confounding factors, we structured our experiments in such a way that the quality-based subgroups had a similar number of speakers and average MMSE score, and included both male and female speakers. Table 3 shows the comparison of EER values between each pair of subgroups per audio quality criterion. Our results indicate that subgroups of audio samples with no background noise, and high participant clarity yielded lower EER than subgroups with varying levels of background noise and poor participant clarity. Therefore, better control of the data collection setting, along with the use of high-clarity audio recordings with minimal background noise, would be recommended in order to improve ASV performance in clinical trials. Our findings are in line with Eskimez et al. (2018), who demonstrated that incorporating a DNN-based speech enhancement technique as a front-end noise reduction module can enhance the ASV performance when applied to noisy speech data obtained from real customers.

Our results also suggest that clinician interference can negatively impact ASV performance with 0.48% increase in EER. Therefore, it is recommended that clinicians refrain from interrupting participants during speech tasks in recording sessions to prevent any decline in performance.

We also evaluated the speakers' accents as a qual-

Audio Quality Criterion	Tuned Thr.	EER(%)	#Spkrs	#Smpls	Avg #Smpls per Spkr	Gender	Avg Age	Avg MMSE Score
Background Noise - No Issue	0.75	**2.90**	125	426	3.40±1.45	M + F	69.60±6.72	16.94±5.83
Background Noise - Minor to Major Issue	0.74	3.54	125	511	4.08±2.08	M + F	69.21±6.45	16.78±5.58
Participant Clarity - No Issue	0.75	**2.85**	112	481	4.29±1.70	M + F	69.80±6.38	16.81±5.62
Participant Clarity - Minor to Major Issue	0.74	3.41	112	432	3.85±1.83	M + F	69.23±6.81	16.04±5.54
Clinician Interference - No Issue	0.75	**2.90**	103	659	4.30±2.08	M + F	69.40±6.77	17.65±5.61
Clinician Interf. - Minor to Major Issue	0.73	3.38	103	399	3.87±1.86	M + F	69.43±6.84	14.77±5.22
Participant Accent - Native	0.74	2.97	188	901	4.79±2.82	M + F	68.63±6.89	17.22±5.01
Participant Accent - Non-Native	0.74	**2.01**	188	594	3.16±1.54	M + F	70.45±6.32	17.19±4.56
All	0.74	3.10	659	7084	10.7±7.0	M + F	69.55±6.75	17.32±4.44

Table 3: Benchmarking TitaNet speaker verification model across different levels of audio quality on the ADCT dataset. '#Spkrs' denotes the number of speakers per each quality subgroup. '#Smpls' denotes the number of audio recordings per each quality subgroup. 'Tuned Thr.' denotes tuned threshold. For each quality criterion, 'No Issue' indicates samples with a rating = 0 and 'Minor to Major Issue' indicates samples with a rating > 0 (maximum is 3). 'M' denotes male speakers and 'F' denotes female speakers. Bold font denotes the subgroup yielding best ASV performance in each quality criterion.

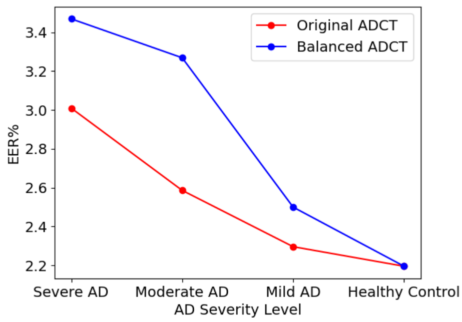
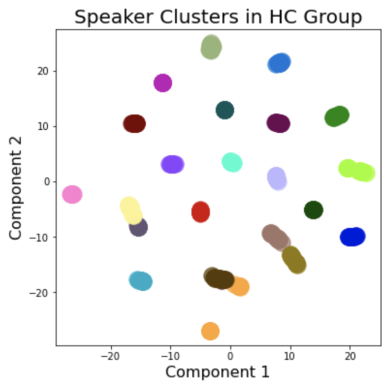
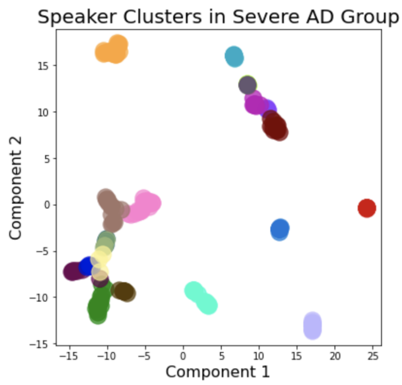


Figure 4: Comparison of the performance of the TitaNet ASV model across different AD severity levels based on EER% metric. Original ADCT refers to the dataset with the original number of speakers per severity level. Balanced ADCT refers to the dataset with each group downsized to the number of speakers in the smallest group, which is the HC group.



(a) Healthy Control



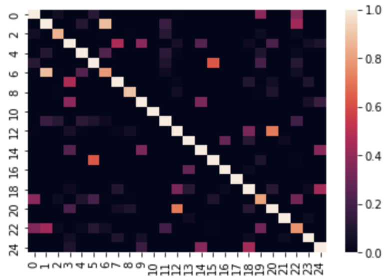
(b) Severe AD

Figure 5: Speaker cluster visualizations of HC and Severe AD groups based on TitaNet embeddings for ADCT dataset, created using the PaCMAP (Wang et al., 2021) dimensionality reduction method, with each colour representing a distinct speaker.

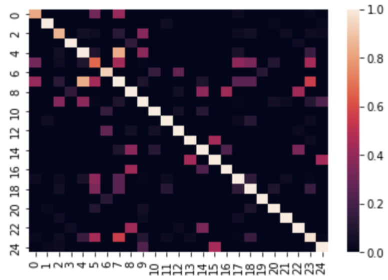
ity indicator, and our results show that ASV performs better on non-native speakers (2.01% EER) than on native speakers (2.97% EER). This suggests that ASV performance can be better in trials that involve participants who speak with diverse non-native English accents as a way to identify unique speech characteristics for each individual.

4.3 Effect of the Severity Level of Alzheimer's Disease on ASV Performance

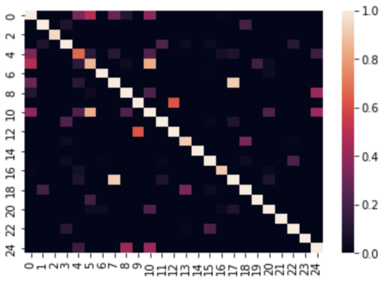
We examined how the AD severity level impacts the performance of the ASV model. For this purpose, we performed a separate recalculation of EER for subgroups of audio samples consisting of Severe AD (Number of speakers = 218), Moderate AD (Number of speakers = 436), Mild AD (Number of speakers = 244), and HC (Number of speakers = 25), while retaining the original number of speakers. To establish a fair comparison, we then balanced the number of speakers across the subgroups by downsizing each to the smallest subgroup size of 25 speakers, which was the size of the HC subgroup. In both scenarios, higher AD severity levels lowered speaker verification performance, by about 1% to 1.5% of EER (Figure 4). Also, EER was lower within the groups where a higher num-



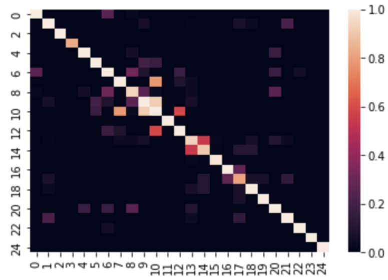
(a) Severe AD



(b) Moderate AD



(c) Mild AD



(d) Healthy Control

Figure 6: Similarity heatmap visualizations for ADCT dataset using TitaNet embeddings for different AD severity levels each downsized to the number of speakers in the smallest group, which is equal to 25. The lighter colours correspond to a higher level of similarity between the associated speaker tuples.

ber of speakers were included. Figure 5a and 5b display the PaCMAP (Wang et al., 2021) visualization of TitaNet speaker embeddings for the HC and Severe AD groups in the balanced ADCT dataset, where each colour representing a unique speaker. In both the HC and Severe AD groups, samples from the same speakers are clustered close to each other, while in the HC group, the samples from different speakers are more distinguishable compared to the Severe AD group. Moreover, there is a higher level of similarity between negative speaker tuples within Moderate and Severe AD subgroups in comparison to Mild AD and HC subgroups, as indicated in the similarity heatmaps for different AD severity levels (Figure 6). Overall, our findings suggest that the unique voice characteristics associated with varying levels of AD severity (Boschi et al., 2017) can be entangled with the identity of the speaker and may introduce a potential source of bias in the ASV models.

5 Conclusion

Large-scale clinical trials require accurate verification of participants, as duplicate participation may lead to substandard data quality and significant financial and health risks. Therefore, developing accurate ASV models for verifying participant identity is essential in these settings. External factors such as participant profile or audio quality can cause errors and biases in ASV performance during the trials, but limited research has been conducted in this area. In the present work, we utilize a longitudinal speech dataset of participants with varying levels of AD severity and investigate the impact of external factors, such as different participant demographic characteristics, audio quality criteria, and AD severity levels, on the performance of an end-to-end TI ASV model. Our findings show that variations in ASV performance can be attributed to the inherent voice characteristics of different subgroups (e.g., different ages, genders, accents, or AD severity levels) that are likely to be confused with the identity of the speaker. Hence, it is critical to reassess this technology to mitigate the risk of potential biases toward certain subgroups. Based on our results, poor audio quality with unclear speech, noisy background, and clinician interference also negatively impacts ASV performance. This highlights the importance of quality assurance for the speech recordings during the trials. In future work, we aim to automate the audio quality assessment process by leveraging existing automated methods such as perceptual evaluation of speech quality (PESQ) (Rix et al., 2001) or short-time objective intelligibility (STOI) (Taal et al., 2010), which would reduce the human effort required for this task.

References

Soroosh Tayebi Arasteh, Tobias Weise, Maria Schuster, Elmar Noth, Andreas Maier, and Seung Hee Yang. 2022. The effect of speech pathology on automatic speaker verification-a large-scale study. arXiv preprint arXiv:2204.06450.

James T Becker, François Boiler, Oscar L Lopez, Judith Saxton, and Karen L McGonigle. 1994. The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis. Archives of neurology, 51(6):585–594.

John G Borkowski, Arthur L Benton, and Otfried Spreen. 1967. Word fluency and brain damage. Neuropsychologia, 5(2):135–140.

Veronica Boschi, Eleonora Catricala, Monica Consonni, Cristiano Chesi, Andrea Moro, and Stefano F Cappa. 2017. Connected speech in neurodegenerative language disorders: a review. Frontiers in psychology, 8:269.

Brian M Bot, Christine Suver, Elias Chaibub Neto, Michael Kellen, Arno Klein, Christopher Bare, Megan Doerr, Abhishek Pratap, John Wilbanks, E Dorsey, et al. 2016. The mpower study, parkinson disease mobile data collected using researchkit. Scientific data, 3(1):1–9.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. Voxceleb2: Deep speaker recognition. arXiv preprint arXiv:1806.05622.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The fisher corpus: A resource for the next generations of speech-to-text. In LREC, volume 4, pages 69–71.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. arXiv preprint arXiv:2005.07143.

Tridibesh Dutta. 2008. Dynamic time warping based approach to text-dependent speaker identification using spectrograms. In 2008 Congress on Image and Signal Processing, volume 2, pages 354–360. IEEE.

Sefik Emre Eskimez, Peter Soufleris, Zhiyao Duan, and Wendi Heinzelman. 2018. Front-end speech enhancement for commercial speaker verification systems. Speech Communication, 99:101–113.

John Godfrey and Edward Holliman. 1993. Switchboard-1 release 2 ldc97s62. Linguistic Data Consortium, page 34.

Harold Goodglass, Edith Kaplan, and Sandra Weintraub. 2001. BDAE: The Boston Diagnostic Aphasia Examination. Lippincott Williams & Wilkins Philadelphia, PA.

Curry I Guinn and Anthony Habash. 2012. Language analysis of speakers with dementia of the alzheimer's type. In 2012 AAAI Fall Symposium Series.

Wei Han, Zhengdong Zhang, Yu Zhang, Jiahui Yu, Chung-Cheng Chiu, James Qin, Anmol Gulati, Ruoming Pang, and Yonghui Wu. 2020. Contextnet: Improving convolutional neural networks for automatic speech recognition with global context. arXiv preprint arXiv:2005.03191.

Cemal Hanilçi and Figen Ertaş. 2013. Investigation of the effect of data duration and speaker gender on text-independent speaker recognition. Computers & Electrical Engineering, 39(2):441–452.

Bin Hao and Xiali Hei. Voice print liveness detection for medical devices.

Carsten Henneges, Catherine Reed, Yun-Fei Chen, Grazia Dell'Agnello, and Jeremie Lebrec. 2016. Describing the sequence of cognitive decline in alzheimer's disease patients: results from an observational study. Journal of Alzheimer's Disease, 52(3):1065–1080.

Amna Irum and Ahmad Salman. 2019. Speaker verification using deep neural networks: A. International Journal of Machine Learning and Computing, 9(1).

Finnian Kelly and Naomi Harte. 2011. Effects of long-term ageing on speaker verification. In Biometrics and ID Management: COST 2101 European Workshop, BioID 2011, Brandenburg (Havel), Germany, March 8-10, 2011. Proceedings 3, pages 113–124. Springer.

Nithin Rao Koluguri, Taejin Park, and Boris Ginsburg. 2022. Titanet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 8102–8106. IEEE.

Brian MacWhinney. 2014. The CHILDES project: Tools for analyzing talk, Volume II: The database. Psychology Press.

Juan JG Meilán, Francisco Martínez-Sánchez, Juan Carro, José A Sánchez, and Enrique Pérez. 2012. Acoustic markers associated with impairment in language processing in alzheimer's disease. The Spanish journal of psychology, 15(2):487–494.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: a large-scale speaker identification dataset. arXiv preprint arXiv:1706.08612.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), pages 5206–5210. IEEE.

Andre Pinho, Kerri Weingard, and Mitchell D Efros. 2021. Improving safety and preventing failure in clinical trials by detecting and preventing duplicate and professional research subjects: the case for use of a research subject database registry. Journal of Clinical Trials, 11(4):1–6.

Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. 2001. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In 2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221), volume 2, pages 749–752. IEEE.

Brian Roark, Margaret Mitchell, John-Paul Hosom, Kristy Hollingshead, and Jeffrey Kaye. 2011. Spoken language derived measures for detecting mild cognitive impairment. IEEE transactions on audio, speech, and language processing, 19(7):2081–2090.

Thomas M Shiovitz, Charles S Wilcox, Lilit Gevorgyan, and Adnan Shawkat. 2013. Cns sites cooperate to detect duplicate subjects with a clinical trial subject registry. Innovations in Clinical Neuroscience, 10(2):17.

Sophie Si, Zhengxiong Li, and Wenyao Xu. 2021. Exploring demographic effects on speaker verification. In 2021 IEEE Conference on Communications and Network Security (CNS), pages 1–2. IEEE.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), pages 5329–5333. IEEE.

Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. 2010. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In 2010 IEEE international conference on acoustics, speech and signal processing, pages 4214–4217. IEEE.

Sammi Taylor, Christopher Dromey, Shawn L Nissen, Kristine Tanner, Dennis Eggett, and Kim Corbin-Lewis. 2020. Age-related changes in speech and voice: spectral and cepstral measures. Journal of Speech, Language, and Hearing Research, 63(3):647–660.

Calvin Thomas, Vlado Keselj, Nick Cercone, Kenneth Rockwood, and Elissa Asp. 2005. Automatic detection and rating of dementia of alzheimer type through lexical analysis of spontaneous speech. In IEEE International conference mechatronics and automation, 2005, volume 3, pages 1569–1574. IEEE.

Tom N Tombaugh, Jean Kozak, and Laura Rees. 1999. Normative data stratified by age and education for two measures of verbal fluency: Fas and animal naming. Archives of clinical neuropsychology, 14(2):167–177.

Shrikant Upadhyay, Mohit Kumar, Ashwani Kumar, Ramesh Karnati, Gouse Baig Mahommad, Sara A Althubiti, Fayadh Alenezi, and Kemal Polat. 2022. Feature extraction approach for speaker verification to support healthcare system using blockchain security for data privacy. Computational and Mathematical Methods in Medicine, 2022.

Qianhui Wan. 2017. Speaker Verification Systems under Various Noise and SNR Conditions. Ph.D. thesis, Université d'Ottawa/University of Ottawa.

Yingfan Wang, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik. 2021. Understanding how dimension reduction tools work: An empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization. Journal of Machine Learning Research, 22(201):1–73.

Fuliang Weng, Taufiq Hasan, and Zhe Feng. 2014. Speaker verification in a health monitoring system. US Patent 8,818,810.

Anders Wimo, Catherine C Reed, Richard Dodel, Mark Belger, Roy W Jones, Michael Happich, Josep M Argimon, Giuseppe Bruno, Diego Novick, Bruno Vellas, et al. 2013. The geras study: a prospective observational study of costs and resource use in community dwellers with alzheimer's disease in three european countries–study design and baseline findings. Journal of Alzheimer's Disease, 36(2):385–399.

Ram H Woo, Alex Park, and Timothy J Hazen. 2006. The mit mobile device speaker verification corpus: data collection and preliminary experiments. In 2006 IEEE Odyssey-The Speaker and Language Recognition Workshop, pages 1–6. IEEE.

Team Cadence at MEDIQA-Chat 2023: Generating, augmenting and summarizing clinical dialogue with large language models

Ashwyn Sharma
Cadence Solutions, USA
ashwyn@cadencerpm.com

David I. Feldman, MD, MPH
Cadence Solutions, USA
Massachusetts General Hospital, Harvard University , USA
david.feldman@cadencerpm.com

Aneesh Jain
Cadence Solutions, USA
Virginia Polytechnic Institute and State University, USA
aneeshjain70@gmail.com

Abstract

This paper describes Team Cadence's winning submission to Task C of the MEDIQA-Chat 2023 shared tasks. We also present the set of methods, including a novel *N-pass* strategy to summarize a mix of clinical dialogue and an incomplete summarized note, used to complete Task A and Task B, ranking highly on the leaderboard amongst stable and reproducible code submissions. The shared tasks invited participants to summarize, classify and generate patient-doctor conversations. Considering the small volume of training data available, we took a *data-augmentation-first* approach to the three tasks by focusing on the dialogue generation task, i.e., Task C. It proved effective in improving our models' performance on Task A and Task B. We also found the BART architecture to be highly versatile, as it formed the base for all our submissions. Finally, based on the results shared by the organizers, we note that Team Cadence was the only team to submit stable and reproducible runs to all three tasks.

1 Introduction

MEDIQA-Chat 2023 Shared Tasks included three tasks on the summarization and generation of doctor-patient conversations to promote research on these topics (Ben Abacha et al., 2023). Task A (*Short Dialogue2Note Summarization*) expected a section summary (section header and content) given a short input conversation. We recognized that generating the summary content was an abstractive summarization (Chopra et al., 2016) task while predicting the section header was a multi-class (twenty normalized section labels) classification task. Task B (*Full Dialogue2Note Summarization*) was another abstractive summarization task that required submissions to generate a complete clinical note from a whole dialogue between a patient and a doctor. The complete clinical note was expected to have the following first-level section headers: *"HISTORY OF PRESENT ILLNESS", "PHYSICAL EXAM", "RESULTS", and "ASSESSMENT AND PLAN"*. Finally, Task C (*Note2Dialogue Generation*), a data augmentation (Shorten et al., 2021) task, was about generating patient-doctor conversations for complete input notes.

Aside from predicting section headers for Task A, all other tasks could be formulated as sequence-to-sequence (Sutskever et al., 2014) learning tasks. Various model architectures based on transformers (Vaswani et al., 2017) have proved to be successful at tackling such tasks. Therefore, leveraging pre-trained model checkpoints from public repositories was considered the right choice. Encouraged by the leaderboard for SAMSum (Gliwa et al., 2019) on HuggingFace (Wolf et al., 2020), a dialogue summarization dataset, we chose BART (Lewis et al., 2019) as the base model for our experiments. Specifically, we picked the *facebook/bart-*

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 228–235
July 14, 2023

large [1] model checkpoint (referenced as *bart-large* in this text from hereon) for its effectiveness on text-generation tasks.

The SAMSum (Gliwa et al., 2019) dataset is intended to train dialogue summarization models. However, we recognized that the input and target labels could be inverted to train a dialogue generation model. We trained/validated *bart-large* on the inverse of SAMSum (Gliwa et al., 2019) dataset followed by the Task C training dataset provided by the task organizers, achieving ROUGE-1 and ROUGE-2 scores of 59.11 and 23.69, respectively, on the validation set. This model was then used to augment datasets for Task A and Task B summarization tasks. In order to generate synthetic patient-doctor conversations, we chose to sample a thousand discharge summary notes from the MIMIC-IV-Note (Johnson et al., 2023; Goldberger et al., 2000) dataset. We then added these dialogue-note pairs to the Task A and Task B training datasets provided by the organizers. The impact of this augmentation technique is noted in Section 5 below.

For Task A summarization, *bart-large* was fine-tuned on the SAMSum (Gliwa et al., 2019) dataset followed by fine-tuning on the augmented dataset for Task A, which achieved ROUGE-1 and ROUGE-2 scores of 50.7 and 21.4, respectively, on the validation set. Our methods yielded an overall improvement (over the baseline) of 13.1% and 14% in ROUGE-1 and ROUGE-2 scores, respectively. Results from fine-tuning *bart-large* on the unaugmented (original) Task A dataset were considered the baseline in this comparison.

Inspired by the significant gains exhibited by the Task A model, we decided to use it as the base model for Task B. Fine-tuning this base model on the augmented Task B dataset yielded ROUGE-1 and ROUGE-2 scores of 54.16 and 26.04, respectively - a 13.7% gain in ROUGE-2 score over the baseline. Results from fine-tuning the base model on the unaugmented (original) Task B dataset were considered the baseline in this comparison. The final submission(run1) achieved ROUGE-1 and ROUGE-2 scores of 49.5 and 23.4 on the test set. Unfortunately, the Task B dataset comprised input conversations almost twice as long as the maximum number of tokens accepted by *bart-large*, which naturally prohibits the model's ability to summarize the entire conversation. To solve this problem, we developed an *N-pass* strategy in which the model attempts to summarize the conversation in multiple steps. Each step (or pass) involves the model taking as input the summary note of the dialogue processed till that step, concatenated with the rest of the dialogue. In other words, we trained the model to summarize a partial mix of an incomplete clinical note and an incomplete patient-doctor conversation. This strategy led to a gain of 6.6% and 8.1% in ROUGE-1 (57.76) and ROUGE-2 (28.15) scores, respectively, on the validation set. We submitted the *N-pass* model as run2, which outperformed the run1 submission by 6.8%, both for ROUGE-1 (52.9) and ROUGE-2 (25) scores, on the test set. It also improved the division-based aggregate score by 16.75%. Overall, our methods improved the baseline ROUGE-2 score by 22.9% on the validation set, while the baseline ROUGE-1 score was found to be slightly better by 0.45%.

Given the promising performance of *bart-large* on the summarization tasks, we also decided to use it for Task A classification. We leveraged the *BartForSequenceClassification* wrapper offered by HuggingFace (Wolf et al., 2020), a BART model with a sequence classification head on top (a linear layer on top of the pooled output). Using this approach, we achieved an accuracy of 78% and an F1 score of 78.37%. The final submission was reported to have an accuracy of 73.5% on the test set.

2 Background and Related Work

Studies like the ones from Alkureishi MA et al. (Alkureishi et al., 2016) and Rathert et al. (Rathert et al., 2017) have presented evidence on EHRs (Electronic Health Records) impacting the quality of patient-doctor conversations. Digital scribes (van Buchem et al., 2021) and summarization tools (Shanafelt et al., 2016) can mitigate some of these problems. However, many challenges are associated with clinical dialogue summarization (Zhu and Penn, 2006). Some significant challenges include omitting key medical concepts (Knoll et al., 2022) and hallucinating unsubstantiated information.

Several attempts have been made to address said inherent challenges and automatically generate high-quality summaries of clinical encounters. Approaches like the ones used by Enarvi et al. (2020) have utilized a transformer (Vaswani et al., 2017) model to summarize doctor-patient conversations. Joshi et al. (2020) and Michalopoulos et al.

[1] `https://huggingface.co/facebook/bart-large`

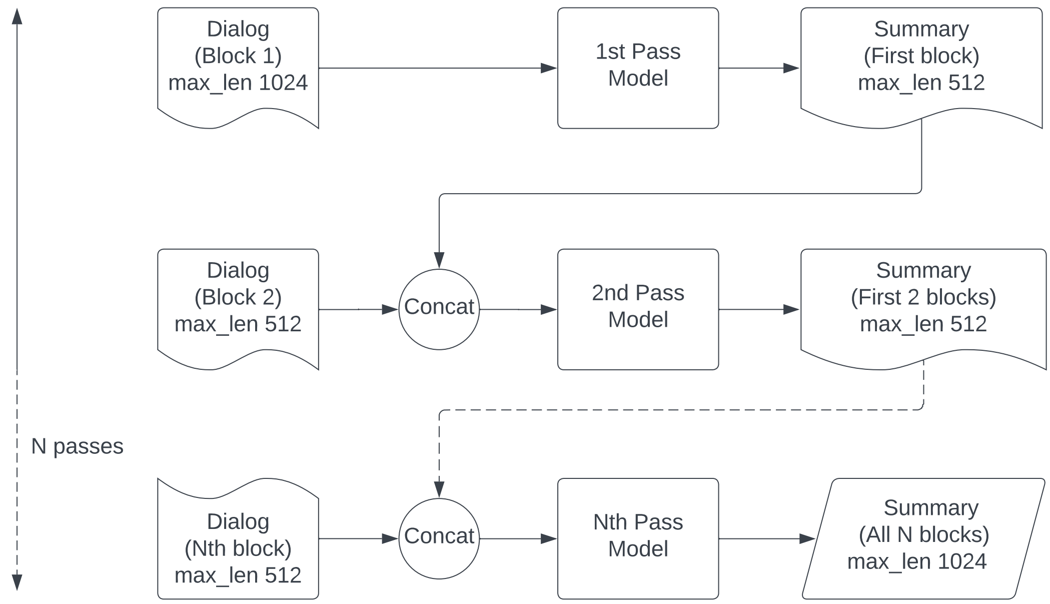


Figure 1: N-pass summarization for handling long conversations.

(2022) have also incorporated medical knowledge into these models. On the data generation front, Chintagunta et al. (2021) showed that large language models can be used for augmenting medical summarization datasets.

To the best of our knowledge, the *N-pass* strategy used to address long input sequences of Task B is novel. However, multiple multi-stage summarization approaches have been proposed so far. For example, Krishna et al. (2020) used modular summarization techniques to produce notes from patient-doctor conversations. Zhang et al. (2021) used multi-stage summarization for long inputs, whereas Gidiotis and Tsoumakas (2020) split a long document and its summary into multiple source-target pairs using sentence similarity. Recursive summarization incorporating human feedback (Wu et al., 2021) even achieved state-of-the-art results in book summarization.

3 Datasets

3.1 MEDIQA-Chat-2023

Task A training (validation) dataset (Ben Abacha et al., 2023) provided by the organizers consists of 1,201 (100) pairs of conversations and associated section headers and summaries. There were 20 unique normalized section headers overall. The Task B and Task C training (validation) set consists of 67 (20) pairs of conversations and full clinical notes (Yim et al., 2023).

3.2 SAMSum

The SAMSum dataset contains 16369 conversations and their summaries (Gliwa et al., 2019), with a train/val/test split of 14732/818/819. Several dialogue summarization models have leveraged this dataset (Ni et al., 2022) and achieved promising results on the task. We note the impact of this dataset in the ablation study (Section 5).

3.3 MIMIC-IV-Note

MIMIC-IV-Note contains 331,794 deidentified free-text clinical notes for patients included in the MIMIC-IV clinical database (Johnson et al., 2023; Goldberger et al., 2000). We sampled a thousand notes from this dataset and used the Task C (dialogue generation) model for downstream data augmentation of Task A and Task B. Ablation study (Section 5) highlights significant contributions of this dataset to improving the results.

4 Methods

4.1 Dialogue Generation

We discovered that by flipping input and target labels, the SAMSum (Gliwa et al., 2019) dataset could also train a dialogue generation model. Our

Table 1: Hyperparameters used for Task A, Task B and Task C

Parameter	**Task A**		**Task B**	**Task C**
	Classification	**Summarization**	**Summarization**	**Generation**
learning_rate	2E-05	5E-05	5E-05	5E-05
per_device_train_batch_size	8	4	4	4
per_device_eval_batch_size	8	4	2	2
weight_decay	0.01	0	0	0
num_train_epochs	30	30	30	10
fp16	TRUE	TRUE	TRUE	TRUE
gradient_accumulation_steps	4	8	8	8
gradient_checkpointing	TRUE	TRUE	TRUE	TRUE
predict_with_generate	-	TRUE	TRUE	TRUE
generation_max_length	-	512	1024	1024
max_target_length	-	512	1024	1024
max_source_length	1024	1024	1024	1024

recipe included fine-tuning *bart-large* on the inverted SAMSum (Gliwa et al., 2019) dataset for 10 epochs, followed by fine-tuning on a dataset that combined training and validation datasets from Task A and Task C for another 10 epochs. Fine-tuning was performed using the Trainer API offered by HuggingFace (Wolf et al., 2020), and the hyperparameters used are described in (Table 1). We did not perform a comprehensive sweep and recognize that a more optimal set of hyperparameters could yield better results. The model yielded by this recipe was also used for generating synthetic data for Task A and Task B summarization. Specifically, patient-doctor conversations were generated for 1000 discharge summary notes sampled from the MIMIC-IV-Note (Johnson et al., 2023; Goldberger et al., 2000) dataset. We used ROUGE-1 and ROUGE-2 scores for evaluating the model's performance on the validation set (Lin, 2004).

4.2 Dialogue Summarization

Summarization models for Task A and Task B leveraged *bart-large* fine-tuned on the SAMSum (Gliwa et al., 2019) dataset for 10 epochs as the base model. The base model was then fine-tuned on the augmented version of the Task A training dataset for 30 epochs. Like dialogue generation, fine-tuning was performed using the Trainer API offered by HuggingFace (Wolf et al., 2020), and the hyperparameters used are described in Table 1. We did not perform a comprehensive sweep and recognize that a more optimal set of hyperparameters could yield better results. With a working hypothesis that the Task A model can capture local themes in conversations with fewer turns, we used the model yielded by the above recipe as the base model for Task B.

Before augmenting the Task B dataset with the dialogue generation model, we sanitized the 1000 notes sampled from the MIMIC-IV-Note (Johnson et al., 2023; Goldberger et al., 2000) dataset. The sanitization process mainly included removing first-level section headers not accepted for evaluation by the organizers, as laid out in (Section 1). The base model was then fine-tuned on the *sanitized-and-augmented* dataset (named Augmented(Sections) in result tables) using the same process as Task A. This fine-tuned version was submitted as run1 and suffered from a significant drawback - the inability to handle input sequences longer than 1024 tokens. To address the shortcoming, we developed a novel *N-pass* approach by training a model that can generate summaries given a partial mix of incomplete summaries and incomplete dialogue. Specifically, a 2-pass version, named run2, was submitted to the shared task.

The *N-pass* approach is illustrated in Figure 1. The idea is to summarize long conversations in multiple passes, where each pass accepts as input the next *block* of the unsummarized dialogue concatenated with the summary output by the previous pass. The intuition behind this approach is to accommodate the limit on the number of input tokens accepted by the model by feeding it the dialogue in *blocks* but still propagating the context by incorporating the summary generated till that point. For run2, the model used for run1 was fine-tuned for

30 epochs on a dataset that concatenated the first *block* summary with the second *block* of the dialogue. The first *block* summaries were generated by the run1 model. A *block size* of 512 tokens was used for both the input and the output (except the final *pass* where output is 1024 tokens). We used a combination of ROUGE-1 and ROUGE-2 scores for evaluating the model's ability to summarize the conversations in the validation set (Lin, 2004).

4.3 Classification

We used a simple yet effective classification approach to producing section headers for Task A. Given the promising results from using *bart-large* on the summarization and dialogue generation tasks, we chose to stick with the same for classification. To be exact, we fine-tuned the model used for Task A submission on the classification task by leveraging the *BartForSequenceClassification* wrapper offered by HuggingFace (Wolf et al., 2020), a BART model with a sequence classification head on top (a linear layer on top of the pooled output). Again, the Trainer API was used with no hyperparameter sweep. Table 1 lists the hyperparameters used for fine-tuning the classifier.

5 Experiments and Ablation Study

Dataset	ROUGE-1	ROUGE-2
MEDIQA	47.5	19.8
Augmented (Sections)	48.12	19.9
Augmented	**50.7**	**21.4**

Table 2: Task A - results with different training datasets. Metrics evaluated on the task validation set.

Model	ROUGE-1	ROUGE-2
bart-large	44.8	18.77
bart-samsum	**47.5**	**19.8**

Table 3: Task A - impact of fine-tuning on SAMSum. Metrics evaluated on the task validation set.

5.1 Task A

In Table 2, we compare the results obtained on the Task A validation set by using three different training datasets - original Task A training data, augmented Task A training data, and *sanitized-and-augmented* (defined in Section 4.2) training data. The augmented version outperforms the original Task A training data by 6.7% (ROUGE-1) and 8% (ROUGE-2). As expected, the *sanitized-and-augmented* training data yields smaller gains because the summary notes for Task A are shorter and do not include first-level section headers in Task B training data.

An ablation study (Table 3) was also conducted on the impact of fine-tuning *bart-large* on the SAMSum(Gliwa et al., 2019) dataset. It was found that fine-tuning on the SAMSum (Gliwa et al., 2019) dataset improved performance on the validation set by 6% (ROUGE-1) and 5.4% (ROUGE-2).

Task A summarization model fine-tuned on classification achieved an accuracy of 78% and an f1 score of 78.37% on the validation set.

Version	ROUGE-1	ROUGE-2
MEDIQA	48.13	19.0
Augmented	51.86	23.42
Augmented (Sections)	54.16	26.04
2-pass	**57.76**	**28.15**

Table 4: Task B - results with different training datasets and the 2-pass strategy. Metrics evaluated on the task validation set.

Model	ROUGE-1	ROUGE-2
bart-large	**58.02**	**22.9**
bart-samsum	48.13	19.0

Table 5: Task B - impact of fine-tuning on SAMSum. Metrics evaluated on the task validation set.

5.2 Task B

Table 4 shows that the *2-pass* summarization strategy leads to a gain of 6.6% (ROUGE-1) and 8.1% (ROUGE-2). Furthermore, training on the *sanitized-and-augmented* dataset yields improvements of 12.5% (ROUGE-1) and 37% (ROUGE-2), driving home the value of data augmentation by clinical dialogue generation. Interestingly, simply fine-tuning on the SAMSum(Gliwa et al., 2019) dataset led to worse results (Table 5) on the Task B validation set, which could be explained by the discrepancy in the length of the conversations and the summaries between the two datasets.

Dataset	bart-large		bart-samsum	
	R-1	R-2	R-1	R-2
MEDIQA	53.6	17.26	56.55	20.64
Combined	58.43	22.74	**59.11**	**23.69**

Table 6: Task C - results with different training datasets and impact of fine-tuning on SAMSum. Metrics evaluated on the task validation set.

5.3 Task C

The ablation study (Table 6) for Task C highlights two significant ideas. First, adding the training data from Task A contributed a hike of 4.5% (9%) in the ROUGE-1 score and 14.7% (31.7%) in the ROUGE-2 score for the model (not) fine-tuned on the inverse of the SAMSum (Gliwa et al., 2019) dataset. Second, fine-tuning on the inverse of SAMSum (Gliwa et al., 2019) led to a gain of 5.5% (1.1%) in ROUGE-1 scores and 19.5% (4.1%) in ROUGE-2 scores when training data from Task C (Task A + Task C) was used. It shows that the additional data from Task A is more critical when fine-tuning on the inverse of SAMSum (Gliwa et al., 2019) is skipped.

6 Results

Team Cadence's submission for Task C earned ***rank-1*** amongst all participants, beating the next-best submission by 28.3% (ROUGE-1) and 99% (ROUGE-2).

The organizers shared test set results (Ben Abacha et al., 2023) along with a *code status* description where a *code status* of 1 meant that the organizers were able to run the submitted code and reproduce the results, and a *code status* of 2 meant that they were able to run the code and found minor differences with no changes in rankings. *Code statuses* 3,4, and 5 meant that the organizers found the submitted code to be unstable or not runnable under their configurations. Amongst *code statuses* 1 and 2, Team Cadence achieved the following ranks: *rank-2* on TaskB-summarization, *rank-3* on TaskA-summarization, *rank-3* on TaskB-summarization(note-divisions), and *rank-5* on TaskA-classification. The code for generating the submitted runs is being shared publicly[2].

[2]https://github.com/ashwyn/MEDIQA-Chat-2023-Cadence

7 System Specification

In the spirit of reproducibility, we share details of the systems used to run these experiments. The models were fine-tuned on *g4dn.12xlarge* AWS Sagemaker notebook instances [3]. HuggingFace's Python package transformers (Wolf et al., 2020) version 4.27.1 was used in a Python3.8 environment. Reported results were aggregated from 4 different runs using 4 different random seeds.

8 Limitations and Future Work

The methods described in this paper do not leverage any external medical knowledge, a technique that has been shown to be effective by other studies (Joshi et al., 2020; Michalopoulos et al., 2022). And like other methods based on large language models, in theory, our models are also prone to hallucinations and omission of key-clinical concepts. We plan to explore constrained beam search[4] as a mitigation strategy for addressing these challenges in the future.

Although the impact of the Task C model as a data augmentation tool is undoubtedly positive (Section 5), qualitative error analysis of patient-doctor conversations produced by the model showed that the output contained a small number of dialogue turns, and each individual turn was too long, packed with information. Producing conversations with a more natural flow should yield an even better boost on downstream tasks, and we leave exploring such methods to future experimentation. We also recognize that *N-pass* summarization for Task B with higher values of *N* should be able to cover the entirety of the input conversations in the Task B datasets, albeit with diminishing returns as *N* increases. We hope to evaluate them in future iterations of similar shared tasks.

9 Conclusion

The two key takeaways from the experiments and results in this paper are significant improvements in summarization results driven by data augmentation and the *N-pass* summarization technique for handling long input patient-doctor conversations. Furthermore, the fact that our submissions to all three tasks share the same base (*bart-large*) model

[3]https://docs.aws.amazon.com/sagemaker/latest/dg/notebooks.html

[4]https://huggingface.co/blog/constrained-beam-search

speaks volumes of its versatility. Finally, the results demonstrate the effectiveness of fine-tuning on custom datasets for specialized domains like medicine.

References

Maria Alcocer Alkureishi, Wei Wei Lee, Maureen Lyons, Valerie G Press, Sara Imam, Akua Nkansah-Amankra, Deb Werner, and Vineet M Arora. 2016. Impact of electronic medical record use on the patient–doctor relationship and communication: a systematic review. *Journal of general internal medicine*, 31:548–560.

Asma Ben Abacha, Wen-wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen. 2023. Overview of the mediqa-chat 2023 shared tasks on the summarization and generation of doctor-patient conversations. In *ACL-ClinicalNLP 2023*.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, Dubrovnik, Croatia. Association for Computational Linguistics.

Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2021. Medically aware gpt-3 as a data generator for medical dialogue summarization. In *Machine Learning for Healthcare Conference*, pages 354–372. PMLR.

Sumit Chopra, Michael Auli, and Alexander M Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 93–98.

Seppo Enarvi, Marilisa Amoia, Miguel Del-Agua Teba, Brian Delaney, Frank Diehl, Stefan Hahn, Kristina Harris, Liam McGrath, Yue Pan, Joel Pinto, et al. 2020. Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In *Proceedings of the first workshop on natural language processing for medical conversations*, pages 22–30.

Alexios Gidiotis and Grigorios Tsoumakas. 2020. A divide-and-conquer approach to the summarization of long documents. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:3029–3040.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. 2000. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023. Mimic-iv-note: Deidentified free-text clinical notes.

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. *arXiv preprint arXiv:2009.08666*.

Tom Knoll, Francesco Moramarco, Alex Papadopoulos Korfiatis, Rachel Young, Claudia Ruffini, Mark Perera, Christian Perstl, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. User-driven research of medical note generation software. *arXiv preprint arXiv:2205.02549*.

Kundan Krishna, Sopan Khosla, Jeffrey P Bigham, and Zachary C Lipton. 2020. Generating soap notes from doctor-patient conversations using modular summarization techniques. *arXiv preprint arXiv:2005.01795*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

George Michalopoulos, Kyle Williams, Gagandeep Singh, and Thomas Lin. 2022. Medicalsum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4741–4749.

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. 2022. Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial intelligence review*, pages 1–101.

Cheryl Rathert, Jessica N Mittler, Sudeep Banerjee, and Jennifer McDaniel. 2017. Patient-centered communication in the era of electronic health records: What does the evidence say? *Patient education and counseling*, 100(1):50–64.

Tait D Shanafelt, Lotte N Dyrbye, Christine Sinsky, Omar Hasan, Daniel Satele, Jeff Sloan, and Colin P West. 2016. Relationship between clerical burden and characteristics of the electronic environment with physician burnout and professional satisfaction. In

Mayo Clinic Proceedings, volume 91, pages 836–848. Elsevier.

Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. 2021. Text data augmentation for deep learning. *Journal of big Data*, 8:1–34.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Marieke M van Buchem, Hileen Boosman, Martijn P Bauer, Ilse MJ Kant, Simone A Cammel, and Ewout W Steyerberg. 2021. The digital scribe in clinical practice: a scoping review and research agenda. *NPJ digital medicine*, 4(1):57.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Jeff Wu, Long Ouyang, Daniel M Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Submitted to Nature Scientific Data*.

Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed H Awadallah, Dragomir Radev, and Rui Zhang. 2021. Summˆ n: A multi-stage summarization framework for long input dialogues and documents. *arXiv preprint arXiv:2110.10150*.

Xiaodan Zhu and Gerald Penn. 2006. Summarization of spontaneous conversations. In *Ninth International Conference on Spoken Language Processing*.

Method for Designing Semantic Annotation of Sepsis Signs in Clinical Text

Melissa Y. Yan[1], **Lise Tuset Gustad** [2,3], **Lise Husby Høvik**[3,4], **Øystein Nytrø**[1]
[1] Dept. of Computer Science, Norwegian University of Science and Technology
[2] Faculty of Nursing and Health Sciences, Nord University
[3] Dept. of Circulation and Medical Imaging, Norwegian University of Science and Technology
[4] Clinic of Anaesthesia and Intensive Care, St. Olavs Hospital, Trondheim University Hospital
{melissa.yan, lise.t.gustad, lise.hovik, nytroe}@ntnu.no

Abstract

Annotated clinical text corpora are essential for machine learning studies that model and predict care processes and disease progression. However, few studies describe the necessary experimental design of the annotation guideline and annotation phases. This makes replication, reuse, and adoption challenging.

Using clinical questions about sepsis, we designed a semantic annotation guideline to capture sepsis signs from clinical text. The clinical questions aid guideline design, application, and evaluation. Our method incrementally evaluates each change in the guideline by testing the resulting annotated corpus using clinical questions. Additionally, our method uses inter-annotator agreement to judge the annotator compliance and quality of the guideline. We show that the method, combined with controlled design increments, is simple and allows the development and measurable improvement of a purpose-built semantic annotation guideline. We believe that our approach is useful for incremental design of semantic annotation guidelines in general.

1 Introduction

Annotated clinical text corpora provide natural language processing (NLP) and machine learning (ML) studies the data necessary to find patterns, classify, and predict patient risk and disease progression. Compared to models that only utilize structured data from the electronic health record (EHR), many studies and reviews have shown that model performance can increase by incorporating unstructured clinical text (Soguero-Ruíz et al., 2016; Huddar et al., 2016; Culliton et al., 2017; Assale et al., 2019; Sheikhalishahi et al., 2019; Spasic and Nenadic, 2020).

Pre-existing annotated clinical corpora include the Medical Information Mart for Intensive Care (MIMIC-III) (Johnson et al., 2016), the Clinical E-Science Framework (CLEF) (Roberts et al., 2007), and the Informatics for Integrating Biology and the Bedside (i2b2) challenges and National NLP Clinical Challenges (n2c2) (Uzuner and Stubbs, 2015; Luo et al., 2020). However, studies utilizing pre-existing annotated corpora must limit their research questions to the specific purpose(s) for which the corpus was annotated. Otherwise, the annotations required to answer a research question might be missing or too general. Thus, many studies opt to develop their own annotated clinical corpus tailored to capture and extract the necessary information for their research (Yim et al., 2015; Rama et al., 2018; South et al., 2009; Oliveira et al., 2022).

Methods with lower requirements for supervision, such as information extraction, commonly use keyword search, rule-based algorithms, and ML to detect clinical cases. However, those methods might not consider the context of the clinical case (Ford et al., 2016). For example, different documented signs within a specific situation can describe a medical condition that is not named. Hence, medical expertise is necessary for making annotation judgments and capturing clinical knowledge within the text (Xia and Yetisgen-Yildiz, 2012). Retrieving domain-specific patient knowledge to ascertain or answer clinical questions includes extracting data, information, and knowledge. Data are attributes (e.g., names or dates), information gives meaning to data (e.g., location, cause, and time), and knowledge interprets information based on one's role and responsibility (e.g., clinical document's purpose and effect) (Gudea, 2005).

Making a quality annotated corpus is an iterative process that includes designing an annotation guideline, annotating text with the guideline, and refining the guideline based on inter-annotator agreement (IAA) (Roberts et al., 2009; Xia and Yetisgen-Yildiz, 2012; Deleger et al., 2012; Savkov et al., 2016; Oliveira et al., 2022). Although studies describe how annotated clinical corpora were made, few studies are explicit about the design process.

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 236–246
July 14, 2023 ©2023 Association for Computational Linguistics

We believe that the acquisition and transformation of clinical questions about the patient cohort into corresponding corpus requirements for retrieving information from the actual text of the annotated corpus should drive the annotation process.

2 Related Work

This section provides an overview of studies that describe the design process leading to an annotation guideline and annotated clinical corpus. Studies that share their annotation challenges or offer improvements are also included.

The CLEF Corpus was semantically annotated to help develop and evaluate the CLEF information extraction system (Roberts et al., 2007, 2009). Free-text documents in the corpus are histopathology reports, imaging reports, and clinical narratives (i.e., discharge summaries, reports, case notes, audits, letters, or narratives to the general practitioner, consultant, referrer, or patient). Initially, templates for the documents using ontology-based entities and relationships were manually filled-in. However, the templates did not directly align with text, and ontology complexity made it time-consuming to fill templates. Thus, Roberts et al. (2009) iteratively developed an annotation guideline based on a simplified version of the original ontology and template definitions. Following established standard NLP annotation methodology (Boisen et al., 2000), 2 clinicians annotated 31 documents over 5 sessions, and a third annotator resolved disagreements (Roberts et al., 2009). However, due to workload and time constraints, resigning annotators could have impacted the corpus quality and size. Thus, Roberts et al. (2009) proposed solutions such as pre-annotated documents and a reduced annotation scope.

The i2b2 challenges have annotated corpora for various purposes. For example, in the i2b2 NLP challenge of extracting patient medication from discharge summaries, 79 annotators from 20 teams annotated 251 discharge summaries in a community annotation experiment (Uzuner et al., 2010). The annotation guideline was developed iteratively in 2 phases before the community annotation. For several iterations in phase 1, university students annotated discharge summaries that were measured for IAA and asked questions to aid revisions. This produced a guideline and 17 annotated discharge summaries for phase 2. Finally, during phase 2, teams annotated discharge summaries using the guideline and addressed inconsistencies within the 17 annotated discharge summaries to produce a refined guideline.

The i2b2 temporal relations corpus contains 310 discharge summaries annotated by 8 annotators (Sun et al., 2013). The annotation guideline was based on the TimeML event and temporal expression specification language (Pustejovsky et al., 2003) and the Temporal Histories of Your Medical Event (THYME) project annotation guidelines. The corpus development process included: a guideline development pilot study, data selection, pre-annotation, annotator training session, 2 annotators annotating pre-annotated documents, an adjudicator who resolved disagreements, and evaluation.

The 2014 i2b2/UTHealth de-identification corpus annotation guideline focuses on removing Protected Health Information (PHI) in longitudinal medical records for automatic de-identification system development (Stubbs and Uzuner, 2015). Introduced PHI subcategories enable downstream analyses to adjust the scope or focus on specific categories. Additionally, they compared parallel and serial annotation processes on pre-annotated and unannotated corpora and found that the process does not affect annotation quality (Stubbs and Uzuner, 2017).

Xia and Yetisgen-Yildiz (2012) utilized a variation of the typical annotation process for 3 different studies. Each study's corpus focused on a specific clinical report, such as radiology, chest x-ray, or intensive care unit reports. The process included: defining a study based on clinical needs, selecting data, gaining ethical approval, writing annotation guidelines, creating annotation tools, annotating, building a system with the corpus, and testing if the system meets clinical needs. Physicians were guideline designers and annotators, whereas NLP researchers provided technical support and built NLP systems with the corpora. Suggestions for improvement included more NLP researcher involvement, consideration for guideline granularity versus annotation time, marking rationale or evidence for a label, and estimating time commitment.

Deleger et al. (2012) developed their annotation guideline by building off a previous guideline. The rest of the methods were similar: defining annotation tasks, selecting data from stratified random sampling, and annotating with 2 annotators. During the annotation process, 2 annotators annotated the same documents, IAA was measured, and consensus sessions were held to resolve disagreements and update the guideline. Using the same annotation process, they built gold standard corpora from clinical trial announcements, US Food and Drug Administration (FDA) drug labels, and EHR clinical notes. This included clinical notes such as discharge summaries, referrals, reports, and notes for consultations, procedures, plans, or progress.

Interested in capturing infections caused by central venous catheters, a nurse specializing in infection annotated 2 745 of 22 174 inspected notes (Røst et al., 2018). Before inspection, duplicate notes were removed. The guideline was a table containing events for annotation. Defined by computer scientists, nurses, and an NLP domain expert, the annotation labels formed a hierarchy starting with generalized events at the top level and more specific events below. They also provided information about data access restrictions to promote patient confidentiality and clinical record extraction. Record extraction included physician and nurse notes for admissions, care, plans, evaluations, transfers, and discharge summaries.

In this study, we focus on a method of incremental annotation guideline design by intertwining acquisition with testing of corpus requirements and corresponding annotation phases. This ensures that the guideline produces an annotated corpus that fulfills corpus requirements derived from clinical questions, even if the clinical questions are not answerable by the actual data. To the best of our knowledge, there lacks a study that describes this approach in detail.

3 Objective

This study aims to describe our method for designing a semantically annotated corpus for signs of sepsis by starting from clinical questions that formulate the corpus requirements. Hence, the main contributions are: (1) providing a detailed description of the guideline design process before annotation, (2) illustrating the systematic and iterative annotation process taken, and (3) discussing insights from the design and annotation process.

3.1 Clinical Problem

Sepsis leads to life-threatening multi-organ failure and is caused by a dysregulated host immune response to an infection (Singer et al., 2016). One infectious agent is the *Staphylococcus aureus* (*S. aureus*) bacteria found on skin that is known to cause serious bloodstream infections (BSIs). There is a known overlap between sepsis and BSI, as BSI is found in 30–58% of sepsis patients depending on which sepsis definition is used (Phua et al., 2013; Mellhammar et al., 2021). An estimated 7.6%–35% of *S. aureus* BSIs are related to peripheral intravenous catheters (PIVCs), and the presence of phlebitis can indicate infection via PIVC (Mermel, 2017). A PIVC is a medical device inserted into a vein for administering intravenous (IV) fluids, medication, and blood transfusions. Unfortunately, improperly managed PIVCs can become gateways that lead to phlebitis, BSI, or sepsis (Zhang et al., 2016).

Despite the high sepsis mortality rates and routine usage of PIVCs, both sepsis and PIVCs are poorly documented in clinical text and rarely available as structured data in the EHR (Rohde et al., 2013; Alexandrou et al., 2018). This makes it challenging for hospitals to perform retrospective systematic quality surveillance of PIVC-related BSIs to lower sepsis incidents. Additionally, the lack of explicit documentation inhibits the opportunities for clinicians to learn from and improve PIVC care practices to lower BSI and sepsis rates.

4 Original Adverse Event Dataset

We had access to 18 555 Norwegian adverse event (AE) reports extracted from a hospital's electronic incident reporting system (Yan et al., 2021). Extracted AE reports described procedural deviations, misunderstandings, resource needs, and risky patient behavior. Each report has structured data (i.e., identifier, registration date, reporting hospital unit, if the event is patient-related or security-related, event type, and event severity) and an unstructured free-text note.

5 Semantic Annotation Design Process

This section presents the semantic annotation design process leading up to the annotation process and guideline development. A summary can be found in Figure 1.

5.1 Clarify and Operationalize Clinical Questions to Form Corpus Requirements

Curious about PIVC-related BSI or phlebitis that can lead to sepsis and opportunities to improve patient care, nurses proposed the clinical question: "Is there a connection between PIVCs and BSIs or PIVCs and phlebitis at the hospital?" Thus, the

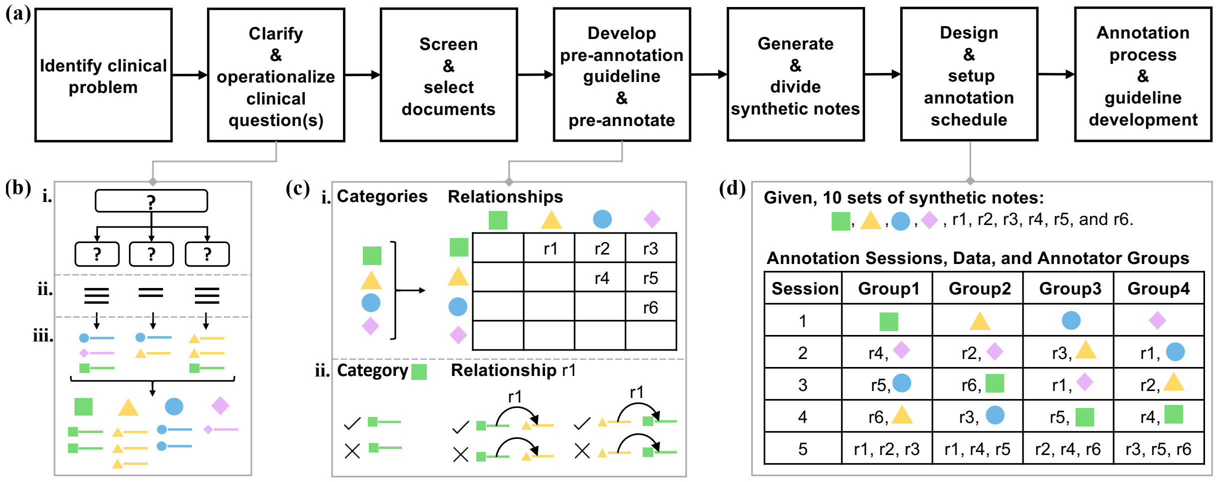


Figure 1: Semantic annotation design process. **(a)** Overview of the process until annotation and guideline development. **(b)** Clarify and operationalize clinical questions into corpus requirements to form annotation categories or entities. **i.** Clarify and operationalize clinical questions by expanding them to derive corpus requirements. **ii.** List examples to answer each question. **iii.** Sort examples into different categories to form the annotation categories. **(c)** Develop the pre-annotation guideline and pre-annotate. **i.** Find relationships using unique category combinations. **ii.** Create the pre-annotation guideline using concrete examples and counterexamples for categories and relationships. **(d)** Determine the annotation sessions and annotator groups to create a schedule. Divide synthetic notes into sets based on the number of categories and relationships. Each group annotates each category at least once in a different session. Additionally, each relationship is annotated at least twice by a different group throughout the sessions. Thus, the sets can be reused in different sessions by different groups, and guideline revisions can be tested on a different group using the same data.

clinical need is to identify PIVC-related BSI and phlebitis or sepsis signs, preferably by automatically classifying patients with PIVCs requiring follow-up care. Through iterative discussions with nurses and computer scientists, the clinical question was clarified to ensure data, information, and knowledge could be extracted to answer the clinical question (Figure 1 **(b)**i). Thus, the clinical question was clarified by expanding it into:

1. How can sepsis or BSIs be identified when the symptoms are similar to other diseases?
2. What signs or symptoms does PIVC-related phlebitis have?
3. How can poorly documented PIVCs be identified?

Those clinical questions were further modified based on the nurses' perspectives. For example, certain types of catheters are distinctly documented (for data extraction). Other catheters can be distinguished based on anatomical insertion sites (for information extraction) or procedures (for knowledge extraction). This resulted in the following questions that also operationalize and form the corpus requirements:

1. What are the different documented signs of infections or phlebitis, specifically those related to PIVCs, BSIs, or sepsis?
2. What can distinguish catheter types in the notes?
3. Where are the documented anatomical insertion sites of catheters?
4. What procedures, interventions, and activities can be related to catheter use from text content or report structured data (e.g., ward type or care situation)?

Figure 2 shows how clinical questions guide the design, application, and evaluation of the annotated corpus, annotation guideline, and corpus requirements.

Creating an annotated clinical corpus is time-consuming and labor-intensive (Wei et al., 2018). However, discussions revealed that we could not reuse a corpus and needed a new annotation guideline. Corpus requirements provided the annotation purpose and can be viewed as "information requests" to develop procedures for extracting data, information, and knowledge through annotation. Extracted data can be facts and observations, such

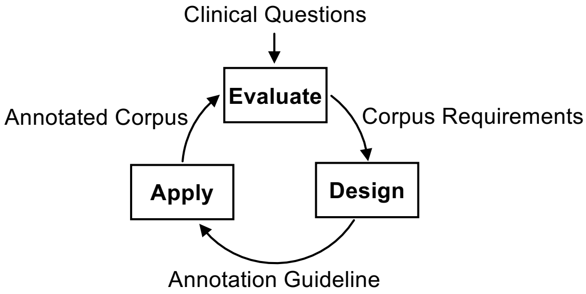


Figure 2: Design driven by clinical questions. Evaluating clinical questions forms the annotated corpus requirements used to design the annotation guideline. Annotators apply the annotation guideline to make a new (sub)corpus. The corpus is evaluated using inter-annotator agreement for annotator compliance and guideline comprehension. Clinical questions are used separately afterward to evaluate the corpus and requirements.

as dates, signs, or symptoms (e.g., purple skin). Information extracted can be phrases for specific signs and symptoms of a case (e.g., purple skin is a sign of a bruise). Furthermore, knowledge extracted can be other signs or symptoms that indicate something not necessarily mentioned (e.g., bruise color can indicate the stage).

Clarifying and operationalizing clinical questions helped determine corpus requirements about documented patient features, patient states, and care features. Including clinicians and computer scientists when clarifying questions was essential because it helped identify requirements for representing knowledge populated by text processing. Furthermore, these questions can be used to evaluate if the annotated corpus can answer the clinical questions.

5.2 Form Annotation Categories or Entities

Clinicians provided examples for the corpus requirements by listing keywords, phrases, and sentences (Figure 1 **(b)**ii). Computer scientists asked clarifying questions to resolve confusion and ambiguity. They also inquired about clinical actions versus actual documented actions to understand what is documented in the text. After generating a list of answers, answers were sorted into different categories (technically known as entities) through discussions (Figure 1 **(b)**iii). Each category is a label for a single word or phrase.

Answers were sorted into the 4 categories: **Sign**, **Location**, **Device**, and **Procedure**. Two additional categories, **Sensitivity** and **Person**, were included to ensure that data is de-identified and that the 4 categories can be linked to an individual. Thus, the 7 main categories are as follows:

1. **Sign**: infection signs
2. **Location**: anatomical insertion sites
3. **Device**: signs of catheter types
4. **Procedure**: catheter acts or interventions
5. **Sensitivity**: potential patient identifiers
6. **Person**: role (e.g., patient or clinician)
7. **Whole**: AE note topic label for validation (i.e., has patient identifier or is about infection, BSI, sepsis, faulty device, catheter, and/or PIVC).

Excluding the **Whole** category, the remaining 6 categories each form a hierarchy with more specific subcategories underneath. Subcategories are used to capture more detailed granularity from the text (e.g., the **Device** category contains a “Catheter” subcategory with different specific catheter types as subcategories).

Concrete categories made understanding the clinical annotation task easier and less ambiguous for the multidisciplinary research group. Having discussions and generating a list with clinicians helped determine the categories and subcategories needed to extract necessary data, information, and knowledge.

5.3 Screen and Select Notes

To ensure that categories specified above are present in notes, 700 randomly selected AE notes were manually screened and categorized by a computer scientist and nurse. Categorizing notes included providing a comment about the categorization rationale and marking potentially ambiguous notes. In addition, the potentially ambiguous notes were clarified in discussions and used as examples for properly annotating notes. Screening notes identified documented information that could satisfy corpus requirements and help answer clinical questions in downstream analyses. Additionally, it provided examples that drove preliminary guideline development in the next section.

5.4 Develop Pre-Annotation Guideline and Pre-Annotate

Initially, 6 possible relationships were found using a table with unique category combinations (Figure 1 (**c**)i). Then, those 6 relationships were discussed within the research group to evaluate which were required and merged. This resulted in the following 4 relationships for linking categories:

1. **Person** $\xrightarrow{\text{Person has}}$ **Sign**, **Location**, **Device**, or **Procedure**
2. **Procedure** $\xrightarrow{\text{Procedure uses}}$ **Device**
3. **Sign** $\xrightarrow{\text{Caused by}}$ **Device** or **Procedure**
4. **Sign**, **Device**, or **Procedure** $\xrightarrow{\text{Located nearby/on/in}}$ **Location**

Before actual annotation, the preliminary annotation guideline underwent a pre-annotation phase. Two pre-annotation guidelines were created to assess the utility and decide how detailed an annotation guideline should be for consistent annotation. The low granularity guideline was a Word document that provided brief instructions, a hierarchical list of categories, and only annotation examples for 2 categories (i.e., **Sensitivity** and **Person**). In contrast, the high granularity guideline was a static HTML webpage with interactive instructions for using the annotation tool and had links to corresponding sections for each category or relationship. Each category and the relationships in the high granularity guideline contained 1 concrete annotation example and counterexamples as needed (Figure 1 (**c**)ii). A nurse and a computer scientist used both pre-annotation guidelines to annotate 15-27 notes. Afterward, the research group determined a high granularity annotation guideline was more informative and easier to use with the annotation tool.

Capturing relationships between categories ensures that data is not lost in downstream analysis (e.g., infection signs at a specific location). It can also provide additional support to answer the clinical questions. By merging relationships, the complexity of annotation options was simplified and reduced. It is ideal to reduce the complexity of annotation because making the annotation task too difficult and time-consuming can result in annotators resigning (Roberts et al., 2009). The pre-annotation phase allowed the research group to manually evaluate, discuss, revise, and improve the guideline before use. This included the suitable granularity level and ease of use for the annotators.

5.5 Generate and Divide Synthetic Notes

Synthetic notes appear real and could be real. 100 unique synthetic clinical text notes were manually generated through 2 methods. The first method combines parts of the original notes to create a similar synthetic AE note with manually anonymized patient identifiers, and the content was verified by a nurse. Whereas in the second method, a nurse manually created a note based on possible clinical scenarios with synthetic patients to ensure some notes contained information about catheters and/or infections. The mean, minimum, maximum and median tokens per AE note in the corpus were 45, 4, 316, and 36, respectively. Generating synthetic notes took a couple of workdays for the nurse.

Afterward, the 100 unique synthetic notes were divided into 10 distinct sets with 10 notes each. Each set corresponds to either a category or relationship. The categories utilized in dividing the sets are those related to catheters or infections (i.e., **Sign**, **Location**, **Device**, and **Procedure**). The relationships utilized are the 6 initial possible relationship combinations.

AE notes often contain excessive and potentially identifying information irrelevant for annotating catheter-related events. Thus, relevant and closely related AE notes were selected and combined to use annotator time efficiently. Generating synthetic notes ensures the data is anonymized and usage is optimized, as clinical data is scarce. Additionally, it provides more data for ML analyses and makes the data more easily accessible to other researchers. Separating synthetic data into different categories or relationships ensures that specific labels will be annotated within the dataset. Different sets could be given to different annotators to reuse data and test if annotation guideline revisions improved IAA.

5.6 Design and Set Up Annotation Schedule

The same 4 categories and 6 initial relationships used to divide synthetic notes into 10 sets were used to design the annotation schedule (Figure 1 (**d**)). Categories were separated into groups, and relationships were added such that each group would annotate a relationship that excluded the group's category. Additionally, relationships within the groups were organized such that each relationship was annotated at least twice by 2 different annotator groups to evaluate revisions. This resulted in 4 annotator groups, each with 5 annotation sessions that used a different set of notes and could annotate in parallel. Each group had 2 annotators so that IAA could be measured. This design defined the annotation schedule, the number of annotation groups needed, and how to reuse synthetic notes for guideline development. Furthermore, parallelization for each session helped reduce the project timeline.

6 Annotation Process and Annotation Guideline Development

Following the schedule, synthetic notes were annotated by 4 annotator groups over 5 sessions using a systematic, iterative annotation process for guideline refinement. In each session, 2 annotators from each group annotated notes based on an annotation guideline using the Brat rapid annotation tool (BRAT) (Stenetorp et al., 2012). Afterward, annotations were evaluated for IAA and manually inspected to assess if annotations could fulfill corpus requirements and answer the clinical questions. Text was tokenized and annotation labels were assigned to tokens before measuring the IAA F_1-score. Disagreements and ambiguities were discussed within the research group, and comments from annotators were incorporated. Next, a computer scientist revised the guideline based on discussions. Finally, the process was repeated with a new set of notes and the revised guideline. Figure 3 shows an example sentence annotated by 2 different annotators.

7 General Results from Sessions 1–5

Over 5 sessions, 8 annotators annotated 100 unique synthetic AE notes to produce 770 annotated synthetic AE notes. From session 1, it was clear that subcategory and attribute names should not be used in more than one category, and synonyms should be avoided. For example, simultaneously having "Name" as both a **Sensitivity** subcategory and an attribute for the **Person** category raised questions. Furthermore, annotators left relationships, attributes, and notes unannotated because they felt those notes were irrelevant to answering the clinical questions.

The need for annotating relationships, attributes, and all notes for ML was addressed in session 2. Red font emphasized guideline revisions, and the guideline began with an "Overview of Updated Instructions" section to aid annotators in identifying revisions. In sessions 2 and 3, the main revisions were correcting and including missing subcategories to address annotator concerns.

Session 4 provided a structured terminology for the guideline. A terminology was developed from the guideline to give structure and provide users quick insight into the annotated corpus for downstream analysis (Yan et al., 2023). This restructured the annotation guideline for session 5 by removing ambiguities and allowed AE note querying to answer the clinical questions. For example, the new **Observation** category encompasses the **Sign** category's signs and symptoms and the **Procedure** subcategory "Device malfunction signs." The computer scientist who revised the guidelines misinterpreted clinical knowledge and made incorrect assumptions in the previous sessions, so the terminology and restructured guideline were validated by nurses to ensure medical concepts were used correctly before session 5. The session 1–4 annotation guidelines were made available online[1] for Yan et al. (2021), and the session 5 annotation guideline was added online for this study. IAA for different sessions are in Figure 4.

8 Discussion

8.1 Design and Annotation Process

The annotation guideline development design process focuses on identifying the effect of the guideline on different categories, corpus content, and clinical questions. Categories were developed to answer different clinical questions and focus on localized guideline changes. Revising parts of specific category hierarchies made it possible to make controlled changes to specific subcategories in the annotation guideline and observe the impact on the annotated corpus, IAA, and clinical questions.

The annotation process greatly influences and drives guideline development. Clinical questions led to corpus requirements that developed the annotation guideline, which is applied on the annotated corpus and evaluated by the clinical questions. In turn, evaluating the annotated corpus also either indicates if it is possible to fulfill corpus requirements to answer clinical questions or detects a lack of corpus content needed for the clinical questions. Using the iterative process, we uncovered corpus requirements that the corpus content could not fulfill and could revise the requirements to drive guideline development and annotation.

[1]https://folk.ntnu.no/melissay/ae-guidelines/

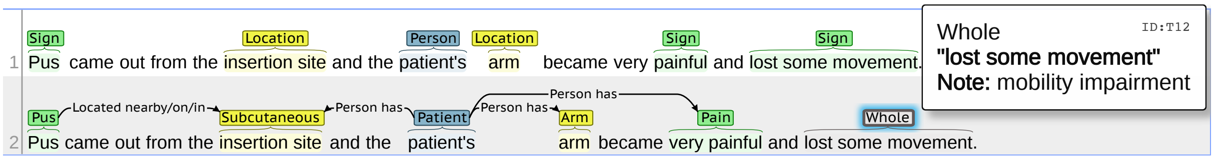


Figure 3: Annotation example for 2 different annotators. Annotator1 on top annotated using only the main categories, whereas Annotator2 on the bottom used subcategories to capture more detail and relationships to link categories. Although the **Whole** category is for indicating if an AE note contains information related to the clinical questions, Annotator2 has misused this label to leave a comment and indicate the phrase is about "mobility impairment". Actual AE notes only contain annotations from 1 annotator, and annotators cannot see the annotations from others.

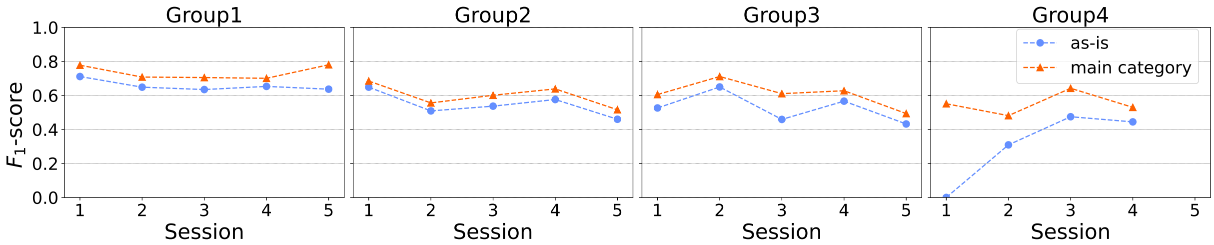


Figure 4: F_1-score for 4 annotator groups over the 5 sessions. The "as-is" F_1-score was calculated using annotator provided labels. Whereas, the "main category" F_1-score converted the labels to the main categories of **Sign**, **Location**, **Device**, **Procedure**, **Sensitivity**, **Person**, or **Whole**. Group4 session 5 has no F_1-score because an annotator withdrew.

8.2 Inter-annotator Agreement

There are several possible reasons for changes in Figure 4's F_1-score. An annotator often misused the **Whole** category to leave comments about clinical knowledge, while this is clinically insightful, it decreases the IAA (e.g., Figure 3 Annotator2 misusing the **Whole** category). The guideline complexity increased and reduced annotator compliance (i.e., sessions 1-5 had 89, 88, 105, 110, and 137 subcategories, respectively). As shown by the "as-is" F_1-score decrease in session 5, the guideline likely became too complex after session 4 revisions. The annotator from Group4 probably withdrew because of the increasing clinical complexity. Another annotator gave feedback that they were uncertain if they annotated some notes correctly. So, increasing the guideline and notes can overwhelm annotators (i.e., sessions 1-5 had 10, 20, 20, 20, and 30 notes, respectively). Group1 was a medical and nursing student, Group2 was a nurse and medical student, Group3 were nurses, and Group4 was a nurse and computer scientist. In general, students followed guidelines well, even if it contained incorrect medical concepts. Thus, paired annotators could have different clinical expertise that impacted results.

Granularity can have an effect on IAA, but granularity can be adjusted to identify problematic subcategories and utilized by those performing downstream analyses. Lower granularity in the annotation guideline leads to higher agreement because it reduces the complexity and level of detail. Annotators usually agree on which main category to annotate a word or phrase, but they had difficulties choosing certain subcategories. For example, in Figure 3 Annotator1 annotated with the main categories whereas Annotator2 was more detailed and annotated almost the same words with subcategories from the same main categories (e.g., "Pus" was annotated by Annotator1 with the **Sign** category and by Annotator2 with **Sign**'s subcategory "Pus"). This is also shown in Figure 4 for Group4 in session 1, where the "as-is" F_1-score is 0, but the "main category" F_1-score is 0.55. It is also possible to perform IAA on different subcategories within a subcategory to identify the most problematic areas after guideline revisions. The granularity in the annotated corpus can also be utilized and adjusted in downstream analyses based on the level of detail required by researchers.

9 Conclusion

Our method captures knowledge about sepsis signs in clinical text. We control changes in the annotation guideline by using hierarchical categories and continuous evaluation. Through applying a systematic, iterative annotation process, we evaluated the changes using the clinical questions and IAA. The clinical questions evaluate corpus quality, and IAA evaluates annotator compliance and guideline complexity. As the guideline is designed to answer different clinical questions, it is possible to adjust the granularity level as needed to answer different clinical questions. By detailing our design process and annotation process, we hope our method can aid other researchers who cannot utilize pre-annotated corpora in developing an annotated corpus for their research.

Limitations

This method for designing and annotating clinical text for a specific clinical use case can be beneficial for researchers needing to annotate a corpus. However, there are some limitations. First, the experiences are based on a specific clinical case and focus on the qualitative aspects. Details of certain parts of the design and annotation process will likely need to be adjusted based on resources available to other researchers. This can include the data selected for annotation, the number of annotators available, and the annotators' level of expertise. For instance, the use case in the design process is based on using 8 annotators to annotate 100 synthetic AE notes over 5 sessions. Second, expertise and additional time are required to generate synthetic notes for annotation. Finally, future work is still needed to replicate the described design and annotation process on other forms of clinical text and problems.

Ethical Considerations

To protect patient privacy when designing and annotating clinical text, synthetic AE notes were manually generated and verified by a nurse to ensure the data is anonymized. Additionally, the annotation guideline includes the **Sensitivity** category to allow annotators to label potential information in the synthetic notes that could identify a patient. This process was described to provide an example for researchers who need to annotate sensitive data.

The Norwegian Regional Committees for Medical and Health Research Ethics (REK) has approved the use of medical data in this study (REK approval no. 26814; 2018/1201/REKmidt). To ensure annotators are protected, collecting and processing personal annotator data has also been approved by the Norwegian Centre for Research Data (NSD reference no. 142683). Furthermore, the annotators have consented to the use of their specified personal information (i.e., profession and years of experience) and their annotations.

Acknowledgements

The authors would like to thank the 8 annotators, this project would not be possible without them. Research is funded by the Computational Sepsis Mining and Modelling project through the Norwegian University of Science and Technology (NTNU) Health Strategic Area.

References

Evan Alexandrou, Gillian Ray-Barruel, Peter J. Carr, Steven A. Frost, Sheila Inwood, Niall Higgins, Frances Lin, Laura Alberto, Leonard Mermel, Claire M. Rickard, and O.M.G. Study Group. 2018. Use of short peripheral intravenous catheters: Characteristics, management, and outcomes worldwide. *Journal of hospital medicine*, 13.

Michela Assale, Linda Greta Dui, Andrea Cina, Andrea Seveso, and Federico Cabitza. 2019. The revival of the notes field: Leveraging the unstructured content in electronic health records. *Frontiers in Medicine*, 6:66.

Sean Boisen, Michael Crystal, Richard M. Schwartz, Rebecca Stone, and Ralph M. Weischedel. 2000. Annotating resources for information extraction. In *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000*, Athens, Greece. European Language Resources Association.

Phil Culliton, Michael Levinson, Alice Ehresman, Joshua Wherry, Jay S Steingrub, and Stephen I Gallant. 2017. Predicting severe sepsis using text from the electronic health record. In *Workshop on Machine Learning For Health at the Conference on Neural Information Processing Systems (NIPS ML4H 2017)*, Long Beach, CA, USA.

Louise Deleger, Qi Li, Todd Lingren, Megan Kaiser, Katalin Molnar, Laura Stoutenborough, Michal Kouril, Keith Marsolo, and Imre Solti. 2012. Building gold standard corpora for medical natural language processing tasks. *AMIA Annual Symposium proceedings. AMIA Symposium.*, 2012:144–153.

Elizabeth Ford, John A. Carroll, Helen E. Smith, Donia Scott, and Jackie Cassell. 2016. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *Journal of the American Medical Informatics Association*, 23:1007 – 1015.

Sorin Gudea. 2005. Data, information, knowledge: a healthcare enterprise case study. *Perspectives in Health Information Management*, 2:8.

Vijay Huddar, Bapu Koundinya Desiraju, Vaibhav Rajan, Sakyajit Bhattacharya, Shourya Roy, and Chandan K. Reddy. 2016. Predicting complications in critical care using heterogeneous clinical data. *IEEE Access*, 4:7988–8001.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad M. Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3.

Yen-Fu Luo, Sam Henry, Yanshan Wang, Feichen Shen, Ozlem Uzuner, and Anna Rumshisky. 2020. The 2019 n2c2/UMass Lowell shared task on clinical concept normalization. *Journal of the American Medical Informatics Association*, 27(10):1529–e1.

Lisa Mellhammar, Fredrik Kahn, Caroline Whitlow, Thomas Kander, Bertil Christensson, and Adam Linder. 2021. Bacteremic sepsis leads to higher mortality when adjusting for confounders with propensity score matching. *Scientific reports*, 11(1):6972.

Leonard A. Mermel. 2017. Short-term peripheral venous catheter-related bloodstream infections: A systematic review. *Clinical Infectious Diseases*, 65(10):1757–1762.

Lucas Emanuel Silva E Oliveira, Ana Carolina Peters, Adalniza Moura Pucca da Silva, Caroline Pilatti Gebeluca, Yohan Bonescki Gumiel, Lilian Mie Mukai Cintho, Deborah Ribeiro Carvalho, Sadid Al Hasan, and Claudia Maria Cabral Moro. 2022. SemClinBr - a multi-institutional and multi-specialty semantically annotated corpus for Portuguese clinical NLP tasks. *Journal of Biomedical Semantics*, 13(1):13.

Jason Phua, Wang Ngerng, Kay See, Chee Tay, Timothy Kiong, Hui Lim, Mei Chew, Hwee Yip, Adeline Tan, Haji Khalizah, Rolando Capistrano, Kang Lee, and Amartya Mukhopadhyay. 2013. Characteristics and outcomes of culture-negative versus culture-positive severe sepsis. *Critical Care*, 17(5):R202.

James Pustejovsky, José M. Castaño, Robert Ingria, Roser Saurí, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. 2003. TimeML: Robust specification of event and temporal expressions in text. In *New Directions in Question Answering, Papers from 2003 AAAI Spring Symposium*, pages 28–34, Stanford University, Stanford, CA, USA. AAAI Press.

Taraka Rama, Pål Brekke, Øystein Nytrø, and Lilja Øvrelid. 2018. Iterative development of family history annotation guidelines using a synthetic corpus of clinical text. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 111–121, Brussels, Belgium. Association for Computational Linguistics.

Angus Roberts, Robert J. Gaizauskas, Mark Hepple, Neil Davis, George Demetriou, Yikun Guo, Jay Kola, Ian Roberts, Andrea Setzer, Archana Tapuria, and Bill Wheeldin. 2007. The CLEF corpus: Semantic annotation of clinical text. *AMIA Annual Symposium proceedings. AMIA Symposium*, pages 625–629.

Angus Roberts, Robert J. Gaizauskas, Mark Hepple, George Demetriou, Yikun Guo, Ian Roberts, and Andrea Setzer. 2009. Building a semantically annotated corpus of clinical texts. *Journal of Biomedical Informatics*, 42(5):950–966.

Jeffrey M. Rohde, Andrew J. Odden, Catherine Bonham, Latoya Kuhn, Preeti N. Malani, Lena M. Chen, Scott A. Flanders, and Theodore J. Iwashyna. 2013. The epidemiology of acute organ system dysfunction from severe sepsis outside of the intensive care unit. *Journal of hospital medicine*, 8:243–247.

Thomas Brox Røst, Christine Raaen Tvedt, Haldor Husby, Ingrid Andås Berg, and Øystein Nytrø. 2018. Capturing Central Venous Catheterization Events in Health Record Texts*. In *IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018*, pages 488–495, Madrid, Spain. IEEE Computer Society.

Aleksandar Savkov, John A. Carroll, Rob Koeling, and Jackie Cassell. 2016. Annotating patient clinical records with syntactic chunks and named entities: the harvey corpus. *Language Resources and Evaluation*, 50:523 – 548.

Seyedmostafa Sheikhalishahi, Riccardo Miotto, Joel T Dudley, Alberto Lavelli, Fabio Rinaldi, and Venet Osmani. 2019. Natural language processing of clinical notes on chronic diseases: Systematic review. *JMIR Medical Informatics*, 7(2):e12239.

Mervyn Singer, Clifford S. Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R. Bernard, Jean-Daniel Chiche, Craig M. Coopersmith, and et al. 2016. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*, 315(8):801–810.

Cristina Soguero-Ruíz, Kristian Hindberg, I. Mora-Jiménez, José Luis Rojo-álvarez, Stein Olav Skrøvseth, Fred Godtliebsen, Kim Mortensen, Arthur Revhaug, Rolv-Ole Lindsetmo, Knut Magne Augestad, and Robert Jenssen. 2016. Predicting colorectal surgical complications using heterogeneous clinical data and kernel methods. *Journal of Biomedical Informatics*, 61:87–96.

Brett R South, Shuying Shen, Makoto Jones, Jennifer Garvin, Matthew H Samore, Wendy W Chapman, and Adi V Gundlapalli. 2009. Developing a manually annotated clinical document corpus to identify phenotypic information for inflammatory bowel disease. *BMC Bioinformatics*, 10 Suppl 9:S12.

Irena Spasic and Goran Nenadic. 2020. Clinical text data in machine learning: Systematic review. *JMIR Medical Informatics*, 8(3):e17984.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Amber Stubbs and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. *Journal of Biomedical Informatics*, 58 Suppl:S20–S29.

Amber Stubbs and Özlem Uzuner. 2017. De-identification of medical records through annotation. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 1433–1459. Springer Netherlands, Dordrecht.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Annotating temporal information in clinical narratives. *Journal of Biomedical Informatics*, 46 Suppl:S5–S12.

Özlem Uzuner, Imre Solti, Fei Xia, and Eithon Cadag. 2010. Community annotation experiment for ground truth generation for the i2b2 medication challenge. *Journal of the American Medical Informatics Association*, 17(5):519–523.

Özlem Uzuner and Amber Stubbs. 2015. Practical applications for natural language processing in clinical research: The 2014 i2b2/UTHealth shared tasks. *Journal of Biomedical Informatics*, 58 Suppl:S1–S5.

Qiang Wei, Amy Franklin, Trevor Cohen, and Hua Xu. 2018. Clinical text annotation - what factors are associated with the cost of time? *AMIA Annual Symposium proceedings. AMIA Symposium*, 2018:1552–1560.

Fei Xia and Meliha Yetisgen-Yildiz. 2012. Clinical corpus annotation: challenges and strategies. In *Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM'2012) in conjunction with the International Conference on Language Resources and Evaluation (LREC 2012)*, pages 32–39.

Melissa Y. Yan, Lise Tuset Gustad, Lise Husby Høvik, and Øystein Nytrø. 2023. Terminology and ontology development for semantic annotation: A use case on sepsis and adverse events. *Semantic Web*, 14(5):811–871.

Melissa Y. Yan, Lise Husby Høvik, André Pedersen, Lise Tuset Gustad, and Øystein Nytrø. 2021. Preliminary processing and analysis of an adverse event dataset for detecting sepsis-related events. In *IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2021, Houston, TX, USA, December 9-12, 2021*, pages 1605–1610. IEEE.

Wen-wai Yim, Sharon Kwan, and Meliha Yetisgen. 2015. In-depth annotation for patient level liver cancer staging. In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*, pages 1–11, Lisbon, Portugal. Association for Computational Linguistics.

Li Zhang, Siyu Cao, Nicole Marsh, Gillian Ray-Barruel, Julie Flynn, Emily Larsen, and Claire M Rickard. 2016. Infection risks associated with peripheral vascular catheters. *Journal of Infection Prevention*, 17(5):207–213.

Prompt Discriminative Language Models for Domain Adaptation

Keming Lu[1], **Peter Potash**[2], **Xihui Lin**[2], **Yuwen Sun**[4], **Zihan Qian**[3], **Zheng Yuan**[5]
Tristan Naumann[2], **Tianxi Cai**[3] and **Junwei Lu**[3]
[1]University of Southern California [2]Microsoft Research [3]Harvard University
[4]Xi'an Jiaotong-Liverpool University [5]Alibaba Damo Academy
[1]keminglu@usc.edu [2]{Peter.Potash,xihlin,tristan}@microsoft.com
[3]{zihanqian,tcai,junweilu}@hsph.harvard.edu
[4]Yuwen.Sun19@student.xjtlu.edu.cn [5]yuanzheng.yuanzhen@alibaba-inc.com

Abstract

Prompt tuning offers an efficient approach to domain adaptation for pretrained language models, which predominantly focus on masked language modeling or generative objectives. However, the potential of discriminative language models in biomedical tasks remains underexplored. To bridge this gap, we develop BioDLM, a method tailored for biomedical domain adaptation of discriminative language models that incorporates prompt-based continual pretraining and prompt tuning for downstream tasks. BioDLM aims to maximize the potential of discriminative language models in low-resource scenarios by reformulating these tasks as span-level corruption detection, thereby enhancing performance on domain-specific tasks and improving the efficiency of continual pertaining. In this way, BioDLM provides a data-efficient domain adaptation method for discriminative language models, effectively enhancing performance on discriminative tasks within the biomedical domain.

1 Introduction

Recent years witnessed the development of biomedical pretrained language models (PLMs) (Kalyan et al., 2022). These domain-specific PLMs contribute to a large number of downstream tasks in the biomedical domain, such as named entity recognition (Yuan et al., 2021; Khandelwal et al., 2022; Watanabe et al., 2022), entity linking (Zhang et al., 2022; Liu et al., 2020), relation extraction (Li et al., 2022a; Sarrouti et al., 2022), and question answering (Jin et al., 2019a; Pappas et al., 2022).

Most existing domain-specific PLMs rely on tremendous in-domain corpus and computing resources for continual pretraining (Lee et al., 2020; Rasmy et al., 2021; Yuan et al., 2022; Alsentzer et al., 2019) or pretraining from scratch (Gu et al., 2021; Yasunaga et al., 2022), which could be infeasible with limited resources. Meanwhile, PLMs

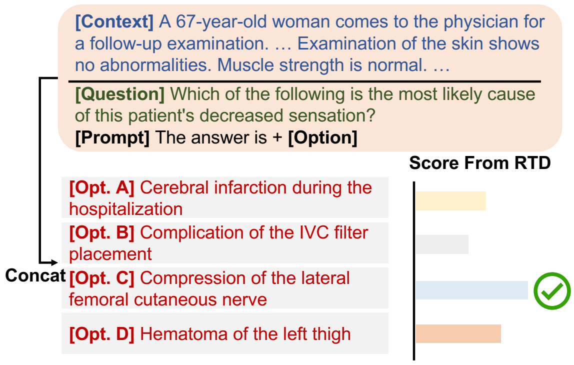


Figure 1: A case for prompting discriminative pretrained language models (DLMs) on multi-choice biomedical question answering. Each option is first concatenated with a predefined hard prompt: "The answer is". They are separately concatenated with the context and question as input. We rank the score from the head of replaced token detection (RTD) in DLMs to determine the best option.

for general purposes usually fails to achieve comparable performance on biomedical tasks with fine-tuning compared with in-domain PLMs at the same model scale (Gu et al., 2021). To combat these issues, exploring a prompt-based domain adaptation method that better leverages existing knowledge learned in pretaining is necessary. Recent research demonstrates that prompts or instructions can activate the hidden abilities of PLMs (Liu et al., 2022; Radford et al., 2019; Brown et al., 2020), including cross-domain inference (Yeh et al., 2022; Fries et al., 2022; Yao et al., 2022b). Therefore, prompt tuning on general PLMs can be a data-efficient domain adaptation method as they are proven promising on various downstream tasks (Wang et al., 2018, 2019).

Existing explorations about prompt-based domain adaptation mainly focus on PLMs with masked language modeling (Lai et al., 2022; Sung et al., 2021) or generative objectives (Luo et al., 2022). However, we identify that discriminative pretrained language models (DLMs) also hold

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 247–258
July 14, 2023

great potential for prompt-based domain adaptation but remains understudied. DLMs are pretrained to distinguish between alternatives and proved to be stronger few-short learners than PLMs with other training objectives (Xia et al., 2022). Therefore, DLMs are better choices for domain adaptation since many downstream tasks in the biomedical domain focus on discriminative objectives (Gu et al., 2021). However, complex model architecture and training recipes hinder DLMs from efficient adaptation to other domains.

To shed light on this topic, we develop BIODLM (Prompt-based Biomedical Domain Adaptation for Discriminative Language Models), which can efficiently take advantage of the state-of-the-art DLMs in the general domain. BIODLM is a prompt-based biomedical domain adaptation method designed explicitly for DLMs, including prompt-based continual pertaining and prompt tuning for downstream tasks. Inspired by Xia et al. (2022), we first formulate discriminative downstream tasks in the biomedical domain, such as multi-choice question answering, as span-level corruption detection.

As shown in Fig. 1, this prompt tuning reformulation allows general-domain DLMs to be used as zero-shot or few-shot learners in biomedical tasks, which is also supported by our probing experiments in §4.2. We develop an efficient prompt-based continual pretraining method to further enhance the performance of DLMs on biomedical tasks. As Bajaj et al. (2022) revealed, the selection of corrupted tokens and the corruption methods play a vital role in pretraining DLMs and is highly related to the performance on downstream tasks. BIODLM selects domain-specific words, defined as different vocabulary between in-domain and general models, as corrupted tokens to lead the continual pretraining focusing on new domain knowledge and improve pretraining efficiency. For corruption, BIODLM employs fixed in-domain PLMs as encoders to corrupt selected tokens instead of co-training encoders and decoders in DLMs. BIODLM is a flexible domain adaptation method that can be applied to any existing DLMs.

The contributions of this work are mainly twofold. First, we explore prompt tuning general-domain DLMs on various biomedical downstream tasks, showing prompting DLMs has significant potential on these tasks under low-resource scenarios. Second, we develop a data-efficient continual pretraining method based on replaced token detection, which employs in-domain PLMs as generators to corrupt domain-specific words in the biomedical corpus. In summary, BIODLM efficiently improves low-resource performance on discriminative tasks in the biomedical domain.

2 Related Works

Discriminative PLMs. Discriminative PLMs (DLMs) incorporate replaced token detection (RTD) or other discriminative objectives during pretraining. Clark et al. (2020) first propose a discriminative pretraining method, which trains a generator to create replaced tokens and a discriminator to distinguish between real and replaced tokens. This approach increases the pretraining efficiency by reducing the computation required in the head compared with previous masked language modeling. Meng et al. (2021) further improves the RTD to corrective language modeling, which requires both RTD and language modeling for correcting the replaced tokens. Bajaj et al. (2022) proposes a more stable and efficient training recipe for DLMs. In this work, we explore domain adaptation for these methods in the biomedical domain. We use METRO-LM (Bajaj et al., 2022) in our experiments of BIODLM since it demonstrates the best performance on general benchmarks, such as GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019).

Prompt tuning for DLMs. Prompt tuning for DLMs is an emerging topic in general and biomedical domains. Ni and Kao (2022) presents empirical evidence showing that ELECTRA can perform well on downstream tasks without fine-tuning or additional training. Xia et al. (2022) introduces a prompt-based fine-tuning approach that leverages discriminative prompts to guide the model towards learning specific downstream tasks with only a few examples. Li et al. (2022b) proposes a few-shot learning approach with pre-trained token-replaced detection models to transform traditional classification and regression tasks into token-replaced detection problems. Yao et al. (2022a) suggests fine-tuning DLMs with prompts for task-specific downstream tasks by adding a small number of task-specific parameters as a prompt to guide the model's output. However, these works are limited to a single method ELECTRA and do not explore biomedical tasks. We follow the recipe of prompt tuning in Xia et al. (2022) and use it on biomedical discriminative tasks.

Biomedical Domain Adaptation. Biomedical domain adaptation of PLMs is a fast-developed topic summarized adequately in the survey from Kalyan et al. (2022). Therefore, we only provide a highly selected review. Alrowili and Vijay-Shanker (2021) propose a novel method for pre-training large biomedical language models that combine BERT, ALBERT, and ELECTRA architectures. Raj Kanakarajan et al. (2021) propose a biomedical domain-specific language encoder model that extends ELECTRA to obtain state-of-the-art performance on numerous biomedical natural language understanding benchmarks. Tinn et al. (2023) propose PubmedELECTRA, a domain-specific version of ELECTRA by continually pertaining ELECTRA on PubMed articles. Luo et al. (2022) propose a generative pre-trained Transformer language model on a large corpus of biomedical articles for biomedical text generation and mining. Our method, BIODLM proposes another perspective that employs prompt-based continual pretraining to adapt DLMs to the biomedical domain, which is understudied in this topic.

3 Methods

We describe preliminaries (§3.1), prompt-based continual pretraining with RTD (§3.2), and prompt tuning for discriminative PLMs (§3.3).

3.1 Preliminaries

Replaced Token Detection. BIODLM is a prompt-based method based on the RTD task. RTD is one of the core pretraining objectives of DLMs (Clark et al., 2020). During the pretaining of DLMs, the input is a sequence of tokens $\mathbf{x} = \{x_i\}_{i=1}^{n}$, where n is the length of input sequences. A random set of tokens in this sequence is selected and corrupted with a generator by masked language modeling. Predictions from the generator will be used to replace the original tokens to obtain a corrupted input $\tilde{\mathbf{x}} = \{\tilde{x}\}_{i=1}^{n}$. At the same time, token-level binary labels are constructed by $\mathbf{y} = \{I(x_i = \tilde{x}_i)\}_{i=1}^{n}$, where $I(\cdot)$ is the indicator function[1]. The discriminator of DLMs is trained with token-level classification on the corrupted input and corresponding labels to detect the replaced tokens.

Method Overview. Similar to the "pretraining-and-finetuning" workflow, BIODLM involves a prompt-based continual pretraining (§3.2) and a prompt-tuning method on downstream tasks (§3.3). As shown in Fig. 2, BIODLM first builds a domain-specific vocabulary for the prompt-based continual pertaining. Then, we corrupt the original biomedical corpus with a fixed in-domain language model as the generator. The corrupted corpus is used to train the general-domain discriminator with RTD for domain adaptation. After the continual pertaining, we explore prompt tuning with RTD to apply BIODLM to biomedical downstream tasks. We reformulate biomedical discriminative tasks into single-token or multi-token RTD, as the example in Fig. 1. BIODLM can also be further tuned on a reformulated training set with RTD objective to enhance downstream performance.

3.2 Prompt-based Continual Pretraining

Continual pretraining on in-domain corpus significantly improve downstream performance on downstream tasks (Gu et al., 2021). However, unlike other training objectives, pretraining with RTD requires self-supervised training corpus construction with corruption. Therefore, we develop a prompt-based continual pretraining method to adapt DMLs to the biomedical domain. The continual pretraining involves a token corruption generator and an RTD discriminator. The recipe of token corruption is essential for both efficiency and effectiveness of the pretraining of DLMs (Bajaj et al., 2022). Therefore, we design a corrupted token selection recipe focusing on in-domain vocabulary and employ fixed in-domain PLMs as generators to corrupt these tokens.

Corrupted Token Selection. Corrupted token selection aims to select the tokens in the in-domain corpus that the generator will corrupt. We first build a domain-specific vocabulary by extracting different tokens from in-domain to general-domain vocabulary. The first challenge is that in-domain and general language models may have very different tokenizers. However, most of them share similar pre-tokenizers to segment context into words. Therefore, we conduct word-level corruption instead of token-level corruption in the traditional design of RTD so that in-domain and general-domain vocabulary can be aligned with each other in the corruption. The detailed selection recipe is described below:

1. We filter tokens that are in in-domain vocabulary but not in the general-domain vocabulary.
2. To conduct word-level corruption, we filter out

[1] The definition of labels may vary in different DLMs. Our introduction follows the recipe in Bajaj et al. (2022).

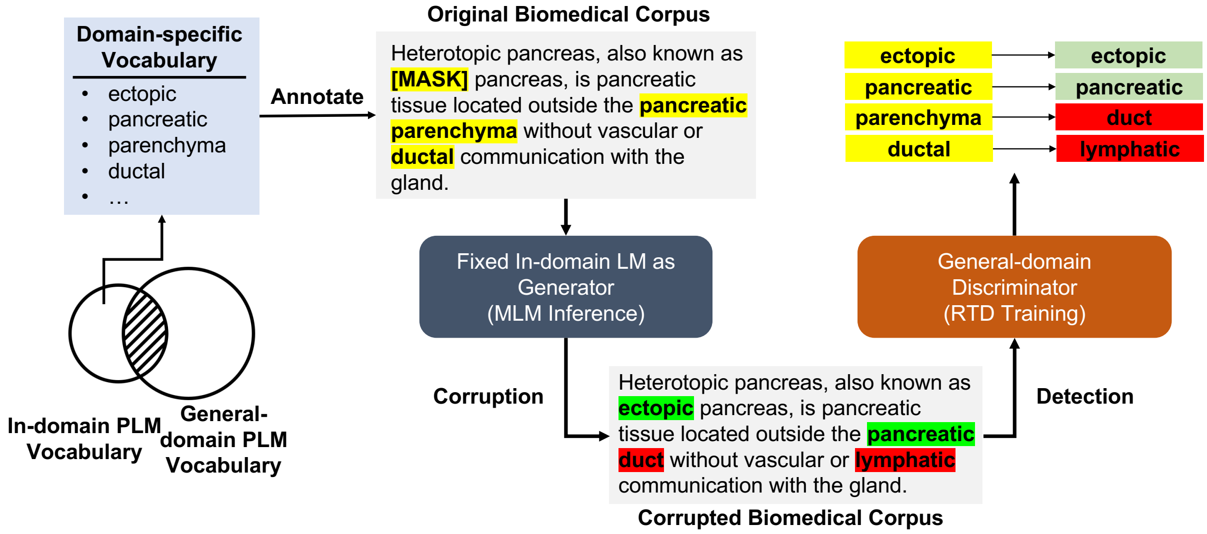


Figure 2: Overview of prompt-based continual petraining in BIODLM. A vocabulary is collected by differing in-domain and general-domain PLMs vocabulary. And we annotate the in-domain corpus with this vocabulary and use this annotation as a set of words for sampling corrupted tokens. Selected tokens are corrupted with a fixed in-domain language model as the generator via masked language modeling inference. The corrupted corpus is then used to continually pretrain a general-domain discriminator with replaced token detection.

all tokens that are not a whole word in the set of tokens we collect in the previous step.

3. We tokenize the remained words in the previous step with the tokenizer of general-domain DLM and filter out any words that contain "unknown" tokens[2]. The rest are our domain-specific vocabulary $\mathcal{D}$.

We use the vocabulary of PubmedBERT (Gu et al., 2021) as our in-domain vocabulary and the vocabulary of MetroLM (Bajaj et al., 2022) as general-domain vocabulary. We eventually have 12,919 words remaining in domain-specific vocabulary $\mathcal{D}$. Most words in $\mathcal{D}$ are biomedical terms, and a sample is listed in §4.3.

Token Corruption. With domain-specific vocabulary $\mathcal{D}$, we employ fixed in-domain LM as a generator to corrupt the in-domain corpus with the inference of masked language modeling. Given an input of the in-domain corpus, such as a PubMed abstract[3], we sample a fixed proportion of words in the input to corrupt. We follow Clark et al. (2020) to set the percentage to 30%. We first pre-tokenize it into words $\mathbf{x} = \{x_i\}_{i=1}^{n}$, where the length of word sequence is n. Then, we identify any domain-specific words in $\mathcal{D}$, denoting them as a bag of words $\mathcal{C}$. The words for corruption are sampled with a strategy that favors domain-specific words:

- $|\mathcal{C}| > \lfloor 0.3n \rfloor$: We randomly select $0.3n$ words from $\mathcal{C}$ as candidates for corruption.
- $|\mathcal{C}| \leqslant \lfloor 0.3n \rfloor$: We randomly select $\lfloor 0.3n \rfloor - |\mathcal{C}|$ words from the rest of the input to meet the requirement of the proportion of corrupted words.

This strategy ensures domain-specific words will be corrupted first, which leads the pretraining to focus on domain knowledge and enhances pretraining efficiency. After identifying the candidates, each word in the candidates will be replaced with a mask token, such as "[MASK]" in the PubmedBERT, and conduct inference of whole-word masked language modeling with the in-domain PLM. The predictions from the in-domain PLM then replace the words in the original inputs to obtain the corrupted in-domain training corpus.

Training. We use the corrupted biomedical corpus for continual pretraining general-domain discriminators with RTD. We conduct word-level corruption—all tokens in corrupted words are labeled with "replaced" and the rest are "original". Otherwise, continual pretraining is the same as §3.1.

3.3 Prompt Tuning with RTD

We explore prompt tuning with RTD on biomedical downstream tasks in BIODLM. Prompt tuning enables DLMs to conduct low-resource inference and

[2]These "unknown" tokens refer to out-of-vocabulary tokens in the general-domain tokenizer, such as the "[UNK]" token in the MetroLM (Bajaj et al., 2022).

[3]PubMed Official Site: https://pubmed.gov

helps DLMs better leverage pretraining knowledge in the general domain. Here, we introduce how to reformulate inputs of biomedical discriminative tasks to conduct low-resource inference with RTD.

Input Reformulation. We follow the recipe from Xia et al. (2022) to prompt DLMs on biomedical downstream tasks. We denote the context as $\mathbf{C}$ and labels as $\mathbf{y} = \{y_i\}_{i=1}^{c}$ of discriminative tasks, where c is the number of labels. We first verbalize labels with predefined words or templates and denote the verbalized templates as $T(\mathbf{y}) = \{t(y_i)\}_{i=1}^{c}$, where $t(\cdot)$ is a manually designed verbalizer for each label. For example, labels from a binary classification task are verbalized as "yes" and "no". As for multi-choice question answering, the labels are already phrases so no verbalization will be applied. Each verbalized label is concatenated with context and a predefined prompt as inputs, denoting as $x = \{C \oplus t(y_i)\}_{i=1}^{c}$, where $\oplus$ is the text concatenation operation. The inputs are fed into the DLMs, and we collect scores from the RTD head within the spans of labels as outputs. The RTD head classifies tokens in labels into "replaced" or "original", where "original" suggests the correct answer to the discriminative problem. The classification scores from the RTD head reveal the semantic correlation between the context and verbalized labels. When verbalized labels are tokenized into more than one token, we use the average RTD scores as the score of these labels. However, the RTD head aims to identify token-level corruption, so averaging multiple tokens do not align well with the pretraining objective and potentially hinders the performance of prompt inference. Therefore, we separately analyze **single-token** and **multi-token** labels in this work. This reformulation allows us to conduct zero-shot prompt inference with DLMs on biomedical discriminative tasks.

Fig. 1 shows a case that we apply prompt inference for DLMs on a multi-choice biomedical question answering dataset. The context is made of a description of the patient background marked in blue and a question marked in green. Then, it is concatenated with four options individually, with a predefined prompt, "The answer is". We consider the average RTD score in each option span as the classification score. And we select the option with the highest average RTD score as the prediction.

Training. In addition to the zero-shot inference, we also conduct prompt tuning on downstream tasks. With the input reformulation described before, discriminative tasks can be reformulated as multi-label binary classification tasks. We further tune the parameters of DLMs in this way to conduct few-shot and fully supervised inference.

4 Experiments

This section introduces an experimental evaluation of prompting discriminative PLMs for biomedical domain adaptation. We describe the experimental setup (§4.1), main results (§4.2), and ablation study (§4.3) on incorporated techniques.

4.1 Experimental Setup

Training corpus. The biomedical corpus for the continual pretraining in this work is the PubMed abstracts in the PubMed Central (PMC) Open Access (OA) Subset[4] (Gamble, 2017; Bethesda, 2003). We process this dump with the open-source tool *pubmed_parser*[5] (Achakulvisut et al., 2020) to extract abstracts of articles. We then follow the preprocessing recipe of Bajaj et al. (2022) and segment the corpus into paragraphs. The original PMC OA Subset contains 21 million paragraphs from biomedical journal articles. We only randomly select three million paragraphs for continual pretraining due to the limitation of computation resources.

Benchmarks. We evaluate BIODLM on five public biomedical datasets: (1) **PubmedQA** (Jin et al., 2019b) contains 1k expert-labeled question-answer pairs based on PubMed abstracts with yes/no/maybe multiple-choice answers. (2) **BioASQ** (Tsatsaronis et al., 2012) is a large question-answering dataset containing biological questions and answers, and related biomedical papers and abstracts. (3) **MedQA(USMLE)** (Jin et al., 2021) is a question-answering dataset containing multiple-choice questions and related answer options in US Medical License Exam (USMLE) format, which were obtained with a choice of 4 or 5 possible answers from the National Medical Board Examination in the United States. (4) **MMLU (Professional Medicine)** (Hendrycks et al., 2020) involves difficult exam questions consisting of four multiple-choice questions with corresponding answers in the biomedical domain. (5) **MedMCQA** (Pal et al., 2022) is a new large-scale

[4] PMC OA Subset: `https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/`

[5] Github repository of *pubmed_parser*: `https://github.com/titipata/pubmed_parser`

	Dataset	Size	Random	Zero-shot (RTD Prompt)			Fully Supervised (CLS Finetuning)			
				MetroLM	Electra	BioElectra	MetroLM	Electra	BioElectra	PubmedBERT
Single	PubmedQA	500	33.3	**64.0**	58.0	48.0	63.8	57.0	62.2	55.8
	BioASQ	140	50.0	**74.3**	73.6	67.1	94.3	73.6	75.7	87.6
Multi	MedQA(USMLE)	1273	25.0	**25.3**	22.1	19.5	28.1	27.4	40.3	39.3
	MMLU	272	25.0	**25.7**	19.9	20.6	27.6	25.8	44.1	29.1
	MedMCQA*	4183	25.0	26.6	**26.8**	20.7	35.5	34.8	40.8	41.2
	Macro Avg.		31.7	**43.2**	40.1	35.2	49.8	43.7	52.6	50.6

Table 1: Probing experiment results display the zero-shot performance with the RTD prompt of various DLMs on the test sets of our benchmark. We also report the CLS-based finetuning performance in the full training setting and involve an in-domain PLM, PubmedBERT, for comparison. We report accuracy on each data split and the macro average accuracy on our benchmark. The best zero-shot performance on each dataset is marked in **bold**. * We report performance on the development set of MedMCQA since we have not received official scores on the test set.

	Dataset	Random	0% (Zero-shot)			10% (Few-shot)			100% (Full)		
			CLS	Prompt	BIODLM	CLS	Prompt	BIODLM	CLS	Prompt	BIODLM
Single	PubmedQA	33.3	31.1	**64.0**	57.0	56.0	**62.8**	58.0	63.8	**69.9**	66.8
	BioASQ	50.0	35.2	74.3	**77.1**	77.9	77.9	**80.0**	**94.3**	85.3	89.8
Multi	MedQA(USMLE)	25.0	9.8	25.3	**27.7**	26.5	25.7	**29.1**	28.1	27.0	**29.6**
	MMLU	25.0	11.0	25.7	**26.8**	21.7	30.8	**32.7**	27.6	31.2	**31.6**
	MedMCQA*	25.0	6.4	26.6	**27.4**	**30.7**	22.9	30.1	**35.5**	27.2	33.2
	Macro Avg.	33.7	18.7	43.1	**43.4**	42.6	44.1	**49.9**	50.0	48.1	**50.2**

Table 2: Results of BIODLM in the zero-shot, few-shot, and full settings compared with finetuning CLS representations on the test sets of our benchmark. We use MetroLM as the backbone in BIODLM for results in this table. The prompt baseline is MetroLM with RTD prompt tuning without continual pretraining in BIODLM. We report accuracy on each data split and the macro average accuracy on our benchmark. Finetuning CLS requires the training of a classification head, so we conduct zero-shot inference of CLS representations by semantic matching between context and options. The best accuracy on each dataset in each setting is marked in **bold**. * We report performance on the development set of MedMCQA since we have not received official scores on the test set.

Multiple-Choice Question Answering dataset containing about 194k 4-option multiple-choice questions from Indian medical entrance exams (AIIMS/NEET). In our benchmark, PubmedQA and BioASQ are single-token datasets as their labels are short as "yes/no/maybe". However, other multi-token datasets, such as MedQA(USMLE), are more challenging since they have longer options and at least four options. We report accuracy scores on the test sets. And we only report the performance on the development set on MedMCQA since we have not received official feedback for the test scores.

Baselines. In the probing experiments, we consider Electra and BioElectra as baselines for MetroLM. We also include PubmedBERT for reference. **Electra** (Clark et al., 2020) is a PLM that uses replaced token detection as a self-supervised task for language representation learning. The central concept of Electra is to train a text encoder to identify input tokens from high-quality negative samples generated by a small generator network, resulting in superior performance on downstream tasks compared to conventional masked language modeling. **BioELECTRA** (Raj Kanakarajan et al., 2021) is a biomedical PLM adapted from the ELECTRA model for the biomedical domain. It is pretrained from scratch on the biomedical domain-specific text and achieves state-of-the-art performance on various biomedical NLP tasks, demonstrating that pretraining from scratch with biomedical domain text enhances the model's capacity. **PubMedBERT** (Gu et al., 2021) is a biomedical PLM that has been pretrained on PubMed abstracts. It achieves state-of-the-art results in several benchmark datasets, making it a strong baseline model for biomedical language understanding tasks. We also include a random baseline, which is the accuracy based on a random guess.

Configurations. We develop BIODLM based on a strong discriminative pretrained language model *MetroLM-base* (Bajaj et al., 2022). This DLM demonstrates the best zero-shot performance in our probing experiments described in §4.2. We run continual pretraining on 8 NVIDIA V100 GPUs for 10 hours and evaluation on each dataset in our benchmark on 1 NVIDIA V100 GPU for less than 1 hour. The hyper-parameters are determined with the grid search based on the accuracy of the development set. Detailed hyper-parameters are shown in Appx. §A.

4.2 Results

We first show the results of a probing experiment to demonstrate DLMs are zero-shot learners on biomedical tasks. Then we present our main results to show the effectiveness of BIODLM on both full- and low-resource scenarios on our benchmark.

Probing Experiments. Tab. 1 shows the probing experiment results about the zero-shot performance of three DLMs with RTD prompt. We also report finetuning results based on the CLS representations of these DLMs, along with the zero-shot prompt tuning performance. First, we notice that models with zero-shot RTD prompt tuning even outperform their finetuning counterparts on several datasets, marked with the underline in Tab. 1. For example, the accuracy of MetroLM with zero-shot RTD prompt tuning in PubmedQA is 64.0, 0.2 absolute percentage higher than its fully supervised finetuning counterpart. Similar cases are also witnessed in other DLMs, such as Electra on the test split of PubmedQA and BioElectra on the development set of BioASQ. These cases show that prompt tuning of general-domain DLMs has great potential as zero-shot learners on biomedical tasks. And these results also provide evidence that reformulating biomedical discriminative tasks as replaced token detection contributes to leveraging general-domain knowledge in pertaining, which is proposed in §3.3. Furthermore, MetroLM significantly outperforms other DLMs on most datasets, achieving 43.2 macro average accuracy. Therefore, we choose MetroLM as the backbone to conduct the following experiments and analyses of BIODLM. Tab. 8 in Appx. §B is an extended version of Tab. 1 containing results on both development and test sets.

Main Results. Tab. 2 shows the main results of BIODLM in the zero-shot, few-shot, and fully supervised settings on the test sets of our benchmark. In the zero-shot setting, BIODLM outperforms MetroLM with only prompt tuning on most datasets, improving macro average accuracy by 0.3 percent. We conduct zero-shot inference with CLS representations by semantic matching between context and options based on CLS representations. However, it can not perform well in the zero-shot setting since the context and options are significantly different. In the few-shot setting, the macro average accuracy of BIODLM is higher than CLS and prompt methods by 7.3% and 5.8%, respectively. These results prove that BIODLM enables general-domain DLMs to conduct inference on biomedical downstream tasks under low-resource scenarios. Furthermore, even though the traditional finetuning method outperforms prompt tuning in the fully supervised setting by 1.9% in accuracy, we notice BIODLM still slightly outperforms the finetuning method by 0.2% on macro average accuracy. This observation suggests that BIODLM benefits from the prompt-based continual pertaining. And we summarize that BIODLM is a better choice under low-resource scenarios, but both traditional CLS finetuning and BIODLM perform well with adequate supervision.

4.3 Study

We provide the following analyses to evaluate further the core components of BIODLM, including corruption methods, prompt templates, and domain-specific vocabulary.

Corruption Methods. In this analysis, we conduct ablation study experiments to demonstrate the effectiveness and data efficiency of the corruption method proposed in BIODLM. We design a random strategy that randomly selects 30% words in the input as the baseline of the domain-specific token selection strategy for corruption. As for the generator, we use the general-domain pretrained language model BERT as the baseline of the in-domain pretrained language model PubmedBERT. We conduct continual pretraining on different combinations of corrupted token selection and generators with 1 million to 3 million samples.

Tab. 3 shows the results of the ablation study on corruption models of BIODLM. Comparing the random and domain-specific token selection, we notice the macro average accuracy on the benchmark of the domain-specific strategy is consistently higher than that of the random strategy. Within each corrupted token selection strategy, PubmedBERT, as the generator, outperforms BERT in most cases, showing that fixed in-domain PLMs with more precise corruption benefit the continual pretraining in BIODLM. Furthermore, we notice the domain-specific token selection strategy with Pub-

Token Selection	Generator	Pretraining Samples 1M	2M	3M
Random	BERT	49.3	49.6	50.0
	PubmedBERT	49.3	49.5	49.9
Domain-specific	BERT	49.7	50.4	51.8
	PubmedBERT	50.2	50.9	52.1

Table 3: Ablation study on corruption methods. We compare two token selection recipes based on Random and In-domain vocabulary and two generators BERT (general-domain) and PubmedBERT (in-domain), with pertaining samples from 1 million to 3 million. We report macro average accuracy scores on our benchmark.

medBERT as the generator used in BIODLM with only 1 million training samples can outperform the random token strategy with 3 million training samples. This result provides valuable insight that corruption methods in BIODLM can significantly improve data efficiency in continual pretraining.

Prompt Templates. Prompt templates play a vital role in prompt tuning. We adopt manually designed prompt templates in BIODLM to verbalize labels and reformulate inputs of discriminative tasks. To better evaluate the influence of manual template design, we construct three prompt templates for two biomedical question answering datasets:

- **Template A**: "[Context]. [Question]? The answer is [prompt label]."
- **Template B**: "[Context] [Question]? The answer is [prompt label]."
- **Template C**: "Context: [Context]. Question: [Question]? The answer is [prompt label]."

There are only minor differences among these templates. Using each prompt template, we then run zero-shot inference with MetroLM and RTD prompt on two datasets. Tab. 5 shows that the design of prompt templates may influence zero-shot performance, which could be related to the specific dataset. It is worth noticing that prompt template B only slightly differs from prompt template A but performance on the test set of BioASQ dropped by half, suggesting an obvious spurious correlation on the punctuation in prompt templates.

We also conduct additional prompt ablation studies on the multi-token prompt datasets. We have manually designed two prompts for multi-choice question-answering datasets in our benchmark:

- **Template D**: "[Context]. [Question]? The answer is [Option]."

Prompt	MedQA(USMLE) dev	test	MMLU dev	test	MedMCQA dev
D	27.6	25.3	38.7	25.7	26.6
E	25.9	25.1	31.2	23.9	24.0

Table 4: Zero-shot accuracy of MetroLM with RTD prompting on multi-token prompt datasets with two manually designed prompt templates.

Prompt	PubmedQA dev	test	BioASQ dev	test
A	50.2	64.0	72.0	74.3
B	50.2	64.0	62.9	38.7
C	48.6	68.0	71.7	72.3

Table 5: Zero-shot accuracy of MetroLM with RTD prompting on BiomedQA and BioASQ with three manually designed prompt templates.

- **Template E**: "[Context] [Question]? The answer [Option] is [right/wrong]."

The underlined spans include tokens for the RTD. Template D is used in our main results, while template E reformulates multi-token prompts into single-token prompts by simply judging whether the option is right or wrong. Tab. 4 shows the results of these two templates. Template D consistently outperforms template E, suggesting direct RTD on the option spans works better in our multi-token prompt datasets. Therefore, prompt templates need to be carefully designed to achieve the best performance on each dataset.

Domain-specific Vocabulary. We present a brief case study of vocabulary differences between in-domain and general-domain PLMs to justify our design in the corrupted token selection. §4.3 shows cases in the domain-specific vocabulary and their corresponding categories. Most words in this vocabulary fall into categories such as Gene, Protein, Disease, Chemical, and Drug. These categories contain rich biomedical terms frequently used in the downstream tasks. Therefore, continual pretraining on the domain-specific vocabulary helps DLMs focus on biomedical knowledge and improves data efficiency of domain adaptation.

5 Conclusion

We study an efficient way to adapt general-domain DLMs to the biomedical domain and propose BIODLM. BIODLM consists of data-efficient continual pretraining that focuses on domain-specific vocabulary and leverages domain knowledge in the

Categories	Words
Gene & Protein	TGFβ1, IGF1R, phosphatases,Synaptophysin
Disease	Adenomatous,malarial, atherosclerotic, cholangiocarcinoma
Chemical & Drug	Phosphatidylcholine, cycloheximide,azithromycin, minocycline,hygromycin, Methylprednisolone

Table 6: A case study of domain-specific vocabulary used for continual pretraining. We present randomly-selected words and their categories in this vocabulary.

in-domain PLMs by employing them as RTD generators. We also conduct experiments on a biomedical benchmark with six biomedical datasets, verifying that prompt tuning is an effective way to adapt DLMs on biomedical discriminative tasks directly. Future works include extending BIODLM to more DLMs, such as ELECTRA (Clark et al., 2020) and COCO-LM (Meng et al., 2021), and experimenting with BIODLM on other discriminative tasks in the biomedical domain.

Limitations

BIODLM adopts DLMs as backbone models. Compared to PLMs with other training objectives, DLMs may miss language modeling benefits and squeeze representation space. Besides, our benchmarks can be extended to more biomedical discriminative tasks, such as relation extraction, document classification, and entity disambiguation. We consider extending our exploration to more DLMs and biomedical tasks as valuable future works.

Ethics Statement

All datasets in our benchmark and continual pretraining are obtained according to each dataset's respective data usage policy.

References

Titipat Achakulvisut, Daniel Acuna, and Konrad Kording. 2020. Pubmed parser: A python parser for pubmed open-access xml subset and medline xml dataset xml dataset. *Journal of Open Source Software*, 5(46):1979.

Sultan Alrowili and K Vijay-Shanker. 2021. Biomtransformers: building large biomedical language models with bert, albert and electra. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 221–227.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Payal Bajaj, Chenyan Xiong, Guolin Ke, Xiaodong Liu, Di He, Saurabh Tiwary, Tie-Yan Liu, Paul Bennett, Xia Song, and Jianfeng Gao. 2022. Metro: Efficient denoising pretraining of large scale autoencoding language models with model generated signals. *arXiv preprint arXiv:2204.06644*.

Bethesda. 2003. PMC Open Access Subset. *National Library of Medicine*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Jason Fries, Leon Weber, Natasha Seelam, Gabriel Altay, Debajyoti Datta, Samuele Garda, Sunny Kang, Rosaline Su, Wojciech Kusa, Samuel Cahyawijaya, et al. 2022. Bigbio: A framework for data-centric biomedical natural language processing. *Advances in Neural Information Processing Systems*, 35:25792–25806.

Alyson Gamble. 2017. Pubmed central (pmc). *The Charleston Advisor*, 19(2):48–54.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019a. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the*

9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019b. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2022. Ammu: a survey of transformer-based biomedical pretrained language models. *Journal of biomedical informatics*, 126:103982.

Anshita Khandelwal, Alok Kar, Veera Raghavendra Chikka, and Kamalakar Karlapalem. 2022. Biomedical NER using novel schema and distant supervision. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 155–160, Dublin, Ireland. Association for Computational Linguistics.

Zhaohong Lai, Biao Fu, Shangfei Wei, and Xiaodong Shi. 2022. Continuous prompt enhanced biomedical entity normalization. In *Natural Language Processing and Chinese Computing: 11th CCF International Conference, NLPCC 2022, Guilin, China, September 24–25, 2022, Proceedings, Part II*, pages 61–72. Springer.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Lishuang Li, Ruiyuan Lian, Hongbin Lu, and Jingyao Tang. 2022a. Document-level biomedical relation extraction based on multi-dimensional fusion information and multi-granularity logical reasoning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2098–2107, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Zicheng Li, Shoushan Li, and Guodong Zhou. 2022b. Pre-trained token-replaced detection model as few-shot learner. *arXiv preprint arXiv:2203.03235*.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2020. Self-alignment pretraining for biomedical entity representations. *arXiv preprint arXiv:2010.11784*.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6).

Yu Meng, Chenyan Xiong, Payal Bajaj, Paul Bennett, Jiawei Han, Xia Song, et al. 2021. Coco-lm: Correcting and contrasting text sequences for language model pretraining. *Advances in Neural Information Processing Systems*, 34:23102–23114.

Shiwen Ni and Hung-Yu Kao. 2022. Electra is a zero-shot learner, too. *arXiv preprint arXiv:2207.08141*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, pages 248–260. PMLR.

Dimitris Pappas, Prodromos Malakasiotis, and Ion Androutsopoulos. 2022. Data augmentation for biomedical factoid question answering. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 63–81, Dublin, Ireland. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Kamal Raj Kanakarajan, Bhuvana Kundumani, and Malaikannan Sankarasubbu. 2021. Bioelectra: pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 143–154.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86.

Mourad Sarrouti, Carson Tao, and Yoann Mamy Randriamihaja. 2022. Comparing encoder-only and encoder-decoder transformers for relation extraction from biomedical texts: An empirical study on ten benchmark datasets. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 376–382, Dublin, Ireland. Association for Computational Linguistics.

Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. 2021. Can language models be biomedical knowledge bases? *arXiv preprint arXiv:2109.07154*.

Robert Tinn, Hao Cheng, Yu Gu, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2023. Fine-tuning large neural language models for biomedical natural language processing. *Patterns*, 4(4).

George Tsatsaronis, Michael Schroeder, Georgios Paliouras, Yannis Almirantis, Ion Androutsopoulos, Eric Gaussier, Patrick Gallinari, Thierry Artieres, Michael R Alvers, Matthias Zschunke, et al. 2012. Bioasq: A challenge on large-scale biomedical semantic indexing and question answering. In *AAAI fall symposium: Information retrieval and knowledge discovery in biomedical text*. Arlington, VA: Citeseer.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Taiki Watanabe, Tomoya Ichikawa, Akihiro Tamura, Tomoya Iwakura, Chunpeng Ma, and Tsuneo Kato. 2022. Auxiliary learning for named entity recognition with multiple auxiliary biomedical training data. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 130–139, Dublin, Ireland. Association for Computational Linguistics.

Mengzhou Xia, Mikel Artetxe, Jingfei Du, Danqi Chen, and Ves Stoyanov. 2022. Prompting electra: Few-shot learning with discriminative pre-trained models. *arXiv preprint arXiv:2205.15223*.

Yuan Yao, Bowen Dong, Ao Zhang, Zhengyan Zhang, Ruobing Xie, Zhiyuan Liu, Leyu Lin, Maosong Sun, and Jianyong Wang. 2022a. Prompt tuning for discriminative pre-trained language models. *arXiv preprint arXiv:2205.11166*.

Zonghai Yao, Yi Cao, Zhichao Yang, and Hong Yu. 2022b. Context variance evaluation of pretrained language models for prompt-based biomedical knowledge probing. *arXiv preprint arXiv:2211.10265*.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016, Dublin, Ireland. Association for Computational Linguistics.

Hui-Syuan Yeh, Thomas Lavergne, and Pierre Zweigenbaum. 2022. Decorate the examples: A simple method of prompt design for biomedical relation extraction. *arXiv preprint arXiv:2204.10360*.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. BioBART: Pretraining and evaluation of a biomedical generative language model. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 97–109, Dublin, Ireland. Association for Computational Linguistics.

Zheng Yuan, Yijia Liu, Chuanqi Tan, Songfang Huang, and Fei Huang. 2021. Improving biomedical pretrained language models with knowledge. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 180–190, Online. Association for Computational Linguistics.

Sheng Zhang, Hao Cheng, Shikhar Vashishth, Cliff Wong, Jinfeng Xiao, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Knowledge-rich self-supervision for biomedical entity linking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 868–880.

A Hyper-parameters

Tab. 7 shows details of hyper-parameters in the experiments of continual pretraining and prompt tuning. Hyper-parameters are determined by grid search.

B Comprehensive Results

We demonstrate extensive results, including performance on development sets in Tab. 8 and Tab. 9.

Parameters	Evaluation				Continual Training
	PubmedQA	BioASQ	MedQA(USMLE)	MedMCQA	PubmedQA
Batch Size	8	8	32	32	8
Learning Rate	2e-5	2e-5	5e-5	5e-5	2e-5
Warmup Steps	500	500	1000	1000	100
Epochs	20	20	10	10	1
Max Sequence Length	512	512	512	512	512

Table 7: Hyper-parameters used for BIODLM evaluation and continual training on PubmedQA, BioASQ, MedQA(USMLE), and MedMCQA.

	Dataset	Split	Size	**Random**	**Zero-shot (RTD Prompt)**			**Fully Supervised (CLS Finetuning)**			
					MetroLM	Electra	BioElectra	MetroLM	Electra	BioElectra	PubmedBERT
Single	PubmedQA	dev	50	33.3	**50.2**	46.9	46.4	62.0	56.0	54.0	52.3
		test	500	33.3	**<u>64.0</u>**	<u>58.0</u>	48.0	63.8	57.0	62.2	55.8
	BioASQ	dev	75	50.0	72.0	78.6	**<u>82.7</u>**	93.3	85.3	81.3	89.3
		test	140	50.0	**74.3**	<u>73.6</u>	67.1	94.3	73.6	75.7	87.6
Multi	MedQA(USMLE)	dev	1272	25.0	**27.6**	24.8	18.2	28.5	27.8	43.5	36.8
		test	1273	25.0	**25.3**	22.1	19.5	28.1	27.4	40.3	39.3
	MMLU	dev	31	25.0	**<u>38.7</u>**	29.0	16.1	25.8	29.7	45.2	32.2
		test	272	25.0	**25.7**	19.9	20.6	27.6	25.8	44.1	29.1
	MedMCQA	dev	4183	25.0	26.6	**26.8**	20.7	35.5	34.8	40.8	41.2
	Macro Avg.			32.4	**44.9**	42.2	37.7	51.0	46.4	54.1	51.5

Table 8: Probing experiment results display the zero-shot performance with the RTD prompt of various DLMs on our benchmark. We also report the CLS-based finetuning performance of these DLMs in the full training setting and involve an in-domain PLM, PubmedBERT, for comparison. We report accuracy on each data split and the macro average accuracy on our benchmark. The best zero-shot performance on each dataset is marked in **bold**.

	Dataset	Split	**Random**	**0% (Zero-shot)**			**10% (Few-shot)**			**100% (Full)**		
				CLS	Prompt	BIODLM	CLS	Prompt	BIODLM	CLS	Prompt	BIODLM
Single	PubmedQA	dev	33.3	28.7	50.2	**52.4**	48.4	62.0	**66.0**	62.0	58.7	**66.0**
		test	33.3	31.1	**64.0**	57.0	56.0	**62.8**	58.0	63.8	**69.9**	66.8
	BioASQ	dev	50.0	34.0	72.0	**75.0**	**89.3**	87.9	88.0	**93.3**	90.6	90.7
		test	50.0	35.2	74.3	**77.1**	77.9	77.9	**80.0**	**94.3**	85.3	89.8
Multi	MedQA(USMLE)	dev	25.0	10.4	**27.6**	26.4	25.6	**28.3**	27.9	28.5	25.4	**29.7**
		test	25.0	9.8	25.3	**27.7**	26.5	25.7	**29.1**	28.1	27.0	**29.6**
	MMLU	dev	25.0	4.2	38.7	**39.4**	29.0	29.3	**32.9**	25.8	30.9	**35.4**
		test	25.0	11.0	25.7	**26.8**	21.7	30.8	**32.7**	27.6	31.2	**31.6**
	MedMCQA	dev	25.0	6.4	26.6	**27.4**	**30.7**	22.9	30.1	**35.5**	27.2	33.2
	Macro Avg.		32.4	19.0	44.9	**45.5**	45.0	47.5	**49.4**	51.0	49.6	**52.5**

Table 9: Results of BIODLM in the zero-shot, few-shot, and full settings compared with finetuning CLS representations. We use MetroLM as the backbone in BIODLM for results in this table. The prompt baseline is MetroLM with RTD prompt tuning but without continual pretraining in BIODLM. We report accuracy on each data split and the macro average accuracy on our benchmark. Finetuning CLS requires the training of a classification head, so it is infeasible in the zero-shot setting. The best accuracy on each dataset in each setting is marked in **bold**.

Cross-domain German Medical Named Entity Recognition using a Pre-Trained Language Model and Unified Medical Semantic Types

Siting Liang[*] and **Mareike Hartmann**[1] and **Daniel Sonntag**[*,2]
[*]German Research Center for Artificial Intelligence, Germany
[1]Saarland University, Germany
[2]University of Oldenburg, Germany
siting.liang|daniel.sonntag@dfki.de
mareikeh@coli.uni-saarland.de

Abstract

Information extraction from clinical text has the potential to facilitate clinical research and personalized clinical care, but annotating large amounts of data for each set of target tasks is prohibitive. We present a German medical Named Entity Recognition (NER) system capable of cross-domain knowledge transferring. The system builds on a pre-trained German language model and a token-level binary classifier, employing semantic types sourced from the Unified Medical Language System (UMLS) as entity labels to identify corresponding entity spans within the input text. To enhance the system's performance and robustness, we pre-train it using a medical literature corpus that incorporates UMLS semantic term annotations. We evaluate the system's effectiveness on two German annotated datasets obtained from different clinics in zero- and few-shot settings. The results show that our approach outperforms task-specific Condition Random Fields (CRF) classifiers in terms of accuracy. Our work contributes to developing robust and transparent German medical NER models that can support the extraction of information from various clinical texts.

1 Introduction

Information extraction from the large volume of unstructured text generated in hospitals and clinics has the potential to facilitate clinical research and enhance personalized clinical care. Especially the narrative notes, such as radiology reports, discharge summaries and clinical notes provide a more detailed and personalized history, assessments, medication and symptoms, offering a better context for clinical decision-making (Chen et al., 2015; Spasic et al., 2020).

In the field of Natural Language Processing (NLP), the problem of automatically and accurately extracting specific terms from text data is approached as a Named Entity Recognition (NER) task. NER methods ranging from rule-based to deep learning methods are the core technologies for automatically identifying medical instances from clinical narratives, such as diseases, diagnosis, drugs, and treatments (Sonntag et al., 2016; Sonntag and Profitlich, 2019; Miotto et al., 2018; Lerner et al., 2020; Wei et al., 2020; Kim and Meystre, 2020; Bose et al., 2021). Building clinical NER systems for non-English languages, e.g. German in our case, is challenging due to data scarcity. Only a few real-world annotated resources in German are publicly available (Starlinger et al., 2017; Kittner et al., 2021). This problem can be overcome by cross-domain transfer learning, where models transfer knowledge learned from data-rich relevant domains to domain-specific target tasks with less or no annotated data (Wang et al., 2019; Xie et al., 2018; Yuan et al., 2020; Plank, 2019; Artetxe et al., 2020; Lauscher et al., 2020).

We propose a simple but effective transfer learning framework based on a German BERT[1] encoder that is given a prompt consisting of a semantic type from UMLS semantic network[2], followed by a separator token and the medical text, e.g. *"[CLS]Clinical Drug[SEP]Zofran 4mg for nausea."*. On top of the encoder is a binary token classifier which predicts a probability for each token to determine whether it belongs to the given semantic type or not. Our approach, denoted as BERT-SNER (code[3]) and depicted in Figure 1, is based on three insights from recent research in transfer learning: i) Pre-trained Language Models (PLMs), e.g. BERT (Devlin et al., 2019), facilitate downstream tasks in specific domains (Lee et al., 2020; Alsentzer et al., 2019; Rasmy et al., 2021). ii) Prompting PLMs is becoming increasingly popular for solving low-resource NER tasks, as it can successfully exploit

[1]https://www.deepset.ai/german-bert
[2]https://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html
[3]https://github.com/sitingGZ/bert-sner.git

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 259–271
July 14, 2023 ©2023 Association for Computational Linguistics

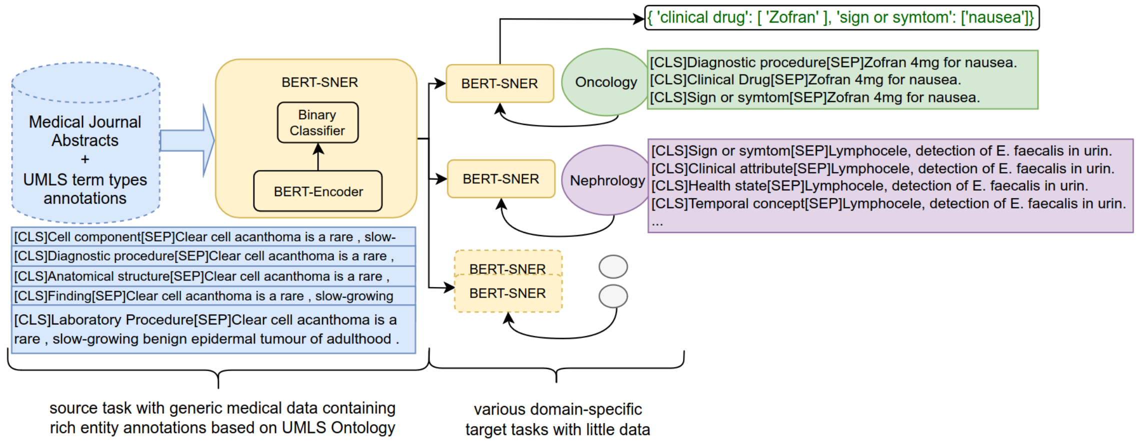


Figure 1: An overview of the transfer learning framework with BERT-SNER. We first train the model using a generic medical corpus with UMLS semantic term types as entity labels and further apply the model to different clinical domain-specific NER tasks with no or limited annotated training data.

generic knowledge learned in the pre-training tasks (Cui et al., 2021; Chen et al., 2021; Wang et al., 2022). iii) The Unified Medical Language System (UMLS) Metathesaurus (Bodenreider, 2004) is a useful knowledge source for mining medical terms in both biomedical and clinical documents (Aronson, 2001, 2006; Savova et al., 2010; Perez-Miguel et al., 2018; Kang et al., 2021; Michalopoulos et al., 2020).

The lack of domain-specific annotations is our motivation to develop models that can easily be adapted after pre-training on non-domain-specific annotated data. In our transfer learning experiments, we first derive training data from the open-source MUCHMORE corpus[4] to train BERT-SNER. MUCHMORE consists of German abstracts from 41 medical journals and entities are annotated with 134 UMLS semantic types (Archive 2001[5]). For more details on the annotation process of this corpus, please refer to Volk et al. (2002). After that, we map the entity labels of the two clinical target tasks to UMLS semantic types to be consistent with the annotations in the MUCHMORE corpus and perform zero- and few-shot experiments with 10, 50 and 100 shots for the two clinical target tasks.

The contributions of our work can be summarized as follows: 1) Our approach addresses low-resource German clinical NER tasks effectively. 2) We identify effective ways of transferring open-source medical knowledge for improving the performance of German clinical NER models.

[4]https://muchmore.dfki.de/resources1.htm
[5]https://lhncbc.nlm.nih.gov/semanticnetwork/SemanticNetworkArchive.html

2 Approach

Our approach explores the feasibility of knowledge transfer between different datasets by incorporating UMLS semantic term types to unify the entity labels. Table 1 shows how we construct training data from different domains to train BERT-SNER.

Input	Target
[CLS] Clinical Drug [SEP] Zofran 4mg for nausea	[0, 1, 1, 0, 1, 0, 0, 0]
[CLS] Sign or Symptom [SEP] Zofran 4mg for nausea	[0, 1, 1, 1, 0, 0, 0, 0, 1]
[CLS] Diagnostic Procedure [SEP] Zofran 4mg for nausea	[0, 0, 0, 0, 0, 0, 0, 0]

Table 1: Examples of training data (translated from German to English) using UMLS semantic types as entity labels. For each preceding entity label, if corresponding entity phrases (highlighted in orange) are found in the medical text, the tokens of the entity label and the entity phrases are annotated as class 1. The remaining part of the input is marked as class 0. If no entity phrase can be extracted for a given entity label (here *Diagnostic Procedure*), the entire target sequence contains only class 0 labels.

We compare the resulting NER system to a baseline architecture of BERT encoder combined with a task-specific conditional random fields (CRF) classifier (Wallach, 2004), i.e. BERT-CRF (Chaudhary et al., 2019; Souza et al., 2019; Pang et al., 2019; Liu et al., 2022; Mahendran and McInnes, 2021). In contrast to BERT-CRF models, BERT-SNER does not require the introduction of new task-specific parameters for solving the cross-domain target tasks, which benefits few-shot fine-tuning, while the BERT-CRF models fail if there are less than 100 samples in the target domain available.

Entity Type	Description	Semantic Type
DIAG	A disease, a symptom or a medical observation that can be matched with the German Modification of the International Classification of Diseases.	Sign or Symptom; Disease or Syndrome; Finding
TREAT	A diagnostic procedure, an operation or a systemic cancer treatment that can be found in the Operation.	Diagnostic Procedure; Therapeutic or Preventive Procedure
MED	A pharmaceutical substance or a drug that can be related to the Anatomical Therapeutic Chemical Classification System.	Pharmacologic Substance; Clinical Drug

Table 2: Original entity types and descriptions in BRONCO, and the best-matched selected semantic types from UMLS semantic network.

Entity Type	Description	Semantic Type
Condition	A pathological medical condition of a patient can describe for instance a symptom or a disease.	Sign or Symptom; Disease or Syndrome; Finding
DiagLab	Particular diagnostic procedures have been carried out.	Laboratory Procedure; Diagnostic Procedure
LabValues	Mentions of lab values.	Clinical Attribute
HealthState	A positive condition of the patient.	Health State*
Measure	Mostly numeric values, often in the context of medications or lab values, but can also be a description if a value changes, e.g. raises.	Quantitative Concept
Medication	A medication.	Pharmacologic Substance
Process	Describes particular process, such as blood pressure, or heart rate, often related to vital parameters.	Physiologic Function
TimeInfo	Describes temporal information, such as 2 weeks ago or January.	Temporal Concept

Table 3: Entity types, descriptions in Ex4CDS and the matched semantic types (*except for *HealthState*, where no proper semantic type is found and retained the natural words from the original entity type).

We use two datasets from different German clinical domains as target tasks: the Berlin-Tübingen-Oncology Corpus BRONCO (Kittner et al., 2021) and Ex4CDS (Roller et al., 2022). BRONCO consists of German discharge summaries for cancer patients annotated with medical entities of interest, such as *Medication (MED)*, *Diagnosis (DIAG)* and *Treatment (TREAT)*. Ex4CDS is a corpus of textual explanations for supporting system predictions of three possible outcomes (rejection, infection, graft failure) after kidney transplantation in the nephrology clinic. It focuses on entities that indicate the patient's *Health State* as well as *Laboratory Measures* after a Process. Table 7 presents the number of training samples and Table 9 presents the most frequent annotated semantic types in Appendix A and D. In order to achieve effective cross-domain transferability, we replace the original entity types of the target tasks with the best-matched UMLS semantic types during training. The matching to semantic types is determined by the ranking of the cosine similarity scores between the hidden representations of the entity types and the semantic types. The English descriptions of entity types are provided with the BRONCO and Ex4CDS datasets, and the hidden representations of type descriptions are obtained from the final hidden states of the encoder output from an English pre-trained language model[6]. The matched semantic types are validated by domain experts. Table 2 and Table 3 show the matched results. All English words of the selected semantic types are manually translated into German in our experiments.

3 Results and Discussion

The binary classifier of BERT-SNER predicts a probability for each token in the input sequence affected by the preceding semantic type and the sentence that follows. The classification result for each token is determined by setting a threshold. If the predicted probability is less than the threshold, the token is assigned to class 0, otherwise to class 1. The lower the threshold, the higher the false positive prediction rate, and conversely, a high threshold may result in a lower recall rate. We determine the threshold value for each entity label by finding the optimal precision-recall trade-off on the validation set of both target tasks based on the calculation results using the $sklearn.metrics.precision\_recall\_curve$ function. Figure 2 presents the range of thresholds in different shot settings. In the 10-shot case in both target tasks, the predicted probabilities of each token are smaller and the thresholds for individual entity types as a result are set lower. Figures in Appendix C show more details about the range of thresholds and Precision-Recall curves in different few-shot settings. In the case where a token is assigned multiple semantic types as several classi-

[6]microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext

	BRONCO				Ex4CDS			
shots	0	10	50	100	0	10	50	100
BERT-SNER	0.56±0.011	0.43±0.023	0.63±0.018	0.70±0.014	0.31±0.015	0.41±0.011	0.66±0.024	0.72±0.013
BERT-CRF	-	-	-	0.34±0.014	-	-	-	0.24±0.012
BERT-SNER(*)	-	0.26±0.020	0.33±0.013	0.36±0.024	-	0.18±0.012	0.27±0.024	0.38±0.016

Table 4: Macro-averaged F-scores of few-shot results on two target datasets. BERT-CRF is initialized with a 135-class classifier for the source task including the 134 semantic types adding an *OUT* (outside of the entity span) class, and is first pre-trained on MUCHMORE. Then, the encoder of BERT-CRF is further fine-tuned with domain-specific classifiers for BRONCO and Ex4CDS when switching domains and datasets. BERT-SNER(*) is our proposed NER framework without pre-training on MUCHMORE, i.e. trained only on data of each target task. '-' indicates a classification failure with an F-score < 0.1. '±' indicates the variance in scores caused by 2 different seeds, 3 times of random sampling and selection of semantic types in cases with multiple best-fit semantic types for individual entity types in each target task.

	BRONCO			
	0	10	50	100
MED	0.54±0.03	0.21±0.05	0.71±0.01	0.81±0.03
TREAT	0.31±0.01	0.22±0.03	0.39±0.03	0.43±0.03
DIAG	0.48 ±0.02	0.42±0.02	0.45±0.03	0.56±0.03

Table 5: F-scores of individual entity type for BRONCO test data and the BERT-SNER model with optimal thresholds in different settings.

	Ex4CDS			
	0	10	50	100
Condition	0.30±0.03	0.50±0.03	0.67±0.01	0.72±0.03
DiagLab	0.43±0.04	0.65±0.01	0.73±0.05	0.81±0.02
LabValues	0.20 ±0.03	0.64±0.02	0.78±0.03	0.88±0.01
HealthState	0.31±0.04	0.40±0.02	0.86±0.02	0.90±0.02
Measure	0.20±0.02	0.24±0.03	0.62±0.01	0.66±0.03
Medication	0.14 ±0.02	0.22±0.02	0.22±0.03	0.22±0.01
Process	0.19±0.02	0.24±0.01	0.78±0.01	0.83±0.03
TimeInfo	0.16 ±0.02	0.16±0.02	0.41±0.01	0.60±0.02

Table 6: F-scores of individual entity type for Ex4CDS test data and the BERT-SNER model with optimal thresholds in different few-shot settings.

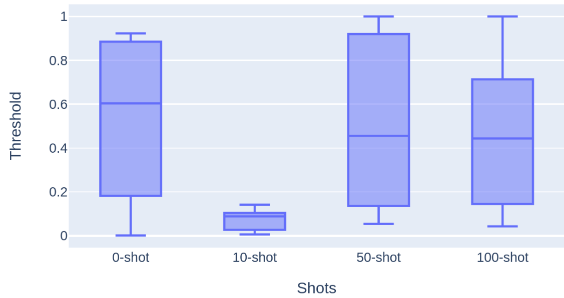


Figure 2: Ranges of thresholds by finding the best precision-recall trade-off on validation datasets. In the case of 10-shot, the prediction scores for each token in both target tasks are low, and therefore the thresholds are found lower compared to the other settings.

fication probabilities exceed the threshold, we rank the semantic types assigned to the token by their probabilities and retain the first type as the final classification result. Figure 3 in Appendix B provides an interpretation of the token-level prediction using BERT-SNER for an input sentence preceded by various semantic types.

Table 4 presents macro-averaged F-scores for BERT-SNER and baseline BERT-CRF on the two target datasets for different numbers of shots for fine-tuning the models. BERT-SNER first trained on MUCHMORE performs much better than the BERT-CRF models trained with the same resource in few-shot settings. Even without additional source data, BERT-SNER(*) shows comparable or better performance than BERT-CRF in both clinical domains. Applying the CRF classifier of the source task directly to the target tasks in the BERT-CRF framework shows worse performance than resetting the classifier with a specific label set on different target tasks.

Tables 5 and 6 present F-scores per individual entity types. When comparing the results between zero- and few-shot, we find some semantic types that can not be generalized well to the target domains, such as *(TREAT -> Diagnostic Procedure)* in BRONCO vs *(DiagLab -> Diagnostic)* in Ex4CDS, and *(MED -> Clinical Drug)* in BRONCO vs *(Medication -> Clinical Drug)* in Ex4CDS. In contrast, domain-specific entity types, *HealthState*, *LabValues* and *Process*, which are unseen or infrequent in the source task, can benefit the most from the increasing number of shots in the self-domain. These results suggest that in future work, there is a need to investigate more deeply the semantic differences of domain-specific entities matched to the same unified semantic type when experimenting with the BERT-SNER system for more diverse clinical domains. In addition, we need to examine more the impact of the amount of training data from the MUCHMORE corpus on individual entity types in target tasks.

4 Related Work

Our work focuses on solving low-resource NER tasks in the clinical domain leveraging additional resources from related domains, and in non-English languages. A common solution is to perform downstream tasks for non-English languages, especially typologically close to English through cross-lingual transfer from large-scale pre-trained multilingual BERT models (Lauscher et al., 2020; Souza et al., 2019; Jørgensen et al., 2021; Hakala and Pyysalo, 2019; Souza et al., 2019) or English language models (Artetxe et al., 2020; Plank, 2019). Frei and Kramer (2022) and Schäfer et al. (2022) attempt to use synthesised data through translation from English resources (Henry et al., 2019) to train a German medical NER model. Most of the previous works in this field have focused on a single task and it's unclear if these task-specific approaches can easily be extended to other clinical datasets with different label sets.

Sequence-to-Sequence (Seq2Seq) PLMs with prompt-based methods in another line have been shown to be useful for solving low-resource NER problems (Han et al., 2021; Gao et al., 2020; Cui et al., 2021; Yan et al., 2021; Chen et al., 2021; Wang et al., 2022). Other previous work of this line (Cui et al., 2021; Chen et al., 2021; Wang et al., 2022) utilized NER data from a resource-rich domain to fine-tune the Seq2Seq models on NER tasks before applying them to low-resource NER tasks. Although no new parameters are introduced to the pre-trained Seq2Seq language model when formulating the NER tasks in a generative framework, these methods require much effort for finding the optimal prompts and framework to transform an input sequence of tokens (words or characters) into an output sequence of entity labels. Unlike Seq2Seq NER models, our BERT-SNER model uses semantic types as prompts in front of the input directly, and the binary classifier is more efficient in terms of computational requirements, inference time and post-processing needs.

5 Conclusion

Our results suggest that transferring knowledge from publicly available medical resources with BERT-SNER is more effective than with BERT-CRF in low-resource scenarios. The overall benefit of the BERT-SNER in real-world use cases is that it can be used as an initial model to effectively develop domain-specific models in a variety of clinical applications, as it requires much less fine-tuning data than training a NER model from scratch. In future work, we will explore transfer learning more to generalize BERT-SNER to more different clinical NER tasks in low-resource situations. To apply BERT-SNER to new clinical applications without annotated samples, we will use active learning strategies such as Least Confidence oracle (Settles and Craven, 2008) to query the most informative samples to obtain annotations for fine-tuning.

Limitations

Due to strict data protection regulations and a high annotation workload in the clinical domain, obtaining more diverse target tasks to validate our approach is a challenge. In this work, we focused on only two use cases in German clinical applications and need to extend our experiments to English or other non-English languages in the field. In addition, we need to conduct more experiments in future work in order to achieve a better balance between the amount of training data required for the source and target tasks.

Acknowledgements

This work is funded by the pAItient project (BMG, 2520DAT0P2).

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Alan R Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. page 17. American Medical Informatics Association.

Alan R Aronson. 2006. Metamap: Mapping text to the umls metathesaurus. *Bethesda, MD: NLM, NIH, DHHS*, 1:26.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Priyankar Bose, Sriram Srinivasan, William C Sleeman IV, Jatinder Palta, Rishabh Kapoor, and Preetam Ghosh. 2021. A survey on recent named entity recognition and relationship extraction techniques on clinical texts. *Applied Sciences*, 11(18):8319.

Aditi Chaudhary, Jiateng Xie, Zaid Sheikh, Graham Neubig, and Jaime Carbonell. 2019. A little annotation does a lot of good: A study in bootstrapping low-resource named entity recognizers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5164–5174, Hong Kong, China. Association for Computational Linguistics.

Xiang Chen, Ningyu Zhang, Lei Li, Xin Xie, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2021. Lightner: A lightweight generative framework with prompt-guided attention for low-resource ner. *arXiv preprint arXiv:2109.00720*.

Yukun Chen, Thomas A Lasko, Qiaozhu Mei, Joshua C Denny, and Hua Xu. 2015. A study of active learning methods for named entity recognition in clinical text. *Journal of biomedical informatics*, 58:11–18.

Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. 2021. Template-based named entity recognition using BART. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1835–1845.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Johann Frei and Frank Kramer. 2022. GERNERMED: An open German medical NER model. *Software Impacts*, 11:100212.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Kai Hakala and Sampo Pyysalo. 2019. Biomedical named entity recognition with multilingual BERT. In *Proceedings of the 5th workshop on BioNLP open shared tasks*, pages 56–61.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, et al. 2021. Pre-trained Models: Past, Present and Future. *AI Open*, 2:225–250.

Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2019. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

Rasmus Kær Jørgensen, Mareike Hartmann, Xiang Dai, and Desmond Elliott. 2021. mDAPT: Multilingual Domain Adaptive Pretraining in a Single Model. *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3404–3418.

Tian Kang, Adler Perotte, Youlan Tang, Casey Ta, and Chunhua Weng. 2021. UMLS-based data augmentation for natural language processing of clinical research literature. *Journal of the American Medical Informatics Association*, 28(4):812–823.

Youngjun Kim and Stéphane M Meystre. 2020. Ensemble method–based extraction of medication and related information from clinical texts. *Journal of the American Medical Informatics Association*, 27(1):31–38.

Madeleine Kittner, Mario Lamping, Damian T Rieke, Julian Götze, Bariya Bajwa, Ivan Jelas, Gina Rüter, Hanjo Hautow, Mario Sänger, Maryam Habibi, et al. 2021. Annotation and initial evaluation of a large annotated german oncological corpus. *JAMIA open*, 4(2):ooab025.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers. *arXiv preprint arXiv:2005.00633*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Ivan Lerner, Jordan Jouffroy, Anita Burgun, and Antoine Neuraz. 2020. Learning the grammar of prescription: recurrent neural network grammars for medication information extraction in clinical texts. *arXiv preprint arXiv:2004.11622*.

Mingyi Liu, Zhiying Tu, Tong Zhang, Tonghua Su, Xiaofei Xu, and Zhongjie Wang. 2022. LTP: a new active learning strategy for CRF-based named entity recognition. *Neural Processing Letters*, pages 1–22.

Darshini Mahendran and Bridget T McInnes. 2021. Extracting adverse drug events from clinical notes. *AMIA Summits on Translational Science Proceedings*, 2021:420.

George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen H. Chen, and Alexander Wong. 2020. UmlsBERT: Clinical domain knowledge augmentation of contextual embeddings using the unified medical language system metathesaurus. *CoRR*, abs/2010.10391.

Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. 2018. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246.

Na Pang, Li Qian, Weimin Lyu, and Jin-Dong Yang. 2019. Transfer learning for scientific data chain extraction in small chemical corpus with BERT-CRF model. *arXiv preprint arXiv:1905.05615*.

Naiara Perez-Miguel, Montse Cuadros, and German Rigau. 2018. Biomedical term normalization of ehrs with umls. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Barbara Plank. 2019. Neural cross-lingual transfer and limited annotated data for named entity recognition in Danish. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 370–375, Turku, Finland. Linköping University Electronic Press.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):1–13.

Roland Roller, Aljoscha Burchardt, Nils Feldhus, Laura Seiffe, Klemens Budde, Simon Ronicke, and Bilgin Osmanodja. 2022. An annotated corpus of textual explanations for clinical decision support. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, page 2317–2326.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Henning Schäfer, Ahmad Idrissi-Yaghir, Peter Horn, and Christoph Friedrich. 2022. Cross-language transfer of high-quality annotations: Combining neural machine translation with cross-linguistic span alignment to apply ner to clinical texts in a low-resource language. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 53–62.

Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 1070–1079.

Daniel Sonntag and Hans-Jürgen Profitlich. 2019. An architecture of open-source tools to combine textual information extraction, faceted search and information visualisation. *Artificial intelligence in medicine*, 93:13–28.

Daniel Sonntag, Volker Tresp, Sonja Zillner, Alexander Cavallaro, Matthias Hammon, André Reis, Peter A Fasching, Martin Sedlmayr, Thomas Ganslandt, Hans-Ulrich Prokosch, et al. 2016. The clinical data intelligence project. *Informatik-Spektrum*, 39(4):290–300.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2019. Portuguese named entity recognition using bert-crf. *arXiv e-prints*, pages arXiv–1909.

Irena Spasic, Goran Nenadic, et al. 2020. Clinical text data in machine learning: systematic review. *JMIR medical informatics*, 8(3):e17984.

Johannes Starlinger, Madeleine Kittner, Oliver Blankenstein, and Ulf Leser. 2017. How to improve information extraction from german medical records. *it - Information Technology*, 59(4):171–179.

Martin Volk, Bärbel Ripplinger, Špela Vintar, Paul Buitelaar, Diana Raileanu, and Bogdan Sacaleanu. 2002. Semantic annotation for concept-based cross-language medical information retrieval. *International Journal of Medical Informatics*, 67(1-3):97–112.

Hanna M Wallach. 2004. Conditional Random Fields: An introduction. *Technical Reports (CIS)*, page 22.

Liwen Wang, Rumei Li, Yang Yan, Yuanmeng Yan, Sirui Wang, Wei Wu, and Weiran Xu. 2022. Instructionner: A multi-task instruction-based generative framework for few-shot ner. *arXiv preprint arXiv:2203.03903*.

Zihan Wang, Stephen Mayhew, Dan Roth, et al. 2019. Cross-lingual ability of Multilingual BERT: An empirical study. *arXiv e-prints*, pages arXiv–1912.

Qiang Wei, Zongcheng Ji, Zhiheng Li, Jingcheng Du, Jingqi Wang, Jun Xu, Yang Xiang, Firat Tiryaki, Stephen Wu, Yaoyun Zhang, et al. 2020. A study of deep learning approaches for medication and adverse drug event extraction from clinical text. *Journal of the American Medical Informatics Association*, 27(1):13–21.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime G. Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. *CoRR*, abs/1808.09861.

Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, and Xipeng Qiu. 2021. A Unified Generative Framework for Various NER Subtasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5808–5822.

Michelle Yuan, Mozhi Zhang, Benjamin Van Durme, Leah Findlater, and Jordan Boyd-Graber. 2020. Interactive Refinement of Cross-Lingual Word Embeddings. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5984–5996.

A Data statistic

	train	valid	test
MUCHMORE	10000	4000	-
BRONCO	100	100	100
Ex4CDS	100	100	100

Table 7: Number of sentences in datasets used for training, where MUCHMORE is a training source annotated with UMLS semantic types. Training data from BRONCO and Ex4CDS are limited to maximum 100 samples in each subset in few-shot experiments.

shots	10	50	100	test
MED	4	8	17	17
TREAT	3	17	40	57
DIAG	5	31	62	101
Condition	20	95	189	163
DiagLab	3	8	17	11
LabValues	8	28	60	78
HealthState	7	35	65	69
Measure	7	33	65	97
Medication	2	17	23	8
Process	7	25	44	60
TimeInfo	15	63	102	48

Table 8: Average number of annotated tokens of individual entity types from both target tasks in different few-shot samplings and test sets.

B Interpretation of the Model Outcome

In our proposed NER framework, each input sentence is iterated once with a semantic type by the BERT-SNER model. The matched semantic types based on the entity types of each task are shown in Table 2 and 3. Given an example *"stabile Funktion, keine Protenurie noch nie NTX-Versagen"* (In English: stable function, no proteinuria not ever NTX failure) from Ex4CDS, it is tokenized as ['stabile', 'Funktion,', 'keine', 'Protenurie,', 'noch', 'nie', 'NTX-Versagen'] and contains the following token-level entity annotation: ['HealthState', 'Process', 'O', 'LabValues', 'O', 'O', 'Condition'] from the original entity type set.

Predictions of the BERT-SNER model are made by a binary classifier, which are probabilities in the range of (0, 1). The scores predicted for the tokens of the semantic types are depending on the text input. The predicted probabilities for each token in an input sentence are affected by the semantic type in front. Figure 3 illustrates that the salience variation of each token in the same input sentence is influenced by the preceding semantic type. As a result, the final probability of each token of the input sentence is multiplied by the probability score of the first token of the given semantic type. We need to rank the scores across the applied semantic types and set a threshold to determine the final entity class for each token. In the following section, we show how to find the optimal threshold ranges to allocate the classification to each token in different few-shot settings based on the final probability scores.

C Precision-Recall Trade-off and Finding the Optimal Thresholds

The thresholds are used to determine the final classification result of a binary classifier. If the probability values are less than the threshold, assigned to class 0, while values greater than or equal to the threshold are assigned to class 1. In order to find the optimal threshold ranges in different few-shot settings, we explore the Prediction-Recall Curves and the correlations between the thresholds and F-scores according to entity types and trained shots. We can find similar phenomena in both Ex4CDS and BRONCO data, as shown in Figures 4-11.

D Most frequent annotated UMLS semantic types

134 semantic types from UMLS semantic network ontology in 2001 are annotated in MUCHMORE corpora. However, the number of annotations of each semantic type is extremely imbalanced ranging from less than 10 terms to at most 8202. We show the most frequent annotated semantic types in Table 9.

[CLS] Zeichen oder Symptom [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Diagnostisches Verfahren und Laborverfahren [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Klinisches Attribut [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Gesunder Zustand [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Quantitatives Konzept [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Klinisches Arzneimittel [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Physiologische Funktion [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Zeitliches Konzept [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen

(a) zero-shot

[CLS] Zeichen oder Symptom [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Diagnostisches Verfahren und Laborverfahren [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Klinisches Attribut [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Gesunder Zustand [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Quantitatives Konzept [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Klinisches Arzneimittel [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Physiologische Funktion [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Zeitliches Konzept [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen

(b) 10-shot

[CLS] Zeichen oder Symptom [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Diagnostisches Verfahren und Laborverfahren [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Klinisches Attribut [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Gesunder Zustand [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Quantitatives Konzept [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Klinisches Arzneimittel [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Physiologische Funktion [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Zeitliches Konzept [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen

(c) 50-shot

[CLS] Zeichen oder Symptom [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Diagnostisches Verfahren und Laborverfahren [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Klinisches Attribut [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Gesunder Zustand [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Quantitatives Konzept [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Klinisches Arzneimittel [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Physiologische Funktion [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen
[CLS] Zeitliches Konzept [SEP] stabile Funktion, keine Protenurie, noch nie NTX-Versagen

(d) 100-shot

Figure 3: Predicted outcomes of zero-shot or few-shot fine-tuning for the example sentence from Ex4CDS dataset corresponding to various preceding semantic types. These eight semantic types (translated into German words) are used to replace the eight entity types during fine-tuning and inference in the target task with BERT-SNER. The color intensity indicates the value of the prediction score; the darker the color, the higher the value.

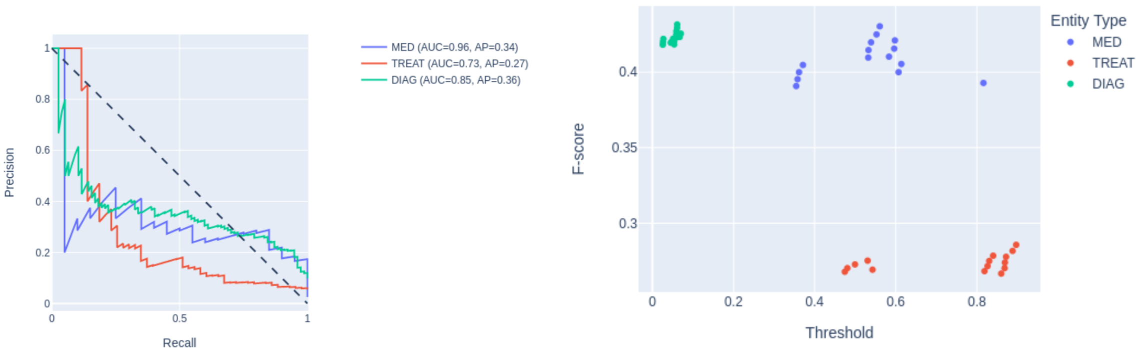


Figure 4: Zero-shot with BRONCO data. Domain-shift presents in types *TREAT* and *DIAG*. The optimal thresholds of each entity types lie in different ranges.

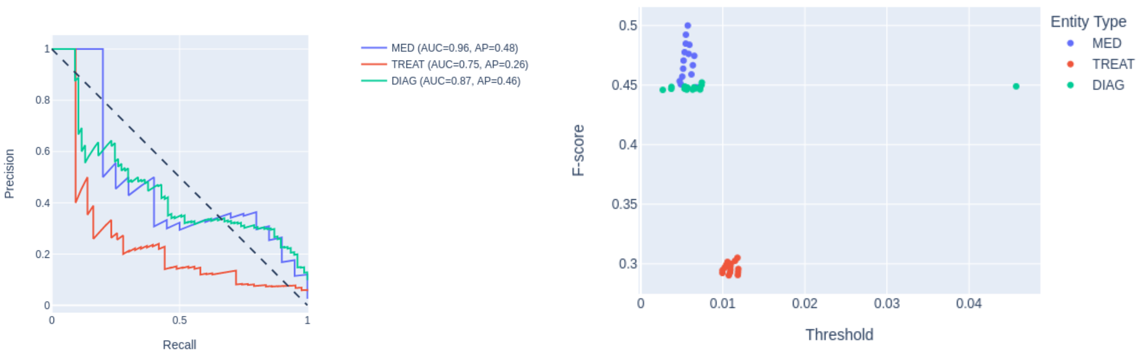


Figure 5: 10-shot with BRONCO data. The optimal thresholds for best F-scores are lowered as the BERT-SNER model has been fine-tuned with 10 samples from the target task compared to zero-shot.

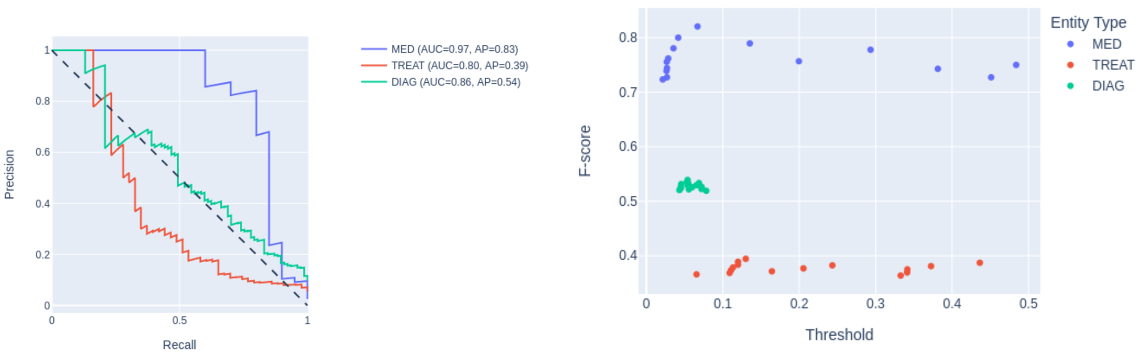


Figure 6: 50-shot with BRONCO data. The AUC scores are improved after fine-tuning with 50 samples from the target task. The optimal thresholds for best F-scores are increased compared to 10-shot fine-tuning.

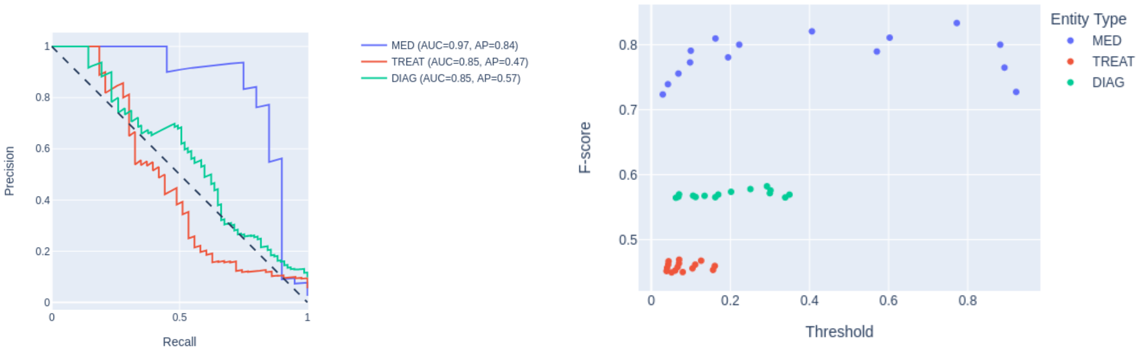


Figure 7: 100-shot with BRONCO data. The optimal thresholds for obtaining the best F-scores are increased for *MED* and *DIAG* types as the BERT-SNER model has been fine-tuned with 100 samples from the target task. From the results of F-scores and AUC scores, we find that identifying the entities of type *TREAT* in BRONCO task is a challenge for BERT-SNER.

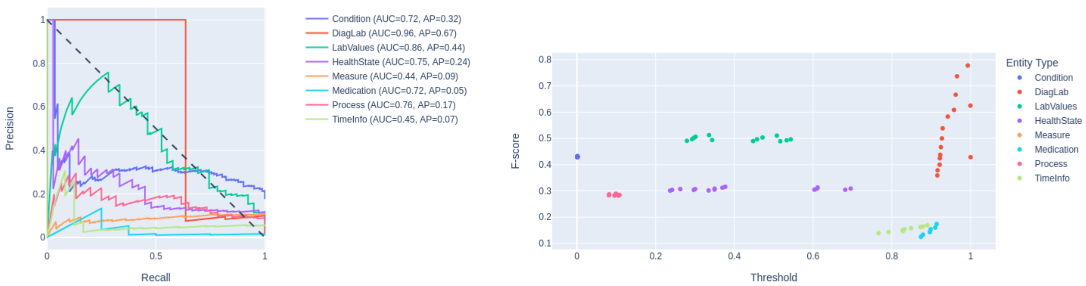


Figure 8: Zero-shot with Ex4CDS data.

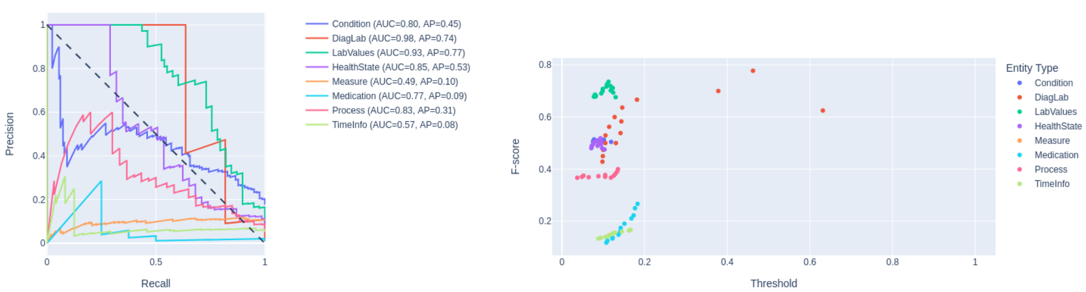


Figure 9: 10-shot with Ex4CDS data.

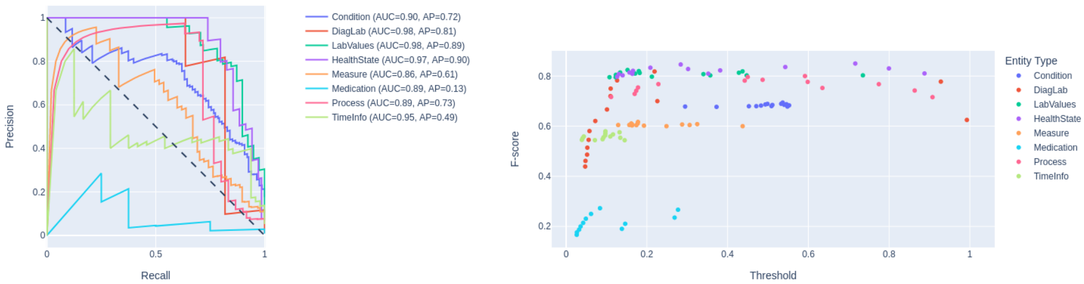


Figure 10: 50-shot with Ex4CDS data.

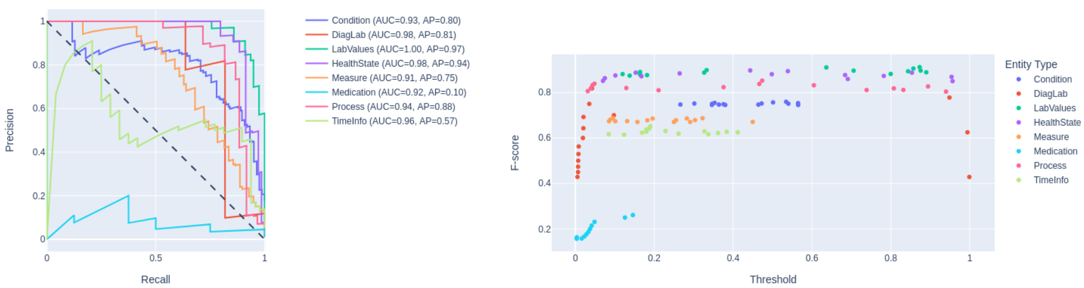


Figure 11: 100-shot with Ex4CDS data. The most challenging type for BERT-SNER in Ex4CDS is *Medication*.

ID	Type Name	Description	Amount
T101	Patient or Disabled Group	An individual or individuals classified according to a disability, disease, condition or treatment.	8202
T047	Disease or Syndrome	A condition which alters or interferes with a normal process, state, or activity of an organism. It is usually characterized by the abnormal functioning of one or more of the host's systems, parts, or organs. Included here is a complex of symptoms descriptive of a disorder.	7636
T023	Body Part, Organ, or Organ Component	A collection of cells and tissues which are localized to a specific area or combine and carry out one or more specialized functions of an organism. This ranges from gross structures to small components of complex organs. These structures are relatively localized in comparison to tissues.	7070
T169	Functional Concept	A concept which is of interest because it pertains to the carrying out of a process or activity.	5569
T061	Therapeutic or Preventive Procedure	A procedure, method, or technique designed to prevent a disease or a disorder, or to improve physical function, or used in the process of treating a disease or injury.	5542
T046	Pathologic Function	A disordered process, activity, or state of the organism as a whole, of a body system or systems, or of multiple organs or tissues. Included here are normal responses to a negative stimulus as well as pathololologic conditions or states that are less specific than a disease. Pathologic functions frequently have systemic effects.	3974
T191	Neoplastic Process	A new and abnormal growth of tissue in which the growth is uncontrolled and progressive. The growths may be malignant or benign.	3806
T170	Intellectual Product	A conceptual entity resulting from human endeavor. Concepts assigned to this type generally refer to information created by humans for some purpose.	3266
T081	Quantitative Concept	A concept which involves the dimensions, quantity or capacity of something using some unit of measure, or which involves the quantitative comparison of entities.	3049
T033	Finding	That which is discovered by direct observation or measurement of an organism attribute or condition, including the clinical history of the patient. The history of the presence of a disease is a 'Finding' and is distinguished from the disease itself.	2621
T060	Diagnostic Procedure	A procedure, method, or technique used to determine the nature or identity of a disease or disorder. This excludes procedures which are primarily carried out on specimens in a laboratory.	2621
T184	Sign or Symptom	An observable manifestation of a disease or condition based on clinical judgment, or a manifestation of a disease or condition which is experienced by the patient and reported as a subjective observation.	2547
T024	Tissue	An aggregation of similarly specialized cells and the associated intercellular substance. Tissues are relatively non-localized in comparison to body parts, organs or organ components.	2533
T121	Pharmacologic Substance	A substance used in the treatment or prevention of pathologic disorders. This includes substances that occur naturally in the body and are administered therapeutically.	2403
T037	Injury or Poisoning	A traumatic wound, injury, or poisoning caused by an external agent or force.	2080
T029	Body Location or Region	An area, subdivision, or region of the body demarcated for the purpose of topographical description.	1865
T040	Organism Function	A physiologic function of the organism as a whole, of multiple organ systems, or of multiple organs or tissues.	1540
T041	Mental Process	A physiologic function involving the mind or cognitive processing.	1429
T078	Idea or Concept	An abstract concept, such as a social, religious or philosophical concept.	1309
T032	Organism Attribute	A property of the organism or its major parts.	1281
T073	Manufactured Object	A physical object made by human beings.	1226
T091	Biomedical Occupation or Discipline	A vocation, academic discipline, or field of study related to biomedicine.	1213
T123	Biologically Active Substance	A generally endogenous substance produced or required by an organism, of primary interest because of its role in the biologic functioning of the organism that produces it.	1187
T100	Age Group	An individual or individuals classified according to their age.	1149
T062	Research Activity	An activity carried out as part of research or experimentation.	1148
T079	Temporal Concept	A concept which pertains to time or duration.	1124

Table 9: Most frequent UMLS semantic types annotated in the MUCHMORE data. The numbers in the third column are the amount of annotated terms of the semantic type.

Reducing Knowledge Noise for Improved Semantic Analysis in Biomedical Natural Language Processing Applications

Usman Naseem[1], Surendrabikram Thapa[2], Qi Zhang[3,4], Liang Hu[3,4], Anum Masood[5], Mehwish Nasim[6,7]
[1]University of Sydney, Australia [2]Virginia Tech, USA
[3]Tongji University, China [4]DeepBlue Academy of Sciences, China
[5]Norwegian University of Science and Technology, Norway
[6]University of Western Australia [7]Flinders University, Australia

Abstract

Graph-based techniques have gained traction for representing and analyzing data in various natural language processing (NLP) tasks. Knowledge graph-based language representation models have shown promising results in leveraging domain-specific knowledge for NLP tasks, particularly in the biomedical NLP field. However, such models have limitations, including knowledge noise and neglect of contextual relationships, leading to potential semantic errors and reduced accuracy. To address these issues, this paper proposes two novel methods. The *first method* combines knowledge graph-based language model with nearest-neighbor models to incorporate semantic and category information from neighboring instances. The *second method* involves integrating knowledge graph-based language model with graph neural networks (GNNs) to leverage feature information from neighboring nodes in the graph. Experiments on relation extraction (RE) and classification tasks in English and Chinese language datasets demonstrate significant performance improvements with both methods, highlighting their potential for enhancing the performance of language models and improving NLP applications in the biomedical domain.

1 Introduction

Language models (LM) have become increasingly popular in a wide range of applications (Adhikari et al., 2023; Shah et al., 2023). LMs have also shown great promise in assisting clinical decision-making. LMs for clinical decision-making are based on massive amounts of text data, including medical literature, clinical notes, and patient records (Lewis et al., 2020; Adhikari et al., 2021). By analyzing medical records and identifying patterns, language models can help doctors identify potential health risks and recommend appropriate treatment options (Kalyan et al., 2022; Naseem et al., 2022a). Additionally, LMs can be used to assist with drug discovery by analyzing vast amounts of scientific literature and identifying potential drug targets (Naseem et al., 2022b).

LMs can also improve patient outcomes by assisting with patient education. LMs can analyze patient records and suggest personalized educational resources, such as videos or articles, to help patients better understand their conditions and treatments. Likewise, language models can help clinicians communicate more effectively with patients by providing real-time translation services for patients who speak different languages. However, these language models are not without their shortcomings. For example, the cross-entropy loss used in fine-tuning language models can lead to poor generalization of performance, as pointed out by Liu et al. (2016). Similarly, LMs can be prone to overfitting as the predictions are made through linear classifiers added directly to the top of pretrained LMs (Li et al., 2021). Moreover, they may neglect the relationship between textual contexts, impacting performance. Given the importance of the medical field, errors in language models can have significant consequences. Thus, efforts to improve the performance of language models are being made. A possible scope for improvement can be using knowledge from nearest neighbors.

Khandelwal et al. (2019) employed a k-nearest neighbor (kNN) approach to enhance the performance of LMs. The kNN-LM approach uses nearest-neighbor models to improve language modeling by explicitly memorizing rare patterns and improving performance, indicating that the representation learning problem is easier than the prediction problem. Similarly, Kassner and Schütze (2020) approached an open-domain question-answering problem with kNN-BERT, a combination of BERT's prediction for a given question with a kNN search. The authors demonstrated through their experimental results and evaluation that incorporating kNN into the BERT-based question-answering model was effective in retriev-

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 272–277
July 14, 2023 ©2023 Association for Computational Linguistics

ing accurate factual information. This model particularly excelled in providing answers to less frequent and difficult questions and was able to handle even recent events that were not included in the training data of the BERT model. Unlike other studies that use kNN to generate augmented samples based on pre-trained language models, Li et al. (2021) utilized a kNN classifier as the decision maker. It was demonstrated that incorporating kNNs with traditional fine-tuning of BERT-like models can significantly improve accuracy in both rich-source and few-shot settings and improve robustness against adversarial attacks. Nearest neighbors have shown immense success in improving the interpretability of models as well (Wallace et al., 2018). The use of kNN in explaining the model behavior in language models is a topic that has been attracting the research community (Rajani et al., 2020). Interpretable machine learning is key to building trustworthiness in AI systems used in healthcare. Interpretability can provide clinicians and patients with insights into the reasoning behind a particular decision made by a machine learning model, making it easier to understand and trust. This can help to improve patient outcomes and increase the acceptance of AI systems in healthcare. Additionally, interpretability can help identify and address biases in machine learning models.

While LMs like BERT (Devlin et al., 2018) have limited ability to capture global information, Graph Neural Networks (GNNs) have proved to be better at it (Wu et al., 2023). Language models like BERT are good at capturing contextual information. In order to make the best out of a language model, it needs to have global information as well as proper contextual information. Utilizing the strengths of both graph neural networks and language models, Lu et al. (2020) proposed Vocabulary Graph Convolutional Network (VGCN)-BERT. The motivation for combining VGCN and BERT is to allow them to collectively build an optimal representation while performing tasks such as classification. VGCN models the relationships between words in a text by representing them as nodes in a graph, while BERT is a state-of-the-art language model that can effectively capture the contextual information of text data. This enhanced representation led to improved performance in text classification tasks, as demonstrated by the results of the study.

Global knowledge is very important in the medical domain because medical data is inherently complex and interconnected. In order to effectively analyze medical data, it is essential to capture the relationships between different medical concepts and understand their context within the larger medical knowledge network (Rasmy et al., 2021). For example, in the field of medical diagnosis, a patient's symptoms and medical history need to be considered in the context of the larger medical knowledge network to accurately diagnose their condition (Lin et al., 2021). By leveraging global knowledge, medical professionals can better understand the relationships between different medical concepts and make more well-informed decisions regarding patient care. Furthermore, medical research often involves analyzing large datasets containing vast medical information (Naseem et al., 2021a). By utilizing global knowledge, researchers can better identify patterns and relationships within the data, leading to more accurate and insightful findings. Thus, there is a pressing need to build representation comprising contextual and global information.

Pre-trained language models are mostly generic in nature and lack domain-specific knowledge (Liu et al., 2023). This can be a problem for tasks that require access to domain-specific knowledge, such as medical text classification and medical relation extraction (Naseem et al., 2021b). To mitigate this problem, Liu et al. (2020) proposed K-BERT, a knowledge-enabled language representation model that addresses this problem by incorporating knowledge from a knowledge graph into its language representation. This knowledge incorporation allows K-BERT to perform better on tasks that require access to domain-specific knowledge. K-BERT allows the triplets to be injected into sentences as knowledge, making it useful for domain-specific tasks. Thus, leveraging K-BERT for domain knowledge and kNN and GNN for global knowledge, we present methods for integrating (i) K-BERT with GNN and (ii) K-BERT with kNN. Our contributions are as follows:

- We present a method to improve the performance of knowledge graph-based language models by integrating semantic information from neighboring instances.
- We demonstrate significant performance improvements in relation extraction (RE) and classification tasks using our methods, showcasing their potential for improving NLP applications in the biomedical domain.

2 Methodology

As discussed above, our aim is to utilize the information of the neighbors in the dataset to help improve the performance of the language model in various downstream tasks. We use K-BERT as a language model because of its flexibility to adapt to domain knowledge.

2.1 Nearest Neighbor Enriched Language Model

Our nearest neighbor algorithm, kNN, extracts instances similar to the test samples from the feature library and uses the information of the instances to assist in prediction. kNN uses the classification label information of the neighbors. In order to combine K-BERT with kNN, the K-BERT model is first trained on the training data. The K-BERT model is then evaluated on the validation and test sets. During the prediction phase, kNN finds the k most similar instances in the training set to a given test sample. The classification labels of these instances are then used to predict the label of the test sample. The K-BERT model is also used to predict the given test sample.

The final prediction is obtained by combining the results of the kNN and K-Bert models using a weighted sum. This is done by applying a sigmoid function to the individual scores predicted by each model and taking their weighted average. The category with the highest score is chosen as the final prediction.

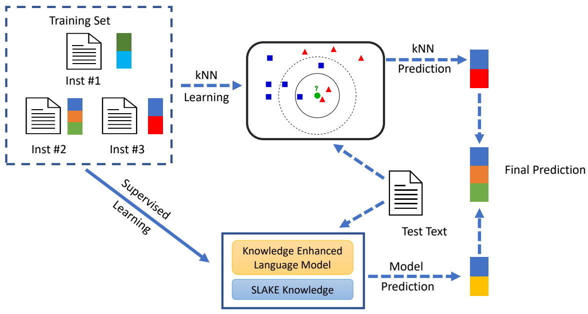


Figure 1: The overall flow of obtaining predictions through the information from nearest neighbors and language model.

2.2 Graph-NN Enriched Language Model

The integration of Graph Neural Networks (GNNs) with the feature level of the data has been shown to be effective in utilizing the feature information of the neighboring nodes. One of the widely used GNN networks is Graph Attention Network (GAT). In the early stages of data processing, a knowledge graph is added to the dataset and GAT is employed to construct a graph between nodes. The construction of this graph is achieved through the utilization of heuristic rules where edges are added to instances that belong to the same label category.

During the training process, the representation of each node obtained using K-BERT and GAT is subsequently employed to aggregate the representations of the neighboring nodes. The classification result of each instance is then predicted using GAT. In the prediction process, the graph is composed by obtaining the result through K-BERT and then using the result to connect nodes that belong to the same category. Finally, GAT is employed to aggregate the representations of the neighboring nodes to predict the classification result.

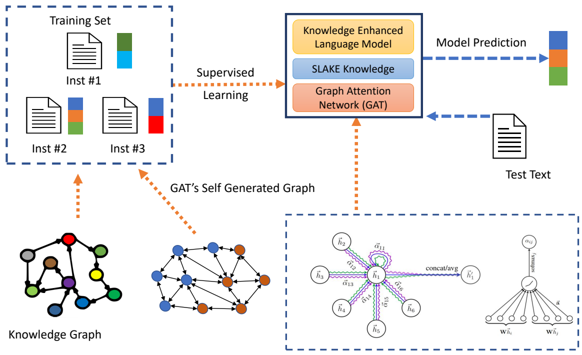


Figure 2: The overall methodology of obtaining prediction using an aggregate representation of language model and neighboring nodes.

2.3 Domain-specific Medical Knowledge

In order to add medical domain-specific knowledge to K-BERT, the triples are injected into sentences as domain knowledge. We leverage the SLAKE (Liu et al., 2021) to get triplets to be injected in K-BERT. SLAKE, a large multilingual dataset, contains rich ensemble semantic labels and a new structural medical knowledge base. For the relation extraction task, we chose SLAKE in the English language, whereas for the classification tasks, we used SLAKE in the Chinese language.

2.4 Datasets

In the experimentation, we have tested the performance of K-BERT + kNN and K-BERT + GNN in two downstream tasks–relation extraction and classification tasks.

Relation Extraction Datasets: For relation extraction task, we use GAD (Bravo et al., 2015), EU-ADR (Van Mulligen et al., 2012), and i2b2 (Uzuner et al., 2011) datasets. The Genetic Association Database (GAD) is a collection of studies that examine the association between genetic variations and the risk of developing complex diseases and disorders in humans (Bravo et al., 2015). Similarly, the EU-ADR corpus has been annotated for drugs, disorders, genes, and their interrelationships (Van Mulligen et al., 2012). i2b2 dataset is a dataset of patient medical problems, treatments, and tests used in the 2010 i2b2/VA Workshop on Natural Language Processing Challenges for Clinical Records (Uzuner et al., 2011).

Classification Tasks Datasets: For the classification task, we use cMedIC and cMedTC datasets (Zhang et al., 2020) in the Chinese language. cMedTC dataset consists of biomedical texts with multiple labels. Similarly, the cMedIC dataset consists of queries with three intent labels (no intention, weak intention, and firm intention).

3 Results and Discussion

We use accuracy as a measure to evaluate the performance of our algorithms for both downstream tasks (relation extraction and text classification). From Table 1, it can be seen that the integration neighbor information has improved performance in relation extraction. Integration of KNN into K-BERT has significantly improved the model performance significantly, with an 8% increase in performance with both the GAD and EU-ADR datasets. Similarly, there is an increase of 6.4% in performance for the i2b2 dataset when KNN is integrated into K-BERT. Integration of GNN to create the aggregated representations has helped K-BERT to improve its performance significantly. There is an increased performance across all the datasets using the aggregated representations from GNN and K-BERT.

	GAD	EU-ADR	i2b2
K-BERT	0.634	0.807	0.814
K-BERT + KNN	0.687	**0.871**	0.866
K-BERT + GNN	**0.696**	0.860	**0.875**

Table 1: Results of different models with various datasets in relation extraction domain. The results show that the addition of KNN and GNN to K-BERT improves the performances in relation extraction significantly.

Similar to the relationship extraction task, the performance of K-BERT with our approach to integrating additional information has shown better performance than the base K-BERT model. From Table 2, it can be observed that adding information on neighboring nodes has improved the accuracy by 1% to 2%. The performance improvement of 1% to 2% is very important in the medical domain, where the decisions impact the lives of people.

	cMedIC	cMedTC	Δ w.r.t. K-BERT CMedIC	CMedTC
K-BERT	0.927	0.609	-	-
K-BERT + KNN	0.939	0.615	**+1.3%**	**+0.99%**
K-BERT + GNN	**0.941**	**0.621**	**+1.51%**	**+1.97%**

Table 2: Results of different models in datasets related to classification tasks. The results show that there is around a 1% to 2% increment in performance by integrating GNN and KNN.

The results show that the base model K-BERT improved performance by integrating supplementary information. The classification and relation extraction tasks are important in the medical domain. The ability of our proposed methodology to get improved performance can also be adapted to other tasks in the medical domain.

4 Conclusion

In this paper, we propose a methodology for integrating additional information, such as neighboring nodes and instances, to improve the performance of a language model in the medical domain. The additional information also provides models with the ability to learn beyond contextual information. The methodology proposed in our work is generic and can be adapted to multiple tasks and domains. The work can be extended to solving other critical problems in the medical domain like report generation, medical dialogue generation, etc. It would also be interesting to integrate both nearest neighbor and graph information into the language model and evaluate the model performance. Another important future direction can be exploring how adding information in our framework contributes to the explainability of the model. Overall, our proposed methodology shows promising results and has the potential to enhance the performance and interpretability of language models in various domains.

Limitations

While the proposed methods offer promising results in improving the performance of knowledge graph-based language representation models, there are some limitations to this work that should be noted. Firstly, the experiments were conducted on a limited number of datasets, and the results may not be generalized to other datasets or domains. Therefore, further experiments are needed to validate the effectiveness of the proposed methods on a broader range of datasets and NLP tasks. Secondly, the proposed methods require additional computation and may increase the complexity of the models. Therefore, it is important to consider the trade-off between performance improvement and computational cost when applying these methods in real-world applications. Lastly, while the proposed methods address some of the limitations of existing knowledge graph-based language representation models, they still may not capture all the contextual relationships and nuances of natural language, leading to potential semantic errors and reduced accuracy. Therefore, it is essential to continue exploring new approaches and techniques to further improve the performance of NLP models.

References

Surabhi Adhikari, Surendrabikram Thapa, Usman Naseem, Hai Ya Lu, Gnana Bharathy, and Mukesh Prasad. 2023. Explainable hybrid word representations for sentiment analysis of financial news. *Neural Networks*, 164:115–123.

Surabhi Adhikari, Surendrabikram Thapa, Priyanka Singh, Huan Huo, Gnana Bharathy, and Mukesh Prasad. 2021. A comparative study of machine learning and nlp techniques for uses of stop words by patients in diagnosis of alzheimer's disease. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Àlex Bravo, Janet Piñero, Núria Queralt-Rosinach, Michael Rautschka, and Laura I Furlong. 2015. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC bioinformatics*, 16:1–17.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2022. Ammu: a survey of transformer-based biomedical pretrained language models. *Journal of biomedical informatics*, 126:103982.

Nora Kassner and Hinrich Schütze. 2020. Bert-knn: Adding a knn search component to pretrained language models for better qa. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3424–3430.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2019. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained language models for biomedical and clinical tasks: understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157.

Linyang Li, Demin Song, Ruotian Ma, Xipeng Qiu, and Xuanjing Huang. 2021. Knn-bert: fine-tuning pretrained models with knn classifier. *arXiv preprint arXiv:2110.02523*.

Shuai Lin, Pan Zhou, Xiaodan Liang, Jianheng Tang, Ruihui Zhao, Ziliang Chen, and Liang Lin. 2021. Graph-evolving meta-learning for low-resource medical dialogue generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13362–13370.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE.

Jiachang Liu, Qi Zhang, Chongyang Shi, Usman Naseem, Shoujin Wang, and Ivor Tsang. 2023. Causal intervention for abstractive related work generation. *CoRR*, abs/2305.13685.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2901–2908.

Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. 2016. Large-margin softmax loss for convolutional neural networks. In *International Conference on Machine Learning*, pages 507–516. PMLR.

Zhibin Lu, Pan Du, and Jian-Yun Nie. 2020. Vgcn-bert: augmenting bert with graph embedding for text classification. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part I 42*, pages 369–382. Springer.

Usman Naseem, Ajay Bandi, Shaina Raza, Junaid Rashid, and Bharathi Raja Chakravarthi. 2022a. Incorporating medical knowledge to transformer-based language models for medical dialogue generation. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 110–115.

Usman Naseem, Adam G Dunn, Matloob Khushi, and Jinman Kim. 2022b. Benchmarking for biomedical natural language processing tasks with a domain specific albert. *BMC bioinformatics*, 23(1):1–15.

Usman Naseem, Matloob Khushi, Shah Khalid Khan, Kamran Shaukat, and Mohammad Ali Moni. 2021a. A comparative analysis of active learning for biomedical text mining. *Applied System Innovation*, 4(1):23.

Usman Naseem, Matloob Khushi, Vinay Reddy, Sakthivel Rajendran, Imran Razzak, and Jinman Kim. 2021b. Bioalbert: A simple and effective pre-trained language model for biomedical named entity recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.

Nazneen Fatema Rajani, Ben Krause, Wengpeng Yin, Tong Niu, Richard Socher, and Caiming Xiong. 2020. Explaining and improving model behavior with k nearest neighbor representations. *arXiv preprint arXiv:2010.09030*.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86.

Aditya Shah, Surendrabikram Thapa, Aneesh Jain, and Lifu Huang. 2023. Adept: Adapter-based efficient prompt tuning approach for language models. In *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Erik M Van Mulligen, Annie Fourrier-Reglat, David Gurwitz, Mariam Molokhia, Ainhoa Nieto, Gianluca Trifiro, Jan A Kors, and Laura I Furlong. 2012. The eu-adr corpus: annotated drugs, diseases, targets, and their relationships. *Journal of biomedical informatics*, 45(5):879–884.

Eric Wallace, Shi Feng, and Jordan Boyd-Graber. 2018. Interpreting neural networks with nearest neighbors. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 136–144.

Lingfei Wu, Yu Chen, Kai Shen, Xiaojie Guo, Hanning Gao, Shucheng Li, Jian Pei, Bo Long, et al. 2023. Graph neural networks for natural language processing: A survey. *Foundations and Trends® in Machine Learning*, 16(2):119–328.

Ningyu Zhang, Qianghuai Jia, Kangping Yin, Liang Dong, Feng Gao, and Nengwei Hua. 2020. Conceptualized representation learning for chinese biomedical text mining. *arXiv preprint arXiv:2008.10813*.

Medical knowledge-enhanced prompt learning for diagnosis classification from clinical text

Yuxing Lu
Department of BigData and Biomedical AI
College of Future Technology
Peking University
Beijing, China
yxlu0613@gmail.com

Xukai Zhao
Department of Landscape Architecture
School of Architecture
South China University of Technology
Guangzhou, China
zhaoxukai0208@163.com

Jingzhuo Wang
Department of BigData and Biomedical AI
College of Future Technology
Peking University
Beijing, China
wangjinzhuo@pku.edu.cn

Abstract

Artificial intelligence based diagnosis systems have emerged as powerful tools to reform traditional medical care. Each clinician now wants to have his own intelligent diagnostic partner to expand the range of services he can provide. When reading a clinical note, experts make inferences with relevant knowledge. However, medical knowledge appears to be heterogeneous, including structured and unstructured knowledge. Existing approaches are incapable of uniforming them well. Besides, the descriptions of clinical findings in clinical notes, which are reasoned to diagnosis, vary a lot for different diseases or patients. To address these problems, we propose a Medical Knowledge-enhanced Prompt Learning (MedKPL) model for diagnosis classification. First, to overcome the heterogeneity of knowledge, given the knowledge relevant to diagnosis, MedKPL extracts and normalizes the relevant knowledge into a prompt sequence. Then, MedKPL integrates the knowledge prompt with the clinical note into a designed prompt for representation. Therefore, MedKPL can integrate medical knowledge into the models to enhance diagnosis and effectively transfer learned diagnosis capacity to unseen diseases using alternating relevant disease knowledge. The experimental results on two medical datasets show that our method can obtain better medical text classification results and can perform better in transfer and few-shot settings among datasets of different diseases.

1 Introduction

Clinical notes in Electronic Health Records (EHRs) are the medical texts written by a physician to address the patient's medical history, chief complaints and examinations during a patient's visit. Physicians can get the corresponding diagnosis through their expertise based on the patient's clinical notes. In the past decade, researchers have tried various methods for medical text classification tasks to assist doctors in their treatment.

Text classification models in the generic domain are developing most rapidly. Traditional machine learning methods, such as Naive Bayesian (NB) (Maron, 1961), K-Nearest Neighbor (KNN) (Cover and Hart, 1967), Support Vector Machine (SVM) (Joachims, 1998), and Random Forest (RF) (Breiman, 2001) are first introduced to solve text classification tasks. For deep learning models, TextCNN (Chen, 2015) is widely used, where Convolutional Neural Network (CNN) (Albawi et al., 2017) models are introduced to solving text classification problems. Whereafter, Pre-trained Language Models (PLMs), such as BERT (Devlin et al., 2018) achieve state-of-the-art results on several Natural Language Processing (NLP) tasks and thus has been widely used. However, these approaches are based on generic data and therefore ignore the high reliance on medical knowledge in medical text classification tasks. When applied directly to the medical field, these models often fail to achieve the same performance as in the generic field.

To address the knowledge-dependent medical text classification tasks, researchers have proposed a number of medical text classification models that incorporate knowledge. Garla and Brandt (2013) map medical text to corresponding medical concepts and is the first to conduct feature engineering. Yao et al. (2019a) use medical concept descriptions

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 278–288
July 14, 2023 ©2023 Association for Computational Linguistics

Raw Clinical Notes	Runny nose and coughing with phlegm started 4 days ago.
+ Prompt Template	Patient: Runny nose and coughing with phlegm started 4 days ago. Which disease does the patient get? [MASK]
+ Knowledge Enhanced Prompt Template	Cough with croup and recurrent infections is a symptom of bronchitis, dyspnoea is a symptom of bronchitis. Patient: Runny nose and coughing with phlegm started 4 days ago. Which disease does the patient get? [MASK]

Figure 1: Different template generation methods for clinical notes. Prompt learning method simply adds questions to the clinical notes, our **Med**ical **K**nowledge-enhanced **P**rompt **L**earning method incorporates heterogeneous medical knowledge in the template.

to improve distributed document representations. Gasmi (2022) use external terminology resources to expand and represent the text with a combination of different methods. Nevertheless, these models only learn the relationship between the text and the corresponding knowledge, without having a good generalization ability. Therefore they tend to be less effective when transferring to the medical domains beyond the training data.

In the medical field, there are rich sources of knowledge, such as expert knowledge (Flores et al., 2011), medical knowledge bases (Zuccon et al., 2013), medical knowledge graphs (Li et al., 2019), medical information on the web, etc. These knowledge present a heterogeneous structure(such as triples, SQLs and free texts, etc.) and cannot be well uniformed in the previous methods. Differences among knowledge sources prevent these models from learning by using knowledge prompt from all sources and thus may have bias when dealing with real-world data. Therefore, we hope to propose a model that is compatible with all the sources of medical knowledge.

To solve the above problems in medical text classification, we propose a **Med**ical **K**nowledge-enhanced **P**rompt **L**earing (MedKPL) model that can uniform different knowledge sources. The contribution of this paper can be summarized as follows: 1) We design the MedKPL model to uniform heterogeneous knowledge by transforming knowledge from different sources into free texts. Experiments prove that structured and unstructured texts can be uniform in our model, and both yield good results. 2) We use the MedKPL model to conduct medical text classification tasks on two Chinese EHR datasets and obtain state-of-the-art classification results through knowledge incorporation. 3) We evaluate the MedKPL model for few-shot learning among departments. The results show that our method can obtain good results in both zero-shot and few-shot scenarios, and can effectively transfer between departments that have low text similarity in a robust way.

2 Related Work

2.1 Knowledge Enhancement for PLMs

PLMs has become text representation method in most NLP tasks. Generic PLMs are usually trained on unstructured text corpus without domain knowledge. For example, BERT (Devlin et al., 2018) is trained on BooksCorpus (Zhu et al., 2015) and English Wikipedia, GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020) use Common Crawl (Raffel et al., 2020) and WebText as training corpus. Due to training on generic datasets, most of contextual information learned by these PLMs lack domain knowledge, resulting in their lack of expertise in dealing with domain-specific problems.

Continuous Knowledge-enhancement uses knowledge encoders to get the embedding of knowledge and incorporate them into the process of training contextual representations of text. Know-BERT (Peters et al., 2019) propose Knowledge Attention and Recontextualization (KAR) and entity linking to incorporate knowledge into PLMs. ERNIE-THU (Zhang et al., 2019) introduce an knowledge fusion module, injecting entity embeddings through knowledge encoders. KEPLER (Wang et al., 2021) jointly optimize the knowledge embedding and language modeling objectives within the same PLM. DKPLM (Zhang et al., 2022) use pseudo token representations to embed long-tail entities which relieve computation burdens of previous methods.

Discrete Knowledge-enhancement retrieves knowledge directly from the knowledge graph and add them to training texts. K-BERT (Liu et al., 2020) and CoLAKE (Sun et al., 2020) directly reorganize the triples in the knowledge graph into texts and insert them directly into the training corpus, without pre-training any extra models. We also apply the ideas behind these methods to our work.

2.2 Prompt Learning

Prompt learning refers to transforming the original text via templates to leverage the contextual pattern learned by the PLMs. Brown et al. (2020) first use

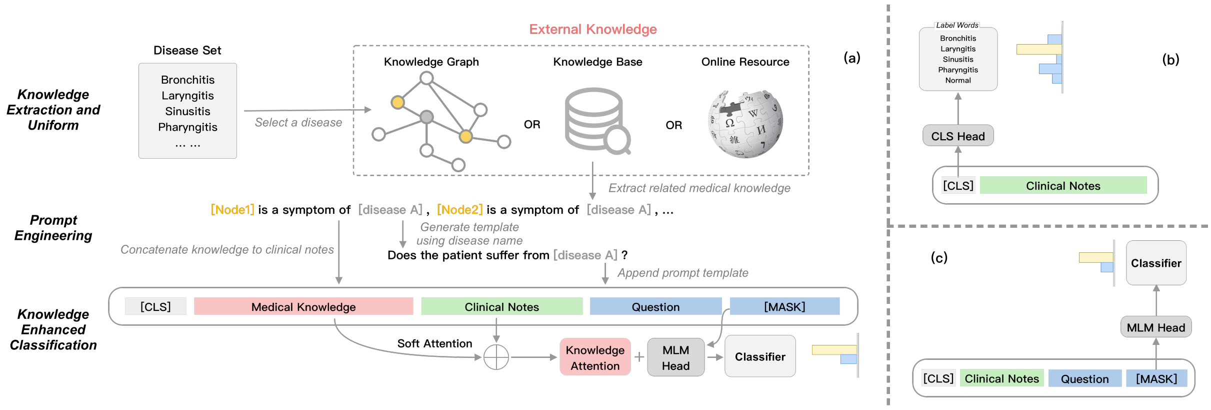


Figure 2: **The illustration of MedKPL and other methods.** (a) is the workflow of MedKPL, knowledge can be obtained from different knowledge sources and then incorporated into clinical notes through template construction, and the classifier (multi-classifier and binary classifier) can be further enhanced by using soft attention on knowledge prompt and clinical notes. (b) is the method for fine tuning at PLM to classify the embedding of the [CLS] token, and (c) is the method for regular prompt learning to predict the probability distribution of the [MASK] token.

the prompt learning method for text classification tasks and find it works well on few-shot learning scenarios. Schick and Schütze (2020) reformulate inputs as cloze questions for text classification. Schick et al. (2020) and Gao et al. (2020) extend previous methods by automatically generating label words and templates, respectively. Recently, some knowledge-related prompt learning methods have been proposed. Hu et al. (2021) incorporate external knowledge into the verbalizer with calibration. Chen et al. (2022) inject latent knowledge into learnable virtual type words and answer words.

Compared with these approaches, our approach can uniform heterogeneous knowledge to build prompt templates, which solves the differences brought by different knowledge formats sources. Our approach also provides a deep integration between clinical notes and knowledge prompts.

2.3 Medical Text Classification

How to apply external knowledge to medical text classification tasks is a topic that has been constantly explored by researchers. Garla and Brandt (2013) map clinical text to Unified Medical Language System (UMLS), and use those UMLS Concept Unique Identifiers (CUIs) as features to train classifiers on medical documents. Yao et al. (2019a) propose to distribute document representations with medical concept descriptions for the classification of traditional Chinese medicine clinical records. Yao et al. (2019b) combine rule-based features and knowledge-guided CNN for effective disease classification. Li and Yu (2020) use multi-filter Residual CNN to predict ICD codes. Chen et al. (2020) propose an attention-based bidirectional LSTM model for classifying outpatient categories according to textual content.

However, none of these works mentioned the model's transferability among departments and few-shot learning ability, which are issues that must be addressed to solve the medical long-tail problem and achieve truly trustworthy medical AI.

3 Method

The overall structure of our model is shown in Figure 2. Our model introduce disease $d \in D$ related medical knowledge prompt k_d into medical text classification tasks, where D is the disease set. The knowledge prompt can come from a variety of sources, e.g. expert knowledge, knowledge graphs, knowledge bases, online resources, etc. We use $p(y|x_i, k)$ to denote the probability of patient i getting disease y, where x_i is the clinical notes for patient i, and k is the set of knowledge prompts used for knowledge incorporation.

Specifically, we decompose the process of knowledge incorporation into three stages. 1) Extract medical knowledge of disease d from different knowledge sources and transform the knowledge into a uniform representation k_d. 2) Construct templates that incorporate knowledge prompts set k with clinical notes. We concatenate the collected medical knowledge prompts into natural text and generate the template based on the disease name d. 3) Predict labels using MLM on the [MASK] token in prompt template. It is also possible to integrate knowledge prompt and clinical notes at a deep level by using PLM to represent knowledge prompt and clinical notes separately and aligning them using soft attention mechanisms to enhance the knowledge representation.

We will then go over our model's methodology and its three stages of knowledge incorporation.

3.1 Knowledge Extraction and Uniform

Unstructured knowledge is naturally available as part of the prompt template, while structured knowledge needs to be pre-processed. For structured medical knowledge, the most common organization form is the medical knowledge graph. Thus we take knowledge graph as our knowledge source and denote it as $\mathcal{G} = (\mathcal{E}, \mathcal{R})$ where $\mathcal{E}$ is the collection of all entities and $\mathcal{R}$ is the collection of all relations. In the knowledge graph $\mathcal{G}$, a relational knowledge triple is denoted as (e_h, r, e_t), where $e_h \in D$ is the head entity and e_t is the tail entity. r is the specific relation between e_h and e_t.

In a large-scale medical knowledge graph, a disease may have multiple relations, we denote the relation set of disease e_i as R_i. The distribution of triples related to disease e_i is very diverse and complex, and we need to find those triples $(e_i, r, e_j) \in \mathcal{G}, r \in R_i$ that are suitable for our medical knowledge-enhanced prompt learning method.

Specifically, we determine the refined relation set $R'_i = (r_1, r_2, \ldots, r_k), r_i \in \mathcal{R}_i$ based on the relationships commonly mentioned in the clinical notes for disease e_i. Then with the disease e_i and the refined relation set R'_i, we can retrieve all relevant triples $\mathcal{T}_i$ of disease e_i from the knowledge graph.

$$\mathcal{T}_i = \{(e_i, r_i, e_j) | r_i \in \mathcal{R}'_i, e_i, e_j \in \mathcal{E}\} \in \mathcal{G} \quad (1)$$

For those diseases lacking relevant medical knowledge, we consider using similar entities e_j for replacement, where $(e_i, r_{syn}, e_j) \in \mathcal{G}$ and r_{syn} is the relationship of synonym. For those diseases not in the entity set of the medical knowledge graph, we consider replacing them with other knowledge sources(e.g., online search engines).

Alternatively, we also consider using unstructured medical knowledge, such as knowledge bases and online search engines for replacement. Medical knowledge related to disease d can be represented as k_d. Since this unstructured knowledge is already in the form of text, we apply them directly to the subsequent processes.

3.2 Prompt Engineering

The core idea of the prompt learning method is to construct templates and use the contextual knowledge learned by the PLM during the pre-training process to make predictions on the masked words.

Different from the normal prompt approach, we want our templates to contain medical knowledge extracted from heterogeneous knowledge sources. Therefore, we propose a disease-adaptive template generation method. For a disease d, if the knowledge source is KG, we first extract all the required knowledge triples $\mathcal{T}_d$ from the KG and concatenate all the triples together into free texts. Given an example knowledge triple t =*(dyspnoea, a symptom of, bronchitis)*, the formed free-text knowledge would correspondingly be ***Dyspnoea is a symptom of bronchitis***. By concatenating all the triples, we can get the disease-related knowledge k_d in the text pattern.

The promoting function $f_{prompt}(k_d, x, d)$ contains medical knowledge and manual template engineering. We devise templates for binary classification tasks and multi-classification tasks seperately. These two tasks are different in practical medical application scenarios, where a multi-classification task can quickly determine which disease the patient is most likely to have, and a binary classification task can make predictions about the likelihood of a specific disease more precisely. For binary classification tasks, the prompt learning method will extend the input clinical notes x into

$$x' = [K_d][X] \text{ Does the patient suffer from } [D]? \ [MASK].$$

and for multi-classification tasks, the input clinical notes x will be turned into

$$x' = [K_d][X] \text{ Which disease the patient have? } [MASK].$$

where the slot $[K_d]$, $[X]$, $[D]$ are filled with k_d, x, d respectively. In this way, we convert the sequence classification task into a task of predicting the distribution of masked token [MASK].

By organizing all the heterogeneous knowledge into free texts, we can extend the knowledge sources of MedKPL to almost all types of medical knowledge.

3.3 Knowledge Enhanced Classification

By simply concatenating and adding knowledge to the template, we can use PLM to learn the contextual association between clinical notes and knowledge prompt. However, this approach treats them in a sentence as a whole. To better explore the deeper connection between clinical notes and knowledge prompt, we integrate these texts in a deeper way.

Vector representation of the knowledge prompt $K = (k_1, k_2, \ldots, k_m)$ and clinical notes $C(c_1, c_2, \ldots, c_n)$ can be obtained by PLM, where m and n are the length of knowledge prompt and clinical notes respectively. We use the Soft Attention mechanism (Luong et al., 2015) to align clinical notes with knowledge prompt.

Specifically, we select the [CLS] token $k_1 \in K$ as the vector representation of the whole knowledge prompt and calculate the alignment vector a which is calculated by comparing the knowledge prompt representation k_1 with each clinical note word's hidden state $c_s \in C$:

$$a_s = align(k_1, c_s) = \frac{exp(score(k_1, c_s))}{\Sigma_{s'=0}^{n}(score(k_1, c_{s'}))} \quad (2)$$

where we use dot product function to compute scores.

$$score(k_1, c_s) = k_1^T c_s \quad (3)$$

Given alignment vector a as weight, the integrated vector i_t is computed as weighted average over all the words' representations in clinical notes. The integrated vector $i_t = \sum_{s=0}^{n} a_s c_s$ can enhance the most relevant part of the clinical notes with the knowledge prompt. For medical text classification, we sum the integrated vector i_t with the Masked Language Model (MLM) prediction x_{mlm} on [MASK] to get $xintegrate$ and compute the loss based on the classification tasks.

$$x_{mlm} = f_{MLM}(x', [MASK]) \quad (4)$$

$$x_{integrate} = W_x x_{mlm} + W_i i_t \quad (5)$$

where f_{MLM} is the masked language model of the PLM. For binary classification tasks, the loss function L_{binary} is computed directly between $x_{integrate}$ and the index of label words ("yes" or "no") in the PLM's vocabulary.

$$L_{binary} = CELoss(x_{integrate}, label) \quad (6)$$

where the CELoss is cross entropy loss. For multi-classification tasks, the loss function is computed by first map $x_{integrate}$ into the label space using a fully-connected layer and compute the cross entropy loss.

$$L_{multi} = CELoss(W_x x_{integrate} + b_x, label) \quad (7)$$

where W_x and b_x are learnable parameters in the model, and $label$ represents the categories in mult-classification tasks.

4 Experiments

4.1 Datasets

In this paper, we compare our results against many existing methods on two medical datasets. The first dataset is the Pediatric Patient EHR (PPE) used in (Liang et al., 2019), which contains 1,362,559 outpatient visits from 567,498 pediatric patients across 6 departments, each outpatient visit includes adverse event, chief complaint and history of present illness, some also have physical examination and image report. The second dataset is Adult-EMR, which contains 339,672 EHR records for 2556 diseases across 12 departments, each record includes chief complaint and history of present illness. We use the clinical notes of patients' history of present illnesses for training. In PPE, we use the clinical notes of the Hematology-Immunology department as normal control data and select six diseases from each of the other five departments (Respiratory (Resp.), Gastroenterology (Gast.), Psychiarty (Psy.), Neurology (Neuro.), Gynecology (Gyn.)) for experiments. In Adult-EMR, we use the clinical notes of the Respiratory Department as normal control data and select six diseases from Tumor and Cancer Department and Cardiology Department for experiments. The knowledge graph we use in our experiments is the DiseaseKG, which is an open-source Chinese medical knowledge graph from OpenKG.

4.2 Settings

In the multi-label classification task, we select 1000 samples from each of the k diseases $(k = 2, 4, 6)$ from a department and 1000 samples from normal control data for $k + 1$ classification task. The knowledge prompt is the concatenation of the truncated knowledge from

Table 1: Standard multi-classification accuracy on different departments. "+ERNIE" and "+DKPLM" means using knowledge-enhanced PLMs to replace BERT, "+Attn" means using the attention layer to enhance the classification performance. The results for each department are acquired by averaging the results for disease number k = 2, 4, 6.

	PPE						Adult EMR		
	Resp.	Gast.	Psy.	Neuro.	Gyn.	Overall	Tumor.	Cv.	Overall
LSTM	64.89	77.08	85.63	90.29	77.59	79.10	67.97	64.65	66.31
LSTM+Attn	65.49	78.89	85.58	87.58	79.59	79.42	62.87	74.35	68.61
CNN	69.31	81.98	85.76	91.41	81.38	81.97	71.25	73.05	72.15
Fine tuning	68.74	80.25	86.96	89.41	81.75	81.42	71.00	74.46	72.73
Prompt	71.05	82.51	89.06	91.47	82.17	83.25	71.43	76.45	73.94
Prompt+ERNIE	70.24	83.27	88.08	91.76	80.84	82.84	69.57	73.96	71.77
Prompt+DKPLM	73.94	84.77	88.76	91.69	81.82	84.20	72.14	76.62	74.38
MedKPL (Ours)	74.06	83.72	89.13	92.29	82.10	84.26	**73.14**	77.44	75.29
MedKPL+DKPLM (Ours)	**75.01**	**85.05**	**90.11**	**92.40**	**83.96**	**85.31**	72.71	**78.61**	**75.66**

Table 2: Standard binary classification accuracy on different departments. "+ERNIE" and "+DKPLM" means using knowledge-enhanced PLMs to replace BERT. The results for each department are acquired by averaging the results for disease number k = 2, 4, 6.

	PPE						Adult EMR		
	Resp.	Gast.	Psy.	Neuro.	Gyn.	Overall	Tumor.	Cv.	Overall
Prompt	88.44	87.77	97.28	92.31	93.56	91.87	86.67	91.13	88.90
Prompt+ERNIE	85.11	81.78	95.22	88.75	91.92	88.56	84.33	90.54	87.44
Prompt+DKPLM	89.92	89.17	92.14	95.58	97.31	92.82	89.17	95.10	92.14
MedKPL (Ours)	94.89	96.06	**98.69**	96.08	99.19	96.98	**96.33**	96.45	**96.39**
MedKPL+DKPLM (Ours)	**95.75**	**97.36**	**98.69**	**96.75**	**99.31**	**97.57**	94.33	**97.78**	96.06

all the diseases, the basic truncation length is 50 per disease. In the binary classification task, we select 500 samples from each of the k diseases $(k = 2, 4, 6)$ from a department and select $k * 500$ samples from the normal control data for binary classification task. The knowledge prompt used for normal control data is randomly selected from all extracted medical knowledge in binary classification tasks.

Standard Settings. For traditional NLP methods, we select LSTM (Liu et al., 2016), CNN (Chen, 2015) and LSTM (Chen et al., 2020) and LSTM with attention (Chen et al., 2020) for comparison.The word embedding for LSTM and CNN models is the 300-dimension skip-gram word embedding (Mikolov et al., 2013) pre-trained on Sogou News corpus (Li et al., 2018), and the word embedding for models applying PLM is BERT-base-chinese (Devlin et al., 2018) if not otherwise stated. For fine tuning, we take the classification token [CLS] and feed it into a fully connected layer for classification, as shown in Figure 2 (b). For prompt learning (Brown et al., 2020), we calculate the probability distribution of the [MASK] token and further predict the classification result, as shown in Figure 2 (c). In addition, we also try different knowledge-integrated PLMs for comparisons, such as ERNIE (Zhang et al., 2019) and medical version of DKPLM (Zhang et al., 2022). These models above are used as the baseline in our experiments. We use BERT and DKPLM as PLMs to conduct experiments on our method, the use of DKPLM on our method can be regarded as using medical knowledge in both the pre-training phase and the prompt learning phase.

Low-Resource Settings. In our experiments, we design a couple of different low-resource scenarios on binary classification tasks. The first is model transferring among departments. We compare the effect of 16-shot transfer learning among the five departments of the PPE dataset, comparing the results of fine tuning, prompt learning, and our method. In addition, we conduct 0-, 2-, 4-, 8-, and 16-shot transfer learning experiments to compare the effectiveness of our method with other methods on few-shot learning tasks.

In all of our experiments, we use Adam (Kingma and Ba, 2014) as the optimizer with a learning rate of $1e-7$. The training epoch is 20, the batch size is 32, and the dropout rate is 0.5. Due to the average length of knowledge prompt is 48, we set the truncation lengths of a disease's knowledge prompt as 50, and we set the truncation length of the input clinical notes with prompt template as 128. We use the cross entropy loss as the loss function.

All experiments are conducted on a single NVIDIA Tesla V100. The evaluation metric is accuracy, which is widely used in text classification tasks (Lee and Dernoncourt, 2016).

4.3 Results

4.3.1 Standard Results

We first evaluate the performance on multi-classification tasks under standard text classification task settings. The results are shown in Table 1, where we compared to a range of baselines. The result shows that all MedKPL methods, consistently outperform traditional NLP methods, fine tuning and prompt learning baselines, indicating the effectiveness of our methods. Moreover, as a pre-trained PLM using medical knowledge, prompt learning method using DKPLM outperforms the standard prompt learning method by 0.95 percent in multi-classification performance on PPE dataset, showing the effectiveness of knowledge-enhanced PLMs. However, the knowledge-enhanced PLM ERNIE, which is trained on generic knowledge, is 0.41 percent weaker than the standard prompt learning method. This demonstrates that the incorporation of medical knowledge in the pre-training phase does benefit the medical downstream tasks. In addition, replacing BERT in our model with DKPLM can further yield better results.

We also conduct experiments on binary classification tasks and the results are shown in Table 2, where MedKPL outperforms other methods in a larger gap compared with multi-classification tasks. We conjecture this is because in the binary classification tasks, the

Table 3: The effect of different methods on transferring between departments, this table selects the results of transfer from Respiratory department (Resp.) to other 4 departments. We choose the sample size $shots = 16$ and the number of diseases $k = 6$ as the parameters in the transfer learning experiment.

Resp.→	Gast.	Psy.	Neuro.	Gyn.
Fine tuning	74.92	53.08	54.75	46.92
Prompt tuning	74.83	73.08	68.58	66.50
MedKPL	**85.83**(+11)	**86.83**(+13.75)	**84.42**(+15.84)	**80.83**(+14.33)

Table 4: The effect of transferring MedKPL from Respiratory department (Resp.) to Gastroenterology department (Gast.) with different sample sizes was tested with $shots = 0, 2, 4, 8, 16$ and number of diseases $k = 6$.

Resp.→ Gast.			
shots	Fine tuning	Prompt learning	MedKPL
0	53.97	72.75	**84.92**
2	60.85	86.24	**89.98**
4	68.25	87.57	**90.48**
8	71.43	89.98	**90.48**
16	71.16	89.42	**91.53**

knowledge prompt only contains knowledge of the selected disease, so the model can learn the relationship between the knowledge prompt and clinical notes in a more targeted way.

For the analysis of each department of classification task on the PPE dataset, we observe that the model's performance in the psychiatry (Psy.) department and the gynecology (Gyn.) department are highest both on multi-classification and binary classification tasks. By looking at the clinical notes in these two departments, we conjecture that the model's good performance is due to the low noise contained in the texts of these two departments.

4.3.2 Low-Resource Results

We conduct experiments on transfer learning across departments in PPE dataset and select the results of transferring from the Respiratory department (Resp.) to other departments in Table 3. The results in Table 3 show that the transferability of our method among departments outperforms the fine tuning and prompt learning methods by a large margin.

According to the results, there is also an in-

Table 5: The impact of different knowledge sources on the effect of MedKPL model, where the **Structured** is obtained from the Knowledge Graph, the **Unstructured** is obtained from online resources such as Wikipedia, the **Plain Text** uses the phrase *The disease requires timely medical attention.* as the text that does not contain medical knowledge, and the **Random** refers to randomly selected knowledge for augmentation. The results for each department are acquired by averaging the multi-classification results for disease number $k = 2, 4, 6$.

	Resp.	Gast.	Psy.	Neuro.	Gyn.	Overall
Structured	**85.17**	84.30	**94.03**	92.24	**95.12**	**90.17**
Unstructured	85.13	**84.90**	90.59	84.93	94.17	87.94
Plain Text	72.21	71.64	86.42	**92.62**	84.03	81.39
Random	54.31	61.37	83.36	81.12	69.17	69.87

teresting phenomenon that departments with lower text similarity have a higher improvement on classification accuracy, we conjecture that this is because our knowledge incorporation approach allows our model to discover the association between knowledge prompt and clinical notes in a more direct way. Also, by calculating the variance for all the results, we get the variance of 104.17 for fine tuning method and 83.55 for prompt learning method, while the variance of our method is 56.15, which is much lower than that of fine tuning and prompt learning. Therefore we speculate that our method can achieve higher classification results while having good robustness at the same time.

Besides transferring to other departments, We have also tested our method under different transfer shots to further demonstrate our model's few-shot learning capability. The results of transferring from Respiratory department (Resp.) to Gastroenterology department (Gast.) with different shots are shown in Table 4. It can be observed that under the zero-shot scenario, our method is far superior to the fine tuning and prompt learning methods. As the sample size rises, all methods witnesses an increase in transfer effect, but our method is still the best among the three methods.

Overall, our MedKPL model is more capable of transferring among departments and can also be better adapted to few or zero-shot scenarios.

4.3.3 Comparison among Knowledge Sources

To demonstrate that our model can uniform heterogeneous knowledge as input, we test different knowledge sources and their corresponding classification effects, results are shown in Table 5.

We begin by contrasting the structured knowledge prompt, derived from the knowledge graph, with the unstructured knowledge prompt, sourced from online search. Our findings demonstrate that the structured knowledge prompt outperforms its unstructured counterpart in terms of classification accuracy. This suggests that there exists a trade-off between the quality and accessibility of knowledge. While the structured knowledge prompt is more refined and contains less noise and irrelevant information, it is also more challenging to access. Conversely, unstructured free-text knowledge prompts offer almost limitless accessibility. For cases involving plain text, we employ the sentence ***The disease necessitates expedient medical attention.*** as the knowledge prompt. However, we observed that this non-medical knowledge prompt yielded significantly lower classification performance than the previous two methods. Furthermore, we conducted an experiment to disrupt the knowledge prompt by augmenting clinical notes with a random, irrelevant piece of knowledge prompt. Our results indicate that this method is the least effective among the four knowledge sources, with some outcomes even lower than the fine-tuning method. These findings reinforce the notion that knowledge prompts can contribute to improved classification outcomes in our approach.

In general, our approach can handle heterogeneous medical knowledge in a uniform way. The structured knowledge pormpt works most effectively, but is relatively difficult to obtain, while the unstructured knowledge can be accessed more easily, but at the expense of some performance.

4.3.4 Ablation Study

To explore how much the knowledge prompt contributes to our model, we conduct some ablation experiments of the impact of two main components: length of knowledge and soft

Table 6: Ablation study on a) knowledge length and b) soft attention. We test the knowledge truncation length from 0 to full length and test methods with or without soft attention mechanism, experimental parameters are kept consistent and the number of diseases $k = 6$.

(a)

Knowledge length	Acc.
full-length	**95.67**
40	95.33
30	94.33
20	95.33
10	94.50
0	93.33

(b)

Attention	Acc.
w/o Attention	93.67
w/ Clinical notes	94.17
w/ Knowledge	93.83
w/ Soft Attention	**95.67**

attention mechanism. Results are shown in Table 6.

It is noteworthy that the average length of medical knowledge in the Respiratory department is 36. The experimental results presented in Table 6a reveal that the model performs optimally when the medical knowledge is not truncated. We hypothesize that this is because larger truncation lengths promote the seamless integration of medical knowledge. Additionally, we evaluated the knowledge-enhanced classification module depicted in Figure 2(a) by comparing the soft attention mechanism with only clinical notes embeddings or knowledge prompt embeddings. The results in Table 6b demonstrate that the soft attention mechanism is instrumental in directing the model's focus towards the knowledge-laden attributes of the clinical notes, thereby leading to superior classification outcomes.

5 Conclusion

In this paper, we propose a MedKPL model and achieve state-of-the-art classification results on two medical EHR datasets. With the advantage of knowledge extraction and uniform process, our model can eliminate the difference among different sources and organize all knowledge into one representation style. The knowledge incorporation and soft attention mechanism between knowledge prompt and clinical notes enable the model to be more robust and achieve appreciable improvement on medical text classification tasks. The introduction of knowledge and prompt learning method exploits better few-shot and zero-shot transferability among departments.

References

Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. 2017. Understanding of a convolutional neural network. In 2017 international conference on engineering and technology (ICET), pages 1–6. Ieee.

Leo Breiman. 2001. Random forests. Machine learning, 45(1):5–32.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901.

Che-Wen Chen, Shih-Pang Tseng, Ta-Wen Kuan, and Jhing-Fa Wang. 2020. Outpatient text classification using attention-based bidirectional lstm for robot-assisted servicing in hospital. Information, 11(2):106.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In Proceedings of the ACM Web Conference 2022, pages 2778–2788.

Yahui Chen. 2015. Convolutional neural network for sentence classification. Master's thesis, University of Waterloo.

Thomas Cover and Peter Hart. 1967. Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1):21–27.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

M Julia Flores, Ann E Nicholson, Andrew Brunskill, Kevin B Korb, and Steven Mascaro. 2011. Incorporating expert knowledge when learning bayesian network structure: a medical case study. Artificial intelligence in medicine, 53(3):181–204.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. arXiv preprint arXiv:2012.15723.

Vijay N Garla and Cynthia Brandt. 2013. Knowledge-based biomedical word sense disambiguation: an evaluation and application to clinical document classification. Journal of the American Medical Informatics Association, 20(5):882–886.

Karim Gasmi. 2022. Medical text classification based on an optimized machine learning and external semantic resource. Journal of Circuits, Systems and Computers.

Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Juanzi Li, and Maosong Sun. 2021. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. arXiv preprint arXiv:2108.02035.

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In European conference on machine learning, pages 137–142. Springer.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Ji Young Lee and Franck Dernoncourt. 2016. Sequential short-text classification with recurrent and convolutional neural networks. arXiv preprint arXiv:1603.03827.

Fei Li and Hong Yu. 2020. Icd coding from clinical text using multi-filter residual convolutional neural network. Proceedings of the AAAI Conference on Artificial Intelligence, 34(05):8180–8187.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical reasoning on Chinese morphological and semantic relations. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 138–143, Melbourne, Australia. Association for Computational Linguistics.

Xuedong Li, Yue Wang, Dongwu Wang, Walter Yuan, Dezhong Peng, and Qiaozhu Mei. 2019. Improving rare disease classification using imperfect knowledge graph. BMC Medical Informatics and Decision Making, 19(5):1–10.

Huiying Liang, Brian Y Tsui, Hao Ni, Carolina Valentim, Sally L Baxter, Guangjian Liu, Wenjia Cai, Daniel S Kermany, Xin Sun, Jiancong Chen, et al. 2019. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. Nature medicine, 25(3):433–438.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. arXiv preprint arXiv:1605.05101.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-bert: Enabling language representation with knowledge graph. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 2901–2908.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025.

Melvin Earl Maron. 1961. Automatic indexing: an experimental inquiry. Journal of the ACM (JACM), 8(3):404–417.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 26.

Matthew E Peters, Mark Neumann, Robert L Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. 2019. Knowledge enhanced contextual word representations. arXiv preprint arXiv:1909.04164.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res., 21(140):1–67.

Timo Schick, Helmut Schmid, and Hinrich Schütze. 2020. Automatically identifying words that can serve as labels for few-shot text classification. arXiv preprint arXiv:2010.13641.

Timo Schick and Hinrich Schütze. 2020. Exploiting cloze questions for few shot text classification and natural language inference. arXiv preprint arXiv:2001.07676.

Tianxiang Sun, Yunfan Shao, Xipeng Qiu, Qipeng Guo, Yaru Hu, Xuanjing Huang, and Zheng Zhang. 2020. Colake: Contextualized language and knowledge embedding. arXiv preprint arXiv:2010.00309.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. Kepler: A unified model for knowledge embedding and pre-trained language representation. Transactions of the Association for Computational Linguistics, 9:176–194.

Liang Yao, Zhe Jin, Chengsheng Mao, Yin Zhang, and Yuan Luo. 2019a. Traditional chinese medicine clinical records classification with bert and domain specific corpora. Journal of the American Medical Informatics Association, 26(12):1632–1636.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019b. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. BMC medical informatics and decision making, 19(3):31–39.

Taolin Zhang, Chengyu Wang, Nan Hu, Minghui Qiu, Chengguang Tang, Xiaofeng He, and Jun Huang. 2022. Dkplm: Decomposable knowledge-enhanced pre-trained language model for natural language understanding. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 11703–11711.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. arXiv preprint arXiv:1905.07129.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In Proceedings of the IEEE international conference on computer vision, pages 19–27.

Guido Zuccon, Amol S Wagholikar, Anthony N Nguyen, Luke Butt, Kevin Chu, Shane Martin, and Jaimi Greenslade. 2013. Automatic classification of free-text radiology reports to identify limb fractures using machine learning and the snomed ct ontology. AMIA Summits on Translational Science Proceedings, 2013:300.

Multilingual Clinical NER: Translation or Cross-lingual Transfer?

Félix Gaschi[1,2*] **Xavier Fontaine**[1*] **Parisa Rastin**[2] **Yannick Toussaint**[2]
[1]SAS Posos, France [2]LORIA, France
{xavier,felix}@posos.fr
{felix.gaschi,parisa.rastin,yannick.toussaint}@loria.fr

Abstract

Natural language tasks like Named Entity Recognition (NER) in the clinical domain on non-English texts can be very time-consuming and expensive due to the lack of annotated data. Cross-lingual transfer (CLT) is a way to circumvent this issue thanks to the ability of multilingual large language models to be fine-tuned on a specific task in one language and to provide high accuracy for the same task in another language. However, other methods leveraging translation models can be used to perform NER without annotated data in the target language, by either translating the training set or test set. This paper compares cross-lingual transfer with these two alternative methods, to perform clinical NER in French and in German without any training data in those languages. To this end, we release MedNERF a medical NER test set extracted from French drug prescriptions and annotated with the same guidelines as an English dataset. Through extensive experiments on this dataset and on a German medical dataset (Frei and Kramer, 2021), we show that translation-based methods can achieve similar performance to CLT but require more care in their design. And while they can take advantage of monolingual clinical language models, those do not guarantee better results than large general-purpose multilingual models, whether with cross-lingual transfer or translation.

1 Introduction

In recent years, pre-trained language models based on the Transformer architecture (Vaswani et al., 2017) have demonstrated high performance on many natural language tasks such as Named Entity Recognition (NER), Natural Language Inference or Question-Answering (Devlin et al., 2018; Liu et al., 2019). These models, which are generally pre-trained on general domain data, can be fine-tuned on downstream tasks to achieve state-of-the-art results. Such models can also be adapted to a specific domain such as the legal (Chalkidis et al., 2020) or the biomedical fields (Gu et al., 2021) and can then outperform the general-domain models on domain-specific tasks.

Extracting medical entities from unstructured texts has become an essential tool to structure medical reports, and pre-trained language models have been naturally used to perform this task (Khan et al., 2020; Yang et al., 2020). These models use training data for fine-tuning that come from biomedical NER datasets like NCBI-disease (Islamaj Doğan and Lu, 2012) or n2c2 (Henry et al., 2019). However most of these datasets are only available in English and consequently the majority of such medical NER algorithms are developed for the English language, while medical reports or drug prescriptions are rather written in the country's language.

In the clinical domain, non-English datasets to fine-tune a model are even rarer than in the general domain. Even large domain-specific unlabeled corpora are mostly found in English. For example, the biomedical scientific literature is mostly written and available in English. Moreover, gathering medical texts and annotating them using expert knowledge is very expensive for low-resource languages and more generally for any non-English language. This restricts the development of medical NER models for non-English languages.

Fortunately, Cross-lingual Transfer (CLT) can work around the absence of training data in the target language, by making use of language models pre-trained on multilingual data. CLT consists in applying on a specific task a multilingual large language model (MLLM) fine-tuned in another language for the same task. For example, models like mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) can be fine-tuned on a general-domain NER task in English and provide competitive results when evaluated in other languages on this task (Pires et al., 2019; Wu and Dredze, 2019).

Before MLLMs, cross-lingual adaptation

*Equal contribution.

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 289–311
July 14, 2023 ©2023 Association for Computational Linguistics

was generally tackled using translation methods (Yarowsky and Ngai, 2001). The lack of annotated data can be overcome by translating an English training set into the target language and by projecting the annotations with an alignment algorithm. Another approach consists in translating a test set into English to apply an English NER on it and in projecting the predicted labels in the original language.

Recently, translation models based on the Transformer architectures (Ott et al., 2019; Kocmi et al., 2022) have demonstrated huge improvements over other machine translation algorithms. Such models can therefore be leveraged to perform cross-lingual learning. However, comparing CLT with translation-based methods has attracted little attention and only comparisons in the general domain (Yarmohammadi et al., 2021) has been done.

Our proposed contribution is three-fold: (1) we perform extensive experiments with CLT and translation-based methods on a medical NER task in French and in German without any training data in those languages, (2) we release MedNERF, a French clinical NER dataset based on drug prescriptions, which serves as a test set for our experiments in French, (3) we demonstrate that CLT and translation provide comparable results for clinical NER, that the choice of the technique depends on the model's size and that using domain-specific models does not necessarily improve the results over using multilingual models.

2 Related Work

Multilingual large language models (MLLMs). MLLMs are the logical multilingual extension of Large Language Models, trained on multilingual corpora. There exist several versions of MLLMs among which the most popular are mBERT (Devlin et al., 2018), which is BERT multilingual version pre-trained on Wikipedia in a hundred different languages instead on only English Wikipedia (and BookCorpus), and XLM-R (Conneau et al., 2020) which is a multilingual Transformer encoder using the same pre-training principles as RoBERTa (Liu et al., 2019), trained in 100 languages on a larger and more diverse corpus than Wikipedia.

XLM-R, mBERT and its distilled version distilmBERT (Sanh et al., 2019) are not trained on any parallel data and show nevertheless strong CLT abilities, for example on NER tasks, even between languages which use different sets of characters (Pires et al., 2019; Wu and Dredze, 2019).

Translation-based cross-lingual learning. Cross-lingual learning can also be achieved by using a translation model for either training an algorithm on translated data or using a NER model on translated texts at inference. These techniques (Yarowsky and Ngai, 2001; Yarowsky et al., 2001) make use of translation models and alignment algorithms and have been compared with CLT on general domain tasks by Yarmohammadi et al. (2021) who have shown that using translated and aligned training data improved over zero-shot learning for tasks and languages with weak CLT performances. However this comparison has not been done for domain-specific tasks that often require specific language models like in the medical field. These techniques seem nevertheless to provide good results since they have been leveraged to propose the GERNERMED (Frei and Kramer, 2021) and GERNERMED++ (Frei et al., 2022) models which are medical German NER trained on a German automatic translation of the English dataset n2c2 (Henry et al., 2019).

Performing cross-lingual adaptation in a domain-specific setting raises other questions: such as whether domain-specific language models could be leveraged in translation-based methods, or if translation and alignment models fine-tuned on in-domain data can benefit those methods. To the best of our knowledge, those questions have not been addressed in the literature.

Neural machine translation models. Current state-of-the-art machine translation algorithms are based on the Transformer architecture (Kocmi et al., 2022) and can be either encoder-decoder models (Tiedemann and Thottingal, 2020; Ng et al., 2020) or decoder-only models (Gao et al., 2022). The quality of the translation can be assessed by different metrics such as the traditional BLEU score (Papineni et al., 2002), which is a statistical algorithm based on matching n-grams between a proposed translation and a reference one. It has been used for years but new scoring algorithms such as COMET (Rei et al., 2020) which leverage large multilingual models seem to provide more accurate evaluations.

Word alignment algorithms. Word alignment algorithms are designed to provide a mapping between the words of a sentence and those of its translation. They originally relied on statistical features like the `fast_align` algorithm (Dyer et al., 2013) and have been outperformed by models using contextualized embeddings (Jalili Sabet et al., 2020; Dou and Neubig, 2021).

Evaluation corpora for CLT. The CLT abilities of MLLMs can be assessed on several tasks in many languages thanks to multilingual benchmarks like XTREME (Hu et al., 2020) which cannot be used to evaluate medical NER models since they contain only general-domain tasks. Despite the existence of non-English medical NER datasets like QUAERO in French (Névéol et al., 2014) or GGPONC in German (Borchert et al., 2022), CLT cannot be evaluated on these datasets as there is no English counterpart annotated with the same guidelines, which is a pre-requisite for CLT evaluation. In order to tackle this issue, Frei et al. (2022) have introduced a small test dataset of 30 German medical sentences from Electronic Health Records (EHR) annotated in the same way as the n2c2 dataset and have used it to assess the performances of their GERNERMED++ model. Following their path we propose to release a medical NER dataset in French based on drug prescriptions.

3 Method and models

We now describe the different methods we compare to perform NER without any annotations in the target language: cross-lingual transfer and two translation-based methods, where either the train set or the test set is translated. All of them only use the English annotations from the n2c2 dataset (Track 2, Adverse Drug Events and Medication Extraction) (Henry et al., 2019), which is an English dataset of medical entities extracted from EHR.

3.1 Cross-lingual transfer (CLT)

The most intuitive way to perform NER without annotations in the target language is CLT which has shown impressive cross-lingual performances (Wu and Dredze, 2019). In our setting, we fine-tuned a multilingual large language model to perform NER on the English n2c2 dataset and evaluate on French and German test sets. In our experiments, XLM-R Base is preferred over mBERT as it has the same number of layers but outperforms it on multilingual benchmarks. XLM-R Large is also used to evaluate the impact of model size and distilmBERT, a smaller MLLM obtained by distillation of mBERT, is used to give insights about what is possible with less resources.

With CLT, these models will only see English NER labels during training and will be evaluated on a German and a French NER test set.

3.2 `translate-train`

The `translate-train` approach consists in constructing a translated version of the n2c2 dataset and in training a NER algorithm in French or German on the translated dataset.

The creation of the synthetic translated dataset is done in two steps. First the whole dataset is translated to the target language using a machine translation algorithm (Tiedemann and Thottingal, 2020; Ng et al., 2020). Then the labels must be aligned, which means identifying in the translated sentences the spans of text corresponding to the original English annotations. This task is tackled using either a statistical model `fast_align` (Dyer et al., 2013) or the neural algorithm `awesome-align` (Dou and Neubig, 2021). These alignment tools are applied to the original English sentence and its translation. They provide a mapping between the words of both sentences which is used to transfer the English annotations to the translated sentences. There are cases, namely in German, where the order of the words in a sentence differ from English, when an English annotation corresponds to several disjoint groups of words in the target language. For example, the sentence "She received an additional three units PRBC overnight"[1] is translated to "Sie erhielt über Nacht drei weitere PRBC-Einheiten" and the entity "three units" which is translated by "drei Einheiten" is split into two disjoint words. In those cases, all parts of the split entities have been labeled as the original entity.

This method is similar to the one proposed by Frei et al. (2022) and we improve over it by fine-tuning the translation and alignment models on a corpus of parallel medical texts. The influence of using fine-tuned models for translation and alignment is studied in Section 6.2.

3.3 `translate-test`

The `translate-test` method consists in translating the data into English at inference time and in applying an English NER model on it. The labels obtained with the NER models can be used to recover the entities in the original text with an alignment algorithm. The major drawback of this method is that it requires to translate the

[1] PRBC stands for "Packed Red Blood Cells".

text at inference time while the translation in the `translate-train` method occurs only once during training.

4 Evaluation data

While the model is trained on the English dataset n2c2 or a translation of it, it must be evaluated on French and German data, annotated similarly to n2c2 to assess its cross-lingual adaptation abilities.

4.1 The MedNERF dataset

We release MedNERF[2], a Medical NER dataset in the French language. It has been built using a sample of French medical prescriptions annotated with the same guidelines as the n2c2 dataset.

Sentences containing dosage instructions were obtained from a private set of scanned typewritten drug prescriptions. After anonymization of the drug prescriptions we used a state-of-the-art Optical Character Recognition (OCR) software[3]. We then discarded low quality sentences from the output of the OCR and we manually identified the sentences containing dosage instructions. For the purpose of this paper only 100 sentences have been randomly sampled and made public through the MedNERF dataset, which is intended to be a test and not a training dataset.

The annotations of the medical sentences use the n2c2 labels `DRUG`, `STRENGTH`, `FREQUENCY`, `DURATION`, `DOSAGE` and `FORM`. We did not use the `ADE` (Adverse Drug Event) label since it is very rare that such entities are present in drug prescriptions. We also discarded the `ROUTE` and `REASON` labels as in (Frei et al., 2022) because of either their ambiguous definition or the lack of diversity of the matching samples. A total of 406 entities were annotated in 100 sentences (cf. Table 1)[4].

NER Tag	Count
DRUG	67
STRENGTH	51
FREQUENCY	76
DURATION	43
DOSAGE	76
FORM	93
Total	406

Table 1: Distribution of labels in MedNERF.

[2]The dataset is available at `https://huggingface.co/datasets/Posos/MedNERF`.

[3]`https://cloud.google.com/vision`

[4]Randomly sampled examples in Appendix G.

4.2 The GERNERMED test dataset

The evaluation of the different cross-lingual adaptation techniques in German is done using the GERNERMED test set released by Frei et al. (2022), which consists of 30 sentences from physicians annotated with the same guidelines as n2c2. Table 2 provides statistics about the different datasets used in this paper.

dataset	lang.	sent.	entities
n2c2	en	16,656	65,495
GERNERMED-test	de	30	119
MedNERF	fr	100	406

Table 2: Statistics about the datasets.

5 Pre-selecting translation and alignment

The translation-based methods require a translation and an alignment models. We present in this section how we fine-tuned translation and alignment algorithms and how we chose which algorithms to use in our experiments. The choice of these algorithms can be seen as an hyper-parameter and for fair comparison with CLT, the selection should not be based on downstream cross-lingual abilities as this would mean cross-lingual supervision.

5.1 Translation models

We perform the automated translation of the n2c2 dataset from English to French and German with the following transformer-based machine translation algorithms: Opus-MT (Tiedemann and Thottingal, 2020) and FAIR (Ng et al., 2020) which we fine-tuned on a corpus of bilingual medical texts proposed in the BioWMT19 challenge [5] (Bawden et al., 2019). We used the UFAL dataset, which is a collection of medical and general domain parallel corpora in 8 languages paired with English, and Medline which is a dataset containing the titles and abstracts of scientific publications from Pubmed in English and a foreign language.

Since the UFAL dataset is orders of magnitude larger than Medline, we downsampled it to have equal proportions of sentences coming from Medline, from the medical part of UFAL and from UFAL general data. This resulted in approximately 90k sentences for the German translation models and 164k sentences for translation into French.

[5]The links of the datasets of the BioWMT19 challenge are available on its page `https://www.statmt.org/wmt19/biomedical-translation-task.html`

We fine-tuned the Opus-MT model (Tiedemann and Thottingal, 2020) for translation to French and German and the FAIR model (Ng et al., 2020) only for translation to German as no version of it is available in French.

The quality of the different translation models is measured on the Medline test set of the BioWMT 19 challenge and on the Khresmoi dataset (Dušek et al., 2017). Results are presented in Tables 3 and 4.

model	BioWMT19		Khresmoi	
	BLEU	COMET	BLEU	COMET
FAIR	<u>32.8</u>	0.628	**33.7**	**0.667**
+ ft	**34.2**	**0.734**	<u>32.4</u>	<u>0.666</u>
Opus	32.2	0.651	<u>32.4</u>	0.608
+ ft	32.5	<u>0.700</u>	30.5	0.619

Table 3: Evaluation of the translation models from English to German. Best model in bold and second underlined. ft for finetuned.

model	BioWMT19		Khresmoi	
	BLEU	COMET	BLEU	COMET
Opus	35.9	0.672	**48.0**	**0.791**
+ ft	**36.7**	**0.786**	46.5	**0.791**

Table 4: Evaluation of the translation models from English to French.

The analysis of these results lead us to choose the FAIR fine-tuned model as the best translation model for English to German translation and the Opus-MT fine-tuned model as the best translation model for English to French translation.

The same pre-selection evaluation was performed for the `translate-test` approach, with translation models from German and French to English. The models were fine-tuned on the same parallel dataset and similar results led to the same choice of best translation models (cf. Appendix A).

5.2 Alignment models

`fast_align` and `awesome-align` are two popular choices for word alignment (Yarmohammadi et al., 2021). Since `awesome-align` can be fine-tuned on parallel data we use the data used for translation to fine-tune the alignment model.

Choosing the right alignment models for the task can be tricky. While parallel corpora for fine-tuning `awesome-align` might be available in several languages and domains, annotated word alignment on parallel data is more scarce. In our case, annotated word alignment test data is not available in the clinical domain. The best alignment models can thus only be selected based on performance on a general-domain dataset. `awesome-align` pre-trained on general-domain data is preferred in French, and the same model with further fine-tuning on biomedical data is selected for German.

model	fr	de
FastAlign	10.5	27.0
AWESOME from scratch	5.6	17.4
+ ft on clinical	4.7	15.4
AWESOME pre-trained	**4.1**	15.2
+ ft on clinical	4.8	**15.0**

Table 5: Average Error Rate (AER) for various aligners.

Table 6 summarizes the choices of the best translation and alignment methods.

lang	translation	alignment
fr	Opus ft	AWESOME
de	FAIR ft	AWESOME pt+ft

Table 6: Pre-selected translation and alignment models.

6 Results

Having selected the best translation and alignment model for each language based on intrinsic evaluation, `translate-train` and `translate-test` approaches can be compared to Cross-lingual Transfer (CLT). The impact of the translation and alignment models can also be analyzed, as well as using the translation fine-tuning data to improve directly the models used in CLT. Since `translate-train` can leverage monolingual domain-specific model, we evaluate the outcome of such a strategy. Five different random seeds were used for each model and results presented in this section show the average performance along with the standard deviation (more implementation details in Appendix B).

6.1 Comparison of the different methods

For a fair comparison between the three methods detailed in Section 3, we use the pre-selected translation and alignment models of Table 6 for the `translate-train` and `translate-test` methods. We report the F1-scores of the different methods in Table 7. The translation and alignment models providing the best test scores are also

method	fr	de
distilmBERT		
CLT	$65.9_{\pm 3.3}$	$64.6_{\pm 2.4}$
translate-train select.	$66.5_{\pm 1.9}$	$\mathbf{68.3}_{\pm 1.3}$
translate-test select.	$\mathbf{69.2}_{\pm 1.4}$	$\mathbf{68.3}_{\pm 1.8}$
translate-train best	$\underline{69.2}_{\pm 1.2}$	$\underline{69.2}_{\pm 1.2}$
translate-test best	$\underline{69.7}_{\pm 1.5}$	$\underline{68.3}_{\pm 1.8}$
XLM-R Base		
CLT	$\mathbf{79.1}_{\pm 0.8}$	$72.2_{\pm 0.7}$
translate-train select.	$74.6_{\pm 0.9}$	$\mathbf{73.7}_{\pm 0.9}$
translate-test select.	$74.2_{\pm 1.6}$	$72.7_{\pm 0.8}$
translate-train best	$78.6_{\pm 0.5}$	$\underline{74.8}_{\pm 1.0}$
translate-test best	$74.4_{\pm 1.3}$	$72.7_{\pm 0.8}$
XLM-R Large		
CLT	$\mathbf{77.9}_{\pm 1.7}$	$\mathbf{78.5}_{\pm 0.4}$
translate-train select.	$76.5_{\pm 0.7}$	$77.4_{\pm 1.3}$
translate-test select.	$75.3_{\pm 0.9}$	$76.1_{\pm 2.8}$
translate-train best	$\underline{78.0}_{\pm 0.5}$	$\underline{79.4}_{\pm 1.3}$
translate-test best	$75.3_{\pm 0.9}$	$76.1_{\pm 2.8}$

Table 7: Comparing the three methods with pre-selected translation and alignment models (select.). Best performing pairs are provided for comparison and are underlined when better than CLT.

provided for comparison, revealing what can be missed with the pre-selection.

CLT with a sufficiently large MLLM provides the best results. When compared with translation-based methods with pre-selected translation and alignment models, CLT with XLM-R models gives higher scores, except for XLM-R Base in German.

On the other hand, it seems that using an English NER on a translated version of the test set provides the best results with a small model like distilmBERT. DistilmBERT might be better as a monolingual model than a multilingual one. In the same vein, XLM-R Base struggles in German, while its large version does not. Small language models underperform in CLT and their generalization ability is not sufficient compared to translation-based methods. A first take-away is consequently that translation-based methods should be favored with small language models.

The `translate-test` method is consistently outperformed by `translate-train` for large models. Even using a specific biomedical model like (Gu et al., 2021) for `translate-test` does not improve the results (results in Appendix F). Indeed translation and alignment errors only harm the training set of `translate-train`, which does not prevent a large model from generalizing despite some errors in the training data, while errors of translation or alignment in `translate-test` are directly reflected in the test score. In the rest of this analysis we will consequently compare CLT only with `translate-train`.

Providing a large-enough MLLM, CLT outperforms pre-selected `translate-train` and `translate-test`. However, choosing the translation and the alignment model beforehand does not lead to the best results. To the exception of XLM-R Base in French, there always exists a pair of translation and alignment models that leads to a better score for the `translate-train` method over CLT. This agrees with Yarmohammadi et al. (2021) and encourages practitioners to explore different methods to perform cross-lingual adaptation.

6.2 Influence of the translation model

The choice of the translation and alignment models can have an important impact on the final NER performances as shown on Table 7. This section studies their impact in details. A German NER model was trained using the Opus model instead of the FAIR model for translating the training set into German. Using a worse model (see Table 3) for translation leads to lower NER scores as shown in Table 8: the NER model based on the FAIR translation beats by more than 2 points the one using the Opus translation, whatever aligner is used.

While choosing between different base translation models (like Opus or FAIR) based on their translation scores on in-domain data seems to provide the best results, deciding between the fine-tuned version of a translation model and the base one by comparing the BLEU or COMET scores on biomedical data does not guarantee the best downstream F1 score as Table 9 shows. The translation model was fine-tuned on biomedical data, which improved intrinsic results on the BioWMT19 translation dataset. But this dataset belongs to a specific biomedical sub-domain (PubMed abstracts), and fine-tuning might not improve translation for the clinical sub-domain of the NER dataset.

The takeaway is that, while a small gain in translation accuracy (obtained with further fine-tuning) might not necessarily improve the result of the `translate-train` approach, a completely different model (like FAIR with respect to Opus) has more chance to improve cross-lingual adaptation.

aligner	Opus f1	FAIR f1
FastAlign	$70.9_{\pm1.8}$	$\mathbf{72.8}_{\pm1.6}$
AWESOME	$72.2_{\pm1.7}$	$\mathbf{73.1}_{\pm1.3}$
AWESOME ft	$71.1_{\pm1.2}$	$\mathbf{74.1}_{\pm1.1}$
AWESOME pt+ft	$71.2_{\pm1.1}$	$\mathbf{74.1}_{\pm1.3}$

Table 8: `translate-train` in German with XLM-R Base using either fine-tuned or base Opus model.

aligner	base	fine-tuned
FastAlign	$78.2_{\pm0.8}$	$\mathbf{78.6}_{\pm0.5}$
AWESOME	$\mathbf{76.4}_{\pm2.0}$	$74.2_{\pm0.9}$
AWESOME ft	$\mathbf{74.6}_{\pm1.6}$	$74.5_{\pm1.6}$
AWESOME pt+ft	$75.8_{\pm1.7}$	$\mathbf{76.3}_{\pm1.0}$

Table 9: `translate-train` in French with XLM-R Base using either fine-tuned or base Opus model.

6.3 Influence of the alignment model

While choosing a translation model based solely on intrinsic performance should not harm downstream cross-lingual adaptation performances, the choice of the alignment model seems more tricky. Based on intrinsic performances like Error Rate on annotated alignment (Table 5), `awesome-align` seems to be the right aligner for the task. However, while it provides better downstream results than `fast_align` in German (Table 8), it does not hold for French (Table 9).

Table 10 shows that using different aligners leads to different levels of accuracy according to the types of entity we want to retrieve. While the global F1 score suggests that `fast_align` is better suited for the cross-lingual adaptation, looking at the detailed results for each entity type shows that the gap is mainly due to the `FREQUENCY` class on which `awesome-align` performs poorly. But this is not the case on other classes.

`FREQUENCY` entities are usually more verbose than drugs or dosages. Table 11 shows that `fast_align` make obvious errors like aligning "240 mg" to "in morning and night", but

aligner	Freq.	Strength	Drug	f1
FastAlign	**72.0**	89.7	**83.2**	**78.6**
AWESOME	50.2	**92.5**	82.2	74.2

Table 10: Comparison of `fast_align` and `awesome-align` (pre-trained only) for three different entity types (F1-score), for `translate-train` with XLM-R Base on MedNERF with Opus fine-tuned. (Full results in Appendix F).

original	FastAlign	AWESOME
in morning and night	le matin et la nuit et 240 mg	**le matin et la nuit**
daily	**par jour**	jour
once a day	**une fois par jour**	fois par jour
at bedtime	**au moment du coucher**	au // coucher

Table 11: Examples of frequencies transformed with translation and alignment. Bold indicates the right annotation and // indicates that the entity has been split.

`awesome-align` can miss the preposition when aligning "daily" with "jour" instead of "par jour", leading eventually to a consequent score drop.

The choice of the alignment model must thus be made more carefully than the translation one. Intrinsic performances of alignment models are not sufficient information. Some additional post-processing might be needed, as in Yarmohammadi et al. (2021), where `awesome-align` gives better results, but entities that are split by the aligner like "au moment du coucher" in Table 11 are merged by including all words in between. This would work in that particular case, but could cause problems in others, particularly for languages where the word order is different.

6.4 Using parallel data to realign models

With the right translation and alignment model, it seems that CLT can be outperformed by the `translate-train` method. However the latter relies on additional resources: a translation and an alignment models, trained on parallel data. This parallel data could also be used to re-align the representations of the multilingual models used in CLT.

To improve a multilingual language model with parallel data, it is trained for a contrastive alignment objective following Wu and Dredze (2020). Words aligned with `awesome-align` are trained to have more similar representations than random in-batch pairs of words (details in Appendix E). After this realignment step, CLT can be applied.

Results in Table 12 show that while realignment does not systematically provide improvement over CLT as observed by Wu and Dredze (2020), it does significantly boost results in some cases, allowing to outperform the best `translate-train` baseline in German for XLM-R Base and in French for XLM-R Large. This, yet again, encourages practitioners to explore different methods, including realignment to perform cross-lingual adaptation.

model	fr	de
distilmBERT	65.9$_{\pm 3.3}$	64.6$_{\pm 2.4}$
+ realign	<u>66.4</u>$_{\pm 1.4}$	<u>67.9</u>$_{\pm 1.5}$
translate-train best	**69.2**$_{\pm 1.2}$	**69.2**$_{\pm 1.2}$
XLM-R Base	**79.1**$_{\pm 0.8}$	72.2$_{\pm 0.7}$
+ realign	76.7$_{\pm 0.7}$	<u>**75.8**</u>$_{\pm 1.3}$
translate-train best	78.6$_{\pm 0.5}$	74.8$_{\pm 1.0}$
XLM-R Large	77.9$_{\pm 1.7}$	78.5$_{\pm 0.4}$
+ realign	<u>**78.8**</u>$_{\pm 1.6}$	78.3$_{\pm 1.6}$
translate-train best	78.0$_{\pm 0.5}$	**79.4**$_{\pm 1.3}$

Table 12: F1 scores for CLT from scratch and CLT with realignment. Best F1-score in bold. Results underlined show improvement of realignment over CLT.

6.5 Using domain-specific language models

We evaluate now the relevance of using language-specific models like CamemBERT (Martin et al., 2020) or GottBERT (Scheible et al., 2020) on the translated version of the training dataset or language and domain-specific models like DrBERT (Labrak et al., 2023) or medBERT.de (Bressem et al., 2023) which are BERT models fine-tuned on medical corpora in respectively French and German. We report in Table 13 and 14 the results of the `translate-train` method for the best translation/alignment algorithms pair and for the pre-selected one, compared to using XLM-R Base.

model	pre-selected	best
CamemBERT Base	73.5$_{\pm 1.5}$	76.7$_{\pm 0.9}$
DrBERT 7GB	70.7$_{\pm 1.3}$	73.5$_{\pm 1.4}$
DrBERT Pubmed	**76.1**$_{\pm 1.3}$	**78.8**$_{\pm 1.4}$
XLM-R Base	74.6$_{\pm 1.9}$	78.6$_{\pm 0.5}$

Table 13: Comparison of domain and language specific models for translate-train in French.

model	pre-selected	best
GottBERT	**75.5**$_{\pm 1.4}$	**76.6**$_{\pm 0.8}$
medBERT	72.7$_{\pm 0.5}$	75.0$_{\pm 1.6}$
XLM-R Base	73.7$_{\pm 0.9}$	74.8$_{\pm 1.0}$

Table 14: Comparison of domain and language specific models for translate-train in German.

The `translate-train` approach allows to rely on models that are specific to the language and domain of the target evaluation. However, Table 13 and 14 show that their use does not always bring significant improvement over XLM-R Base. The performances of these models can be explained by the quantity of training data used. XLM-R models are indeed trained on 2.5 TB data while DrBERT and medBERT.de use less than 10GB data, which can explain their low score. Besides, the language-specific models CamemBERT and GottBERT are trained with more data (138 GB and 145 GB) and achieve better performances, even beating XLM-R in German. Finally, it must be noted that the best `translate-train` model in French, DrBERT Pubmed, is actually pre-trained on the English PubMed dataset and then on French clinical texts, which suggests that multilingual models should be preferred, even with a translation-based cross-lingual adaptation.

6.6 Computing times

To conclude the analysis of the different cross-lingual adaptation methods studied in this paper we finally compare their computing times. Table 15 gathers the training and inference times of the three methods using the XLM-R base model and the `awesome-align` alignment model in French.

method	training time (total)	inference time (per sample)
CLT	1.2h	0.04s
translate-train	2.7h	0.04s
translate-test	1.2h	0.32s

Table 15: Training and inference times for the different methods, with the XLM-R Base model in French, on a single GPU.

This comparison shows a longer training time for the `translate-train` method, which is due to the translation of the whole training set before the training of the NER model. On a single GPU with 8GB of RAM this translation step is even longer than the NER training. However, once training is done the `translate-train` method has the same inference time as the CLT method, while the `translate-test` method now suffers from the need of translation at inference time.

7 Conclusion

This paper shows that cross-lingual transfer with general-domain MLLMs is efficient for a domain-specific task like clinical NER, giving comparable results with translating the training set. But CLT has the advantage of working off-the-shelf, while translation-based methods require choosing the translation and alignment models carefully. Selecting these models based on intrinsic domain-specific values, like fine-tuning scores on clinical parallel data, or using a domain-specific language model does not provide significantly better downstream results in the target language. The selection of the alignment model was shown to be particularly crucial, and results of `translate-train` could probably be improved by post-processing the alignment. CLT also has a margin of progression as realigning the representations of MLLMs can increase the results dramatically in some cases.

It is also worth noting that training on translated data provide better results than translating at inference time. The `translate-test` approach should then be used only when large multilingual models cannot be used. While training on translated data allows to leverage domain-specific monolingual language models, those latter models can give better results over multilingual models like XLM-R only if pre-trained with sufficient data.

Pre-training a MLLM with only clinical data is a good lead for further improvements in clinical cross-lingual transfer. While the results show that using a domain-specific monolingual model in `translate-train` or `translate-test` is not on par with general-purpose multilingual models, they also show that the French clinical model DrBERT provides the best results for `translate-train` when it uses the English biomedical model PubmedBERT as initialization.

We finally advocate for the release of more non-English clinical datasets annotated with similar guidelines as English (or other) ones. Even a relatively small dataset like MedNERF or the GERNERMED test set are crucial to evaluate cross-lingual adaptation in the clinical domain.

Limitations

This paper is limited to the study of clinical NER models using an encoder-only architecture. The use of generative models with a zero-shot learning approach (Hu et al., 2023) is another promising approach for low-resource languages that could be compared with CLT and translation-based approaches in a future work. However such methods require a careful prompt selection strategy and cannot be directly compared to supervised models.

This paper is also limited to cross-lingual transfer to French and German. Ideally, this work could have included experiments with other target languages and also other source languages than English, as Yarmohammadi et al. (2021) do in their general-domain comparison of strategies for cross-lingual transfer. However evaluation datasets are lacking for that purpose in the clinical domain. Similarly, more general conclusions about cross-lingual adaptation methods in the clinical domains could be drawn with further studies on various clinical NLP tasks such as relation extraction or sentence classification. However, the lack of evaluation datasets in the clinical domain prevented us from extending the experiments to such other clinical NLP tasks. Finally, the authors assume that the findings will be task-specific and encourage practitioners to explore all methods when facing a new NLP task.

The authors also want to point out that MedNERF is drawn from drug prescriptions while the n2c2 and GERNERMED datasets use clinical reports. This domain difference could have made the cross-lingual generalization more challenging, but in practice we found that the different models used were not really affected by the possible domain-shift, showing similar French and German F1 scores. Moreover, when comparing randomly sampled examples from all three datasets, we do not find any critical differences (see Appendix G). Sentences drawn from MedNERF are shorter and less written, but they contain similar annotated entities as the n2c2 sentences, and the n2c2 dataset also contains some short examples that resemble the MedNERF ones.

Acknowledgments

We would like to thank the anonymous reviewers for their comments which helped enrich the discussion around the results.

Some experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations[6].

References

Rachel Bawden, Kevin Bretonnel Cohen, Cristian Grozea, Antonio Jimeno Yepes, Madeleine Kittner, Martin Krallinger, Nancy Mah, Aurelie Neveol, Mariana Neves, Felipe Soares, Amy Siu, Karin Verspoor, and Maika Vicente Navarro. 2019. Findings of the WMT 2019 Biomedical Translation Shared Task: Evaluation for MEDLINE Abstracts and Biomedical

[6] see https://www.grid5000.fr.

Terminologies. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 29–53, Florence, Italy. Association for Computational Linguistics.

Florian Borchert, Christina Lohr, Luise Modersohn, Jonas Witt, Thomas Langer, Markus Follmann, Matthias Gietzelt, Bert Arnrich, Udo Hahn, and Matthieu-P. Schapranow. 2022. GGPONC 2.0 - the German clinical guideline corpus for oncology: Curation workflow, annotation policy, baseline NER taggers. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3650–3660, Marseille, France. European Language Resources Association.

Keno K Bressem, Jens-Michalis Papaioannou, Paul Grundmann, Florian Borchert, Lisa C Adams, Leonhard Liu, Felix Busch, Lina Xu, Jan P Loyen, Stefan M Niehues, et al. 2023. Medbert. de: A comprehensive german bert model for the medical domain. *arXiv preprint arXiv:2303.08179*.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Ondřej Dušek, Jan Hajič, Jaroslava Hlaváčová, Jindřich Libovický, Pavel Pecina, Aleš Tamchyna, and Zdeňka Urešová. 2017. Khresmoi summary translation test data 2.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Johann Frei, Ludwig Frei-Stuber, and Frank Kramer. 2022. GERNERMED++: Transfer Learning in German Medical NLP.

Johann Frei and Frank Kramer. 2021. GERNERMED – An Open German Medical NER Model.

Yingbo Gao, Christian Herold, Zijian Yang, and Hermann Ney. 2022. Is encoder-decoder redundant for neural machine translation? In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 562–574, Online only. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. 3(1).

Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2019. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Yan Hu, Iqra Ameer, Xu Zuo, Xueqing Peng, Yujia Zhou, Zehan Li, Yiming Li, Jianfu Li, Xiaoqian Jiang, and Hua Xu. 2023. Zero-shot clinical entity recognition using chatgpt. *arXiv preprint arXiv:2303.16416*.

Rezarta Islamaj Doğan and Zhiyong Lu. 2012. An improved corpus of disease mentions in PubMed citations. In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 91–99, Montréal, Canada. Association for Computational Linguistics.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Muhammad Raza Khan, Morteza Ziyadi, and Mohamed AbdelHady. 2020. Mt-bioner: Multi-task learning for biomedical named entity recognition using deep bidirectional transformers. *arXiv preprint arXiv:2001.08904*.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023. Drbert: A robust pre-trained model in french for biomedical and clinical domains. *bioRxiv*, pages 2023–04.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Aurélie Névéol, Cyril Grouin, Jérémy Leixa, Sophie Rosset, and Pierre Zweigenbaum. 2014. The quaero french medical corpus : A ressource for medical entity recognition and normalization.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2020. Facebook FAIR's WMT19 News Translation Task Submission. In *Proc. of WMT*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Raphael Scheible, Fabian Thomczyk, Patric Tippmann, Victor Jaravine, and Martin Boeker. 2020. Gottbert: a pure german language model. *arXiv preprint arXiv:2012.02110*.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Do explicit alignments robustly improve multilingual encoders? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4471–4482, Online. Association for Computational Linguistics.

Xi Yang, Jiang Bian, William R Hogan, and Yonghui Wu. 2020. Clinical concept extraction using transformers. *Journal of the American Medical Informatics Association*. Ocaa189.

Mahsa Yarmohammadi, Shijie Wu, Marc Marone, Haoran Xu, Seth Ebner, Guanghui Qin, Yunmo Chen, Jialiang Guo, Craig Harman, Kenton Murray, Aaron Steven White, Mark Dredze, and Benjamin Van Durme. 2021. Everything is all it takes: A multipronged strategy for zero-shot cross-lingual information extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1950–1967, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*.

A Pre-selection of the translation model for `translate-test`

The `translate-test` method needs translation algorithms from German and French to English. Similarly to Section 5.1 we fine-tuned the Opus-MT and FAIR algorithms on the same medical datasets and obtained the COMET and BLEU scores presented in Tables 16 and 17. These scores are used to select the best translation model for the `translate-test` approach and they lead to the same model choices as for the `translate-train` method.

model	BioWMT19		Khresmoi	
	BLEU	COMET	BLEU	COMET
FAIR	<u>38.2</u>	0.538	**47.1**	**0.764**
+ ft	**38.5**	**0.675**	<u>46.8</u>	**0.764**
Opus	35.3	0.587	43.6	0.723
+ ft	38.1	<u>0.640</u>	44.3	0.729

Table 16: Evaluation of the translation models from German to English. Best model bold and second underlined. ft for finetuned.

model	BioWMT19		Khresmoi	
	BLEU	COMET	BLEU	COMET
Opus	33.9	0.721	**48.3**	0.798
+ ft	**36.3**	**0.749**	48.0	**0.799**

Table 17: Evaluation of the translation models from French to English. Best model bold. ft for finetuned.

B Training and implementation details

All models were written in Pytorch using the Huggingface librairies (Wolf et al., 2020) and were fine-tuned using the AdamW optimizer (Loshchilov and Hutter, 2019) and a learning rate of $6 \cdot 10^{-6}$ with linear decay. We used 4 epochs for the translation models and 8 epochs for the NER models. The NER models were trained on a single GPU (Nvidia GeForce RTX 2070 with 8GB of RAM) for approximately one hour.

C Alignment methods used

`fast_align` was applied asymmetrically, by mapping words from source language (English) to target language. Although it might increase alignment score, symmetrization was not used, because it might remove important links during the labels projection step.

`awesome-align` was used with softmax (instead of the alternative α-entmax function) and without the optional consistency optimization objective. For completeness we added the results with the consistency optimization objective (w/ co lines) in the tables of Appendix F and we observed that they did not improve the NER scores. The base model used is mBERT as in the original paper (Dou and Neubig, 2021). The pre-trained version of `awesome-align` used is the one provided by the authors, fine-tuned on general-domain parallel data. Throughout the paper, in tables, "AWESOME" designates this latter pre-trained version. "AWESOME ft" is `awesome-align` with the raw mBERT model fine-tuned on the clinical parallel data only and "AWESOME pt+ft" is the pre-trained model, fine-tuned again on the clinical parallel data.

D Transformer-base models used

We report in Table 18 the number of parameters and the quantity of training data of the different large language models used in this paper.

model	params (M)	emb. (M)	train (GB)
Multilingual models			
distilmBERT	135	92	42
XLM-R Base	278	192	2.5k
XLM-R Large	560	256	2.5k
Language-specific models			
CamemBERT (fr)	111	25	138
GottBERT (de)	126	40	145
Clinical models			
medBERT (de)	109	23	10
DrBERT 7GB (fr)	111	25	7.4
DrBERT PubMed (fr)	109	23	28

Table 18: Size of the different base models.

Although distilmBERT has more parameters than CamemBERT, it must be noted that it has also more words in its vocabulary, due to its multilingual nature. Hence most of its parameters are embeddings weights that are not necessarily used in our experiments as they might be embeddings of words from other languages. So in our setting, distilmBERT can be considered a smaller model than CamemBERT and GottBERT despite the higher number of parameters.

E Realignment method

The reader might refer to Wu and Dredze (2020) for the realignment method itself. The representations of the last layer of the pre-trained model to be realigned were used in a contrastive loss where pairs of words aligned with `awesome-align` are encouraged to be more similar than the other possible pairs of words in the batch. The strong alignment objective was used, meaning that pair of same-language words were also used as negative examples for the contrastive loss. The version of `awesome-align` was the one pre-trained on general-domain data, released by Dou and Neubig (2021), with softmax and without the optional consistency optimization objective, the same used by the authors of the realignment method used.

The parallel data used for realignment was the same as for fine-tuning the translation and alignment models (Section 5.1). The two datasets (English-German and English-French) were used together to realign a given model, which can then be used either for generalization to French or German. For each base model, five realigned models were obtained for the five random seeds, each of them used in the corresponding fine-tuning by seed.

The realignment was done for 20,000 steps of batches of size 16, with Adam optimizer, a learning rate of 2×10^{-4}, and with linear warm-up for 10% of the total steps. This means that the whole dataset was repeated approximately 1.25 times.

F Additional results

Detailed results are shown in the following tables:

- Summary of results for cross-lingual adaptation, with pre-selected and best pairs of translation and alignment models: Table 19 for French and 20 for German;
- CLT and `translate-train` with multilingual models: Table 21 (fr) and 22 (de);
- `translate-train` with language- and domain-specific models: Table 23 and 24;
- `translate-test` with multilingual language models: Table 25 and 26;
- `translate-test` with PubmedBERT: Table 27 and 28;
- Breakdown of the results class-by-class in French for multilingual models: Table 29.

model	pre-selected	best
translate-train		
distilmBERT	$66.5_{\pm 1.9}$	$69.2_{\pm 1.2}$
XLM-R Base	$74.2_{\pm 0.9}$	$78.6_{\pm 0.5}$
XLM-R Large	$76.5_{\pm 0.7}$	$78.0_{\pm 0.5}$
CamemBERT	$73.5_{\pm 1.5}$	$76.7_{\pm 0.9}$
DrBERT	$70.7_{\pm 1.3}$	$73.5_{\pm 1.4}$
DrBERT Pubmed	$76.1_{\pm 1.3}$	$78.8_{\pm 1.4}$
translate-test		
distilmBERT	$69.2_{\pm 1.4}$	$69.7_{\pm 1.5}$
XLM-R Base	$74.2_{\pm 1.6}$	$74.4_{\pm 1.3}$
XLM-R Large	$75.3_{\pm 0.9}$	$75.3_{\pm 0.9}$
PubmedBERT	$73.3_{\pm 1.3}$	$73.5_{\pm 1.2}$
CTL*		
distilmBERT	$65.9_{\pm 3.3}$	$65.9_{\pm 3.3}$
+ realigned	$66.4_{\pm 1.4}$	$66.4_{\pm 1.4}$
XLM-R Base	$\mathbf{79.1}_{\pm 0.8}$	$\mathbf{79.1}_{\pm 0.8}$
+ realigned	$76.7_{\pm 0.7}$	$76.7_{\pm 0.7}$
XLM-R Large	$77.9_{\pm 1.7}$	$77.9_{\pm 1.7}$
+ realigned	$78.8_{\pm 1.6}$	$78.8_{\pm 1.6}$

Table 19: Summary of results for cross-lingual adaptation to French.

*results are reported twice as there is no pre-selection process

model	pre-selected	best
translate-train		
distilmBERT	$68.3_{\pm 1.3}$	$69.2_{\pm 1.2}$
XLM-R Base	$73.7_{\pm 0.9}$	$74.8_{\pm 1.0}$
XLM-R Large	$77.4_{\pm 1.3}$	$\mathbf{79.4}_{\pm 1.3}$
GottBERT	$75.5_{\pm 1.4}$	$76.6_{\pm 0.8}$
MedBERT.de	$72.7_{\pm 0.5}$	$75.0_{\pm 1.6}$
translate-test		
distilmBERT	$68.3_{\pm 1.8}$	$68.3_{\pm 1.8}$
XLM-R Base	$72.7_{\pm 0.8}$	$72.7_{\pm 0.8}$
XLM-R Large	$76.1_{\pm 2.8}$	$76.1_{\pm 2.8}$
PubmedBERT	$72.6_{\pm 1.5}$	$73.3_{\pm 1.7}$
CTL*		
distilmBERT	$64.6_{\pm 2.4}$	$64.6_{\pm 2.4}$
+ realigned	$67.9_{\pm 1.5}$	$67.9_{\pm 1.5}$
XLM-R Base	$72.2_{\pm 0.7}$	$72.2_{\pm 0.7}$
+ realigned	$75.8_{\pm 1.3}$	$75.8_{\pm 1.3}$
XLM-R Large	$\mathbf{78.5}_{\pm 0.4}$	$78.5_{\pm 0.4}$
+ realigned	$78.3_{\pm 1.6}$	$78.3_{\pm 1.6}$

Table 20: Summary of results for cross-lingual adaptation to German.

*results reported twice as there is no pre-selection process

translation	aligner	precision	recall	micro-f1	macro-f1
		distilmBERT			
Opus	FastAlign	$67.8_{\pm 2.1}$	$68.7_{\pm 1.5}$	$68.3_{\pm 1.6}$	$70.7_{\pm 1.4}$
Opus	AWESOME w/o co	$67.1_{\pm 1.0}$	$68.7_{\pm 1.3}$	$67.9_{\pm 0.3}$	$70.4_{\pm 0.3}$
Opus	AWESOME w/ co	$\mathbf{68.2}_{\pm 1.6}$	$70.1_{\pm 1.1}$	$\mathbf{69.2}_{\pm 1.2}$	$\mathbf{71.7}_{\pm 1.1}$
Opus	AWESOME ft w/o co	$65.0_{\pm 0.8}$	$67.4_{\pm 1.8}$	$66.2_{\pm 1.0}$	$68.8_{\pm 0.8}$
Opus	AWESOME ft w/ co	$65.2_{\pm 1.2}$	$68.4_{\pm 1.8}$	$66.8_{\pm 1.3}$	$69.4_{\pm 1.1}$
Opus	AWESOME pt+ft w/o co	$66.5_{\pm 1.3}$	$69.1_{\pm 1.2}$	$67.8_{\pm 1.0}$	$70.2_{\pm 0.9}$
Opus	AWESOME pt+ft w/ co	$66.6_{\pm 1.2}$	$\mathbf{70.2}_{\pm 1.8}$	$68.3_{\pm 1.1}$	$70.8_{\pm 1.1}$
Opus ft	FastAlign	$66.0_{\pm 1.1}$	$69.5_{\pm 0.4}$	$67.7_{\pm 0.5}$	$70.0_{\pm 0.5}$
Opus ft	AWESOME w/o co	$65.2_{\pm 1.9}$	$67.8_{\pm 1.9}$	$66.5_{\pm 1.9}$	$69.1_{\pm 1.7}$
Opus ft	AWESOME w/ co	$65.7_{\pm 1.2}$	$68.3_{\pm 1.9}$	$66.9_{\pm 1.4}$	$69.9_{\pm 1.2}$
Opus ft	AWESOME ft w/o co	$64.6_{\pm 0.8}$	$67.4_{\pm 1.0}$	$65.9_{\pm 0.4}$	$68.6_{\pm 0.5}$
Opus ft	AWESOME ft w/ co	$64.9_{\pm 0.7}$	$68.8_{\pm 1.5}$	$66.8_{\pm 1.1}$	$69.4_{\pm 1.0}$
Opus ft	AWESOME pt+ft w/o co	$64.5_{\pm 1.1}$	$68.1_{\pm 1.4}$	$66.2_{\pm 1.0}$	$69.0_{\pm 1.0}$
Opus ft	AWESOME pt+ft w/ co	$63.5_{\pm 1.0}$	$68.3_{\pm 1.1}$	$65.8_{\pm 0.9}$	$68.7_{\pm 1.0}$
Cross-lingual Transfer		$64.9_{\pm 2.5}$	$66.8_{\pm 4.1}$	$65.9_{\pm 3.3}$	$68.2_{\pm 3.2}$
Cross-lingual Transfer with realignment		$68.1_{\pm 2.2}$	$64.9_{\pm 1.0}$	$66.4_{\pm 1.4}$	$67.4_{\pm 1.6}$
		XLM-R Base			
Opus	FastAlign	$77.4_{\pm 0.9}$	$79.1_{\pm 0.7}$	$78.2_{\pm 0.8}$	$79.8_{\pm 0.6}$
Opus	AWESOME w/o co	$75.4_{\pm 2.4}$	$77.3_{\pm 1.6}$	$76.4_{\pm 2.0}$	$78.6_{\pm 1.6}$
Opus	AWESOME w/ co	$76.0_{\pm 1.0}$	$77.6_{\pm 1.6}$	$76.8_{\pm 1.2}$	$78.8_{\pm 1.2}$
Opus	AWESOME ft w/o co	$73.8_{\pm 1.7}$	$75.4_{\pm 1.6}$	$74.6_{\pm 1.6}$	$76.8_{\pm 1.4}$
Opus	AWESOME ft w/ co	$75.0_{\pm 0.8}$	$76.9_{\pm 0.6}$	$76.0_{\pm 0.5}$	$78.2_{\pm 0.6}$
Opus	AWESOME pt+ft w/o co	$74.8_{\pm 1.5}$	$76.8_{\pm 2.0}$	$75.8_{\pm 1.7}$	$78.0_{\pm 1.7}$
Opus	AWESOME pt+ft w/ co	$75.1_{\pm 1.3}$	$76.8_{\pm 1.6}$	$76.0_{\pm 1.4}$	$78.2_{\pm 1.2}$
Opus ft	FastAlign	$77.9_{\pm 0.2}$	$79.4_{\pm 1.0}$	$78.6_{\pm 0.5}$	$80.3_{\pm 0.3}$
Opus ft	AWESOME w/o co	$72.9_{\pm 1.4}$	$75.6_{\pm 0.8}$	$74.2_{\pm 0.9}$	$76.7_{\pm 0.7}$
Opus ft	AWESOME w/ co	$74.6_{\pm 1.0}$	$76.6_{\pm 0.8}$	$75.6_{\pm 0.9}$	$77.9_{\pm 0.7}$
Opus ft	AWESOME ft w/o co	$73.4_{\pm 1.7}$	$75.6_{\pm 1.6}$	$74.5_{\pm 1.6}$	$77.0_{\pm 1.4}$
Opus ft	AWESOME ft w/ co	$74.2_{\pm 1.9}$	$76.7_{\pm 2.2}$	$75.5_{\pm 2.0}$	$78.0_{\pm 1.7}$
Opus ft	AWESOME pt+ft w/o co	$75.1_{\pm 1.0}$	$77.5_{\pm 1.1}$	$76.3_{\pm 1.0}$	$78.5_{\pm 0.9}$
Opus ft	AWESOME pt+ft w/ co	$75.1_{\pm 1.9}$	$77.6_{\pm 1.7}$	$76.3_{\pm 1.8}$	$78.7_{\pm 1.5}$
Cross-lingual Transfer		$\mathbf{78.7}_{\pm 1.8}$	$\mathbf{79.6}_{\pm 0.5}$	$\mathbf{79.1}_{\pm 0.8}$	$\mathbf{80.9}_{\pm 0.9}$
Cross-lingual Transfer with realignment		$76.9_{\pm 1.4}$	$76.6_{\pm 0.2}$	$76.7_{\pm 0.7}$	$78.9_{\pm 0.8}$
		XLM-R Large			
Opus	FastAlign	$78.8_{\pm 0.7}$	$77.2_{\pm 0.6}$	$78.0_{\pm 0.5}$	$79.8_{\pm 0.5}$
Opus	AWESOME w/o co	$76.9_{\pm 1.1}$	$76.1_{\pm 1.8}$	$76.5_{\pm 1.2}$	$78.7_{\pm 1.1}$
Opus	AWESOME w/ co	$76.5_{\pm 1.2}$	$75.3_{\pm 1.5}$	$75.9_{\pm 1.3}$	$78.2_{\pm 1.1}$
Opus	AWESOME ft w/o co	$74.6_{\pm 1.0}$	$73.9_{\pm 0.4}$	$74.2_{\pm 0.6}$	$76.8_{\pm 0.5}$
Opus	AWESOME ft w/ co	$74.9_{\pm 0.3}$	$75.7_{\pm 0.8}$	$75.3_{\pm 0.3}$	$78.1_{\pm 0.5}$
Opus	AWESOME pt+ft w/o co	$75.8_{\pm 1.4}$	$75.0_{\pm 1.0}$	$75.4_{\pm 1.0}$	$78.0_{\pm 0.7}$
Opus	AWESOME pt+ft w/ co	$75.7_{\pm 2.2}$	$76.2_{\pm 1.3}$	$75.9_{\pm 1.7}$	$78.3_{\pm 1.8}$
Opus ft	FastAlign	$76.2_{\pm 2.0}$	$76.9_{\pm 2.5}$	$76.6_{\pm 2.1}$	$78.3_{\pm 2.0}$
Opus ft	AWESOME w/o co	$76.1_{\pm 1.0}$	$76.9_{\pm 0.8}$	$76.5_{\pm 0.7}$	$78.9_{\pm 0.5}$
Opus ft	AWESOME w/ co	$74.6_{\pm 1.5}$	$76.0_{\pm 1.3}$	$75.3_{\pm 0.9}$	$77.9_{\pm 0.6}$
Opus ft	AWESOME ft w/o co	$74.1_{\pm 1.4}$	$75.5_{\pm 0.4}$	$74.8_{\pm 0.8}$	$77.4_{\pm 0.8}$
Opus ft	AWESOME ft w/ co	$75.5_{\pm 1.9}$	$75.6_{\pm 1.2}$	$75.5_{\pm 1.4}$	$78.1_{\pm 1.2}$
Opus ft	AWESOME pt+ft w/o co	$75.1_{\pm 0.5}$	$76.0_{\pm 1.2}$	$75.5_{\pm 0.6}$	$78.1_{\pm 0.5}$
Opus ft	AWESOME pt+ft w/ co	$75.0_{\pm 1.8}$	$75.8_{\pm 0.8}$	$75.4_{\pm 1.1}$	$77.7_{\pm 1.0}$
Cross-lingual Transfer		$78.2_{\pm 2.5}$	$77.6_{\pm 1.2}$	$77.9_{\pm 1.7}$	$80.0_{\pm 1.4}$
Cross-lingual Transfer with realignment		$\mathbf{79.7}_{\pm 1.6}$	$\mathbf{77.9}_{\pm 1.8}$	$\mathbf{78.8}_{\pm 1.6}$	$\mathbf{80.8}_{\pm 1.4}$

Table 21: Cross-lingual Transfer and `translate-train` results in French for multilingual base models.

translation	aligner	precision	recall	micro-f1	macro-f1
		distilmBERT			
FAIR	FastAlign	$69.0_{\pm1.2}$	$63.9_{\pm0.5}$	$66.3_{\pm0.8}$	$44.2_{\pm2.4}$
FAIR	AWESOME w/o co	$68.1_{\pm1.6}$	$66.1_{\pm1.3}$	$67.1_{\pm1.3}$	$50.5_{\pm3.0}$
FAIR	AWESOME w/ co	$68.0_{\pm2.0}$	$\mathbf{67.4}_{\pm1.4}$	$67.7_{\pm1.0}$	$50.6_{\pm3.7}$
FAIR	AWESOME ft w/o co	$68.5_{\pm3.6}$	$65.0_{\pm2.2}$	$66.6_{\pm1.2}$	$48.3_{\pm4.0}$
FAIR	AWESOME ft w/ co	$67.6_{\pm2.4}$	$66.1_{\pm3.5}$	$66.7_{\pm1.4}$	$46.4_{\pm5.1}$
FAIR	AWESOME pt+ft w/o co	$69.6_{\pm2.7}$	$65.9_{\pm2.8}$	$67.6_{\pm1.1}$	$46.8_{\pm6.4}$
FAIR	AWESOME pt+ft w/ co	$68.8_{\pm2.6}$	$66.6_{\pm3.7}$	$67.5_{\pm1.4}$	$49.2_{\pm6.0}$
FAIR ft	FastAlign	$70.5_{\pm1.9}$	$65.2_{\pm1.4}$	$67.7_{\pm1.3}$	$\mathbf{52.2}_{\pm1.5}$
FAIR ft	AWESOME w/o co	$69.3_{\pm2.0}$	$66.9_{\pm1.6}$	$68.0_{\pm1.1}$	$48.0_{\pm3.6}$
FAIR ft	AWESOME w/ co	$68.6_{\pm2.8}$	$66.4_{\pm2.0}$	$67.4_{\pm1.5}$	$47.4_{\pm4.0}$
FAIR ft	AWESOME ft w/o co	$70.9_{\pm2.7}$	$67.2_{\pm2.1}$	$69.0_{\pm1.9}$	$49.8_{\pm3.7}$
FAIR ft	AWESOME ft w/ co	$71.4_{\pm2.0}$	$67.2_{\pm1.8}$	$\mathbf{69.2}_{\pm1.2}$	$49.0_{\pm4.3}$
FAIR ft	AWESOME pt+ft w/o co	$69.3_{\pm1.8}$	$\mathbf{67.4}_{\pm1.9}$	$68.3_{\pm1.3}$	$49.2_{\pm3.2}$
FAIR ft	AWESOME pt+ft w/ co	$70.4_{\pm1.9}$	$67.1_{\pm1.6}$	$68.7_{\pm1.2}$	$49.5_{\pm3.1}$
Cross-lingual Transfer		$62.8_{\pm2.9}$	$66.4_{\pm2.0}$	$64.6_{\pm2.4}$	$46.3_{\pm3.1}$
Cross-lingual Transfer with realignment		$\mathbf{72.0}_{\pm2.3}$	$64.2_{\pm1.4}$	$67.9_{\pm1.5}$	$46.5_{\pm5.2}$
		XLM-R Base			
FAIR	FastAlign	$74.0_{\pm2.6}$	$71.6_{\pm1.4}$	$72.8_{\pm1.6}$	$55.5_{\pm5.7}$
FAIR	AWESOME w/o co	$73.3_{\pm1.8}$	$72.9_{\pm0.8}$	$73.1_{\pm1.3}$	$53.2_{\pm3.8}$
FAIR	AWESOME w/ co	$73.9_{\pm2.8}$	$72.8_{\pm2.2}$	$73.3_{\pm2.3}$	$53.1_{\pm4.4}$
FAIR	AWESOME ft w/o co	$74.1_{\pm1.1}$	$74.1_{\pm1.1}$	$74.1_{\pm1.1}$	$57.5_{\pm4.3}$
FAIR	AWESOME ft w/ co	$75.4_{\pm2.3}$	$74.1_{\pm1.4}$	$74.8_{\pm1.8}$	$58.0_{\pm4.2}$
FAIR	AWESOME pt+ft w/o co	$74.6_{\pm1.8}$	$73.6_{\pm1.1}$	$74.1_{\pm1.3}$	$56.3_{\pm2.2}$
FAIR	AWESOME pt+ft w/ co	$74.9_{\pm1.9}$	$73.8_{\pm1.0}$	$74.4_{\pm1.4}$	$57.5_{\pm3.3}$
FAIR ft	FastAlign	$74.7_{\pm2.2}$	$72.1_{\pm0.6}$	$73.3_{\pm1.3}$	$54.2_{\pm4.6}$
FAIR ft	AWESOME w/o co	$76.8_{\pm1.6}$	$72.8_{\pm1.0}$	$74.7_{\pm1.0}$	$53.0_{\pm2.7}$
FAIR ft	AWESOME w/ co	$75.9_{\pm1.7}$	$\mathbf{73.8}_{\pm0.6}$	$74.8_{\pm1.0}$	$57.5_{\pm4.5}$
FAIR ft	AWESOME ft w/o co	$77.0_{\pm1.5}$	$72.1_{\pm0.6}$	$74.5_{\pm0.8}$	$51.0_{\pm0.9}$
FAIR ft	AWESOME ft w/ co	$76.2_{\pm1.2}$	$72.1_{\pm1.3}$	$74.1_{\pm1.1}$	$50.9_{\pm1.0}$
FAIR ft	AWESOME pt+ft w/o co	$75.5_{\pm1.2}$	$71.9_{\pm1.1}$	$73.7_{\pm0.9}$	$52.1_{\pm3.5}$
FAIR ft	AWESOME pt+ft w/ co	$75.5_{\pm0.9}$	$72.6_{\pm1.1}$	$74.0_{\pm1.0}$	$52.6_{\pm3.8}$
Cross-lingual Transfer		$71.1_{\pm1.1}$	$73.3_{\pm1.0}$	$72.2_{\pm0.7}$	$55.1_{\pm5.7}$
Cross-lingual Transfer with realignment		$\mathbf{78.2}_{\pm1.8}$	$73.5_{\pm1.9}$	$\mathbf{75.8}_{\pm1.3}$	$\mathbf{58.1}_{\pm4.9}$
		XLM-R Large			
FAIR	FastAlign	$77.7_{\pm3.6}$	$75.1_{\pm2.3}$	$76.4_{\pm2.7}$	$65.8_{\pm2.8}$
FAIR	AWESOME w/o co	$80.1_{\pm1.1}$	$77.5_{\pm0.6}$	$78.7_{\pm0.5}$	$65.0_{\pm3.1}$
FAIR	AWESOME w/ co	$79.9_{\pm1.3}$	$77.0_{\pm2.4}$	$78.4_{\pm1.7}$	$64.4_{\pm2.2}$
FAIR	AWESOME ft w/o co	$80.7_{\pm1.6}$	$77.3_{\pm0.9}$	$79.0_{\pm0.6}$	$65.8_{\pm0.6}$
FAIR	AWESOME ft w/ co	$80.0_{\pm1.7}$	$\mathbf{78.5}_{\pm1.1}$	$79.2_{\pm1.1}$	$66.3_{\pm1.1}$
FAIR	AWESOME pt+ft w/o co	$79.3_{\pm0.8}$	$77.3_{\pm1.1}$	$78.3_{\pm0.7}$	$64.5_{\pm2.0}$
FAIR	AWESOME pt+ft w/ co	$79.1_{\pm1.8}$	$75.6_{\pm1.2}$	$77.3_{\pm0.8}$	$63.7_{\pm1.6}$
FAIR ft	FastAlign	$78.5_{\pm2.5}$	$75.1_{\pm1.6}$	$76.7_{\pm1.7}$	$61.7_{\pm1.9}$
FAIR ft	AWESOME w/o co	$\mathbf{83.2}_{\pm2.8}$	$76.0_{\pm0.9}$	$\mathbf{79.4}_{\pm1.6}$	$64.3_{\pm2.2}$
FAIR ft	AWESOME w/ co	$83.0_{\pm1.0}$	$76.1_{\pm1.7}$	$\mathbf{79.4}_{\pm1.3}$	$64.7_{\pm1.3}$
FAIR ft	AWESOME ft w/o co	$80.0_{\pm1.5}$	$76.0_{\pm1.5}$	$77.9_{\pm1.2}$	$64.9_{\pm0.9}$
FAIR ft	AWESOME ft w/ co	$81.7_{\pm2.0}$	$75.1_{\pm1.7}$	$78.3_{\pm1.8}$	$64.8_{\pm1.6}$
FAIR ft	AWESOME pt+ft w/o co	$80.3_{\pm1.9}$	$74.6_{\pm1.6}$	$77.4_{\pm1.3}$	$62.9_{\pm3.2}$
FAIR ft	AWESOME pt+ft w/ co	$80.2_{\pm0.9}$	$75.3_{\pm2.0}$	$77.6_{\pm1.1}$	$63.0_{\pm2.9}$
Cross-lingual Transfer		$81.2_{\pm1.2}$	$76.0_{\pm0.9}$	$78.5_{\pm0.4}$	$64.9_{\pm2.5}$
Cross-lingual Transfer with realignment		$80.7_{\pm2.1}$	$76.0_{\pm1.4}$	$78.3_{\pm1.6}$	$\mathbf{66.8}_{\pm1.8}$

Table 22: Cross-lingual Transfer and `translate-train` results in German for multilingual base models.

translation	aligner	precision	recall	mirco-f1	macro-f1
CamemBERT Base					
Opus	FastAlign	**74.9**$_{\pm1.1}$	**78.5**$_{\pm0.9}$	**76.7**$_{\pm0.9}$	**78.7**$_{\pm1.0}$
Opus	AWESOME w/o co	73.2$_{\pm1.2}$	77.7$_{\pm1.3}$	75.4$_{\pm1.2}$	77.9$_{\pm1.2}$
Opus	AWESOME w/ co	74.4$_{\pm0.8}$	77.5$_{\pm0.8}$	75.9$_{\pm0.8}$	78.1$_{\pm1.0}$
Opus	AWESOME ft w/o co	71.9$_{\pm1.1}$	76.5$_{\pm1.1}$	74.1$_{\pm1.1}$	76.7$_{\pm1.0}$
Opus	AWESOME ft w/ co	72.3$_{\pm2.0}$	77.4$_{\pm1.3}$	74.8$_{\pm1.7}$	77.3$_{\pm1.3}$
Opus	AWESOME pt+ft w/o co	74.0$_{\pm1.4}$	77.8$_{\pm1.5}$	75.9$_{\pm1.4}$	78.2$_{\pm1.3}$
Opus	AWESOME pt+ft w/ co	73.3$_{\pm1.0}$	77.9$_{\pm1.0}$	75.5$_{\pm0.9}$	77.8$_{\pm1.0}$
Opus ft	FastAlign	74.2$_{\pm2.1}$	78.4$_{\pm1.3}$	76.2$_{\pm1.7}$	78.3$_{\pm1.5}$
Opus ft	AWESOME w/o co	71.0$_{\pm1.6}$	76.2$_{\pm1.5}$	73.5$_{\pm1.5}$	76.1$_{\pm1.3}$
Opus ft	AWESOME w/ co	72.0$_{\pm1.8}$	77.4$_{\pm1.7}$	74.6$_{\pm1.7}$	77.2$_{\pm1.6}$
Opus ft	AWESOME ft w/o co	70.7$_{\pm1.8}$	75.0$_{\pm1.4}$	72.8$_{\pm1.6}$	75.4$_{\pm1.5}$
Opus ft	AWESOME ft w/ co	72.3$_{\pm1.6}$	76.6$_{\pm1.6}$	74.4$_{\pm1.6}$	76.8$_{\pm1.5}$
Opus ft	AWESOME pt+ft w/o co	72.2$_{\pm2.4}$	77.1$_{\pm1.5}$	74.6$_{\pm1.9}$	76.9$_{\pm1.7}$
Opus ft	AWESOME pt+ft w/ co	71.6$_{\pm1.2}$	77.0$_{\pm0.4}$	74.2$_{\pm0.8}$	76.7$_{\pm0.4}$
DrBERT 7GB					
Opus	FastAlign	70.9$_{\pm2.4}$	72.4$_{\pm2.1}$	71.7$_{\pm2.1}$	73.2$_{\pm2.1}$
Opus	AWESOME w/o co	69.5$_{\pm1.4}$	71.3$_{\pm1.5}$	70.4$_{\pm1.3}$	72.3$_{\pm1.6}$
Opus	AWESOME w/ co	69.5$_{\pm1.7}$	71.7$_{\pm0.7}$	70.6$_{\pm0.8}$	72.6$_{\pm0.7}$
Opus	AWESOME ft w/o co	69.3$_{\pm1.1}$	71.7$_{\pm0.7}$	70.4$_{\pm0.8}$	72.7$_{\pm0.7}$
Opus	AWESOME ft w/ co	68.1$_{\pm0.8}$	70.3$_{\pm1.5}$	69.2$_{\pm0.7}$	71.3$_{\pm0.8}$
Opus	AWESOME pt+ft w/o co	69.4$_{\pm1.4}$	71.7$_{\pm1.3}$	70.5$_{\pm1.2}$	72.6$_{\pm1.1}$
Opus	AWESOME pt+ft w/ co	70.0$_{\pm1.4}$	71.4$_{\pm1.0}$	70.7$_{\pm0.6}$	72.7$_{\pm0.5}$
Opus ft	FastAlign	**73.2**$_{\pm1.9}$	**73.7**$_{\pm1.5}$	**73.5**$_{\pm1.4}$	**74.9**$_{\pm1.4}$
Opus ft	AWESOME w/o co	69.6$_{\pm1.6}$	71.7$_{\pm1.1}$	70.7$_{\pm1.3}$	72.7$_{\pm1.1}$
Opus ft	AWESOME w/ co	70.5$_{\pm1.6}$	71.9$_{\pm1.2}$	71.2$_{\pm1.2}$	73.4$_{\pm1.1}$
Opus ft	AWESOME ft w/o co	69.1$_{\pm1.5}$	70.6$_{\pm1.1}$	69.8$_{\pm0.5}$	71.8$_{\pm0.3}$
Opus ft	AWESOME ft w/ co	70.8$_{\pm1.5}$	72.6$_{\pm1.3}$	71.6$_{\pm1.0}$	73.7$_{\pm0.8}$
Opus ft	AWESOME pt+ft w/o co	70.7$_{\pm1.0}$	71.7$_{\pm2.1}$	71.2$_{\pm1.4}$	73.2$_{\pm1.2}$
Opus ft	AWESOME pt+ft w/ co	70.1$_{\pm0.8}$	70.6$_{\pm1.7}$	70.4$_{\pm1.1}$	72.4$_{\pm1.2}$
DrBERT-PubMedBERT					
Opus	FastAlign	**76.2**$_{\pm1.4}$	79.4$_{\pm0.8}$	77.8$_{\pm1.1}$	79.7$_{\pm0.8}$
Opus	AWESOME w/o co	75.2$_{\pm1.2}$	77.5$_{\pm0.7}$	76.4$_{\pm0.9}$	78.4$_{\pm0.7}$
Opus	AWESOME w/ co	76.0$_{\pm1.0}$	79.2$_{\pm1.3}$	77.6$_{\pm1.1}$	79.6$_{\pm0.9}$
Opus	AWESOME ft w/o co	74.3$_{\pm1.4}$	78.9$_{\pm1.3}$	76.5$_{\pm1.2}$	78.7$_{\pm1.1}$
Opus	AWESOME ft w/ co	74.0$_{\pm1.2}$	77.8$_{\pm1.1}$	75.9$_{\pm1.1}$	78.2$_{\pm1.0}$
Opus	AWESOME pt+ft w/o co	73.9$_{\pm1.1}$	77.7$_{\pm0.6}$	75.8$_{\pm0.9}$	78.1$_{\pm0.7}$
Opus	AWESOME pt+ft w/ co	75.2$_{\pm0.4}$	79.1$_{\pm0.8}$	77.1$_{\pm0.5}$	79.0$_{\pm0.4}$
Opus ft	FastAlign	**76.2**$_{\pm1.8}$	**81.5**$_{\pm1.0}$	**78.8**$_{\pm1.4}$	**80.4**$_{\pm1.3}$
Opus ft	AWESOME w/o co	73.7$_{\pm1.4}$	78.7$_{\pm1.2}$	76.1$_{\pm1.3}$	78.4$_{\pm1.0}$
Opus ft	AWESOME w/ co	75.4$_{\pm1.2}$	81.2$_{\pm0.8}$	78.2$_{\pm0.8}$	80.3$_{\pm0.7}$
Opus ft	AWESOME ft w/o co	74.9$_{\pm0.9}$	80.7$_{\pm0.7}$	77.7$_{\pm0.8}$	79.7$_{\pm0.5}$
Opus ft	AWESOME ft w/ co	74.8$_{\pm1.3}$	79.7$_{\pm0.8}$	77.2$_{\pm1.1}$	79.2$_{\pm0.8}$
Opus ft	AWESOME pt+ft w/o co	75.6$_{\pm1.1}$	79.3$_{\pm1.5}$	77.4$_{\pm1.2}$	79.4$_{\pm1.1}$
Opus ft	AWESOME pt+ft w/ co	75.5$_{\pm1.2}$	80.3$_{\pm1.3}$	77.8$_{\pm1.2}$	79.8$_{\pm1.0}$

Table 23: `translate-train` results in French for domain and language-specific base models.

translation	aligner	precision	recall	micro-f1	macro-f1
GottBERT					
FAIR	FastAlign	$75.9_{\pm3.2}$	$70.3_{\pm2.2}$	$73.0_{\pm2.6}$	$54.8_{\pm4.8}$
FAIR	AWESOME w/o co	$79.5_{\pm1.6}$	$73.4_{\pm2.0}$	$76.3_{\pm1.7}$	$60.5_{\pm4.6}$
FAIR	AWESOME w/ co	$77.9_{\pm1.6}$	$72.9_{\pm1.4}$	$75.3_{\pm1.4}$	$57.2_{\pm5.2}$
FAIR	AWESOME ft w/o co	$78.4_{\pm2.8}$	$73.1_{\pm2.4}$	$75.7_{\pm2.4}$	$57.9_{\pm5.4}$
FAIR	AWESOME ft w/ co	$77.9_{\pm1.8}$	$72.6_{\pm2.3}$	$75.1_{\pm1.8}$	$53.4_{\pm2.7}$
FAIR	AWESOME pt+ft w/o co	$79.0_{\pm1.8}$	$\mathbf{74.1}_{\pm1.6}$	$76.5_{\pm1.6}$	$\mathbf{60.8}_{\pm3.5}$
FAIR	AWESOME pt+ft w/ co	$77.9_{\pm2.8}$	$73.6_{\pm2.5}$	$75.7_{\pm2.6}$	$57.6_{\pm4.4}$
FAIR ft	FastAlign	$76.6_{\pm3.2}$	$70.1_{\pm1.9}$	$73.2_{\pm2.4}$	$53.7_{\pm4.5}$
FAIR ft	AWESOME w/o co	$\mathbf{80.2}_{\pm1.1}$	$73.3_{\pm1.0}$	$\mathbf{76.6}_{\pm0.8}$	$58.7_{\pm6.1}$
FAIR ft	AWESOME w/ co	$79.2_{\pm0.8}$	$73.1_{\pm1.1}$	$76.0_{\pm0.7}$	$58.8_{\pm2.4}$
FAIR ft	AWESOME ft w/o co	$78.5_{\pm0.7}$	$72.4_{\pm1.0}$	$75.3_{\pm0.8}$	$56.1_{\pm6.6}$
FAIR ft	AWESOME ft w/ co	$78.8_{\pm2.3}$	$72.3_{\pm1.9}$	$75.4_{\pm1.8}$	$55.2_{\pm6.4}$
FAIR ft	AWESOME pt+ft w/o co	$78.6_{\pm1.6}$	$72.6_{\pm1.6}$	$75.5_{\pm1.4}$	$55.5_{\pm6.4}$
FAIR ft	AWESOME pt+ft w/ co	$79.4_{\pm1.8}$	$72.3_{\pm0.8}$	$75.6_{\pm1.1}$	$56.5_{\pm3.2}$
medBERT.de					
FAIR	FastAlign	$75.1_{\pm2.8}$	$69.2_{\pm1.6}$	$72.0_{\pm2.0}$	$56.7_{\pm5.3}$
FAIR	AWESOME w/o co	$75.4_{\pm1.6}$	$71.9_{\pm1.8}$	$73.6_{\pm1.2}$	$58.1_{\pm5.0}$
FAIR	AWESOME w/ co	$\mathbf{77.0}_{\pm3.5}$	$72.9_{\pm1.6}$	$74.9_{\pm2.4}$	$59.5_{\pm5.9}$
FAIR	AWESOME ft w/o co	$76.4_{\pm4.5}$	$70.9_{\pm1.6}$	$73.5_{\pm2.7}$	$56.2_{\pm5.4}$
FAIR	AWESOME ft w/ co	$75.8_{\pm4.0}$	$72.4_{\pm2.5}$	$74.1_{\pm3.1}$	$57.2_{\pm6.4}$
FAIR	AWESOME pt+ft w/o co	$74.9_{\pm3.9}$	$71.3_{\pm1.6}$	$73.0_{\pm2.4}$	$56.7_{\pm5.5}$
FAIR	AWESOME pt+ft w/ co	$76.1_{\pm4.1}$	$71.6_{\pm1.9}$	$73.7_{\pm2.6}$	$57.2_{\pm5.7}$
FAIR ft	FastAlign	$72.6_{\pm2.0}$	$68.9_{\pm0.8}$	$70.7_{\pm1.2}$	$60.7_{\pm1.2}$
FAIR ft	AWESOME w/o co	$75.2_{\pm2.0}$	$72.6_{\pm0.9}$	$73.9_{\pm1.0}$	$61.1_{\pm1.9}$
FAIR ft	AWESOME w/ co	$76.4_{\pm2.4}$	$\mathbf{73.6}_{\pm0.9}$	$\mathbf{75.0}_{\pm1.6}$	$62.2_{\pm3.3}$
FAIR ft	AWESOME ft w/o co	$75.1_{\pm3.6}$	$71.8_{\pm2.8}$	$73.4_{\pm3.1}$	$60.5_{\pm3.5}$
FAIR ft	AWESOME ft w/ co	$75.3_{\pm3.2}$	$71.9_{\pm2.2}$	$73.6_{\pm2.4}$	$58.6_{\pm6.1}$
FAIR ft	AWESOME pt+ft w/o co	$74.1_{\pm1.1}$	$71.4_{\pm0.8}$	$72.7_{\pm0.5}$	$57.8_{\pm4.9}$
FAIR ft	AWESOME pt+ft w/ co	$75.5_{\pm1.3}$	$72.3_{\pm1.0}$	$73.8_{\pm1.0}$	$\mathbf{62.6}_{\pm1.2}$

Table 24: `translate-train` results in German for domain and language-specific base models.

translation	aligner	precision	recall	micro-f1	macro-f1
		distilmBERT			
Opus	FastAlign	$65.9_{\pm 1.8}$	$67.0_{\pm 1.5}$	$66.4_{\pm 1.5}$	$68.5_{\pm 1.3}$
Opus	AWESOME w/o co	$\mathbf{70.5}_{\pm 1.7}$	$68.8_{\pm 1.4}$	$69.6_{\pm 1.4}$	$71.8_{\pm 1.3}$
Opus	AWESOME w/ co	$\mathbf{70.5}_{\pm 1.7}$	$\mathbf{69.0}_{\pm 1.4}$	$\mathbf{69.7}_{\pm 1.5}$	$\mathbf{71.9}_{\pm 1.3}$
Opus	AWESOME ft w/o co	$70.0_{\pm 1.8}$	$68.8_{\pm 1.6}$	$69.4_{\pm 1.6}$	$71.6_{\pm 1.4}$
Opus	AWESOME ft w/ co	$69.7_{\pm 1.8}$	$68.8_{\pm 1.6}$	$69.2_{\pm 1.6}$	$71.5_{\pm 1.4}$
Opus	AWESOME pt+ft w/o co	$70.1_{\pm 1.7}$	$68.9_{\pm 1.5}$	$69.5_{\pm 1.5}$	$71.7_{\pm 1.4}$
Opus	AWESOME pt+ft w/ co	$69.0_{\pm 1.8}$	$68.8_{\pm 1.6}$	$68.9_{\pm 1.6}$	$71.3_{\pm 1.4}$
Opus ft	FastAlign	$63.2_{\pm 1.7}$	$66.6_{\pm 1.6}$	$64.8_{\pm 1.5}$	$66.7_{\pm 1.2}$
Opus ft	AWESOME w/o co	$69.9_{\pm 1.6}$	$68.4_{\pm 1.3}$	$69.2_{\pm 1.4}$	$71.4_{\pm 1.2}$
Opus ft	AWESOME w/ co	$69.2_{\pm 1.7}$	$68.7_{\pm 1.4}$	$68.9_{\pm 1.5}$	$71.3_{\pm 1.3}$
Opus ft	AWESOME ft w/o co	$69.4_{\pm 1.8}$	$68.6_{\pm 1.6}$	$69.0_{\pm 1.6}$	$71.2_{\pm 1.4}$
Opus ft	AWESOME ft w/ co	$69.1_{\pm 1.7}$	$68.6_{\pm 1.5}$	$68.9_{\pm 1.5}$	$71.1_{\pm 1.3}$
Opus ft	AWESOME pt+ft w/o co	$69.1_{\pm 1.7}$	$68.6_{\pm 1.5}$	$68.8_{\pm 1.5}$	$71.1_{\pm 1.3}$
Opus ft	AWESOME pt+ft w/ co	$67.3_{\pm 1.7}$	$68.6_{\pm 1.5}$	$67.9_{\pm 1.5}$	$70.5_{\pm 1.3}$
		XLM-R Base			
Opus	FastAlign	$68.7_{\pm 1.7}$	$71.8_{\pm 1.2}$	$70.2_{\pm 1.4}$	$72.4_{\pm 1.2}$
Opus	AWESOME w/o co	$74.9_{\pm 1.5}$	$73.8_{\pm 1.2}$	$74.3_{\pm 1.3}$	$76.5_{\pm 1.1}$
Opus	AWESOME w/ co	$74.8_{\pm 1.6}$	$\mathbf{74.1}_{\pm 1.3}$	$\mathbf{74.4}_{\pm 1.3}$	$\mathbf{76.6}_{\pm 1.2}$
Opus	AWESOME ft w/o co	$74.6_{\pm 1.6}$	$73.8_{\pm 1.2}$	$74.2_{\pm 1.2}$	$76.3_{\pm 1.1}$
Opus	AWESOME ft w/ co	$74.1_{\pm 1.5}$	$73.8_{\pm 1.2}$	$74.0_{\pm 1.3}$	$76.1_{\pm 1.1}$
Opus	AWESOME pt+ft w/o co	$74.6_{\pm 1.6}$	$73.8_{\pm 1.2}$	$74.2_{\pm 1.2}$	$76.3_{\pm 1.1}$
Opus	AWESOME pt+ft w/ co	$73.4_{\pm 1.5}$	$73.8_{\pm 1.2}$	$73.6_{\pm 1.2}$	$75.9_{\pm 1.1}$
Opus ft	FastAlign	$67.7_{\pm 1.6}$	$71.7_{\pm 1.7}$	$69.6_{\pm 1.4}$	$71.4_{\pm 1.3}$
Opus ft	AWESOME w/o co	$\mathbf{75.4}_{\pm 1.7}$	$73.1_{\pm 1.7}$	$74.2_{\pm 1.6}$	$76.3_{\pm 1.5}$
Opus ft	AWESOME w/ co	$74.2_{\pm 1.7}$	$73.3_{\pm 1.7}$	$73.8_{\pm 1.6}$	$76.0_{\pm 1.6}$
Opus ft	AWESOME ft w/o co	$74.6_{\pm 1.8}$	$73.1_{\pm 1.7}$	$73.8_{\pm 1.6}$	$75.8_{\pm 1.5}$
Opus ft	AWESOME ft w/ co	$74.1_{\pm 1.8}$	$73.1_{\pm 1.7}$	$73.6_{\pm 1.6}$	$75.6_{\pm 1.5}$
Opus ft	AWESOME pt+ft w/o co	$74.3_{\pm 1.8}$	$73.2_{\pm 1.8}$	$73.7_{\pm 1.7}$	$75.7_{\pm 1.6}$
Opus ft	AWESOME pt+ft w/ co	$72.3_{\pm 1.7}$	$73.1_{\pm 1.7}$	$72.7_{\pm 1.6}$	$75.0_{\pm 1.5}$
		XLM-R Large			
Opus	FastAlign	$69.6_{\pm 1.3}$	$71.0_{\pm 0.9}$	$70.3_{\pm 1.1}$	$72.7_{\pm 0.9}$
Opus	AWESOME w/o co	$75.9_{\pm 1.0}$	$73.4_{\pm 0.8}$	$74.7_{\pm 0.9}$	$77.0_{\pm 0.8}$
Opus	AWESOME w/ co	$76.0_{\pm 1.0}$	$\mathbf{73.7}_{\pm 0.7}$	$74.8_{\pm 0.8}$	$77.2_{\pm 0.8}$
Opus	AWESOME ft w/o co	$75.5_{\pm 1.0}$	$73.5_{\pm 0.8}$	$74.5_{\pm 0.9}$	$76.8_{\pm 0.8}$
Opus	AWESOME ft w/ co	$75.2_{\pm 1.1}$	$73.5_{\pm 0.8}$	$74.3_{\pm 0.9}$	$76.7_{\pm 0.8}$
Opus	AWESOME pt+ft w/o co	$75.6_{\pm 1.1}$	$73.5_{\pm 0.8}$	$74.5_{\pm 0.9}$	$76.8_{\pm 0.8}$
Opus	AWESOME pt+ft w/ co	$74.5_{\pm 1.1}$	$73.5_{\pm 0.8}$	$74.0_{\pm 0.9}$	$76.5_{\pm 0.8}$
Opus ft	FastAlign	$70.8_{\pm 1.4}$	$72.2_{\pm 0.5}$	$71.5_{\pm 1.0}$	$73.2_{\pm 0.9}$
Opus ft	AWESOME w/o co	$\mathbf{77.3}_{\pm 1.3}$	$73.3_{\pm 0.6}$	$\mathbf{75.3}_{\pm 0.9}$	$\mathbf{77.6}_{\pm 0.7}$
Opus ft	AWESOME w/ co	$76.3_{\pm 1.1}$	$73.6_{\pm 0.6}$	$75.0_{\pm 0.9}$	$77.4_{\pm 0.6}$
Opus ft	AWESOME ft w/o co	$76.4_{\pm 1.3}$	$73.4_{\pm 0.6}$	$74.9_{\pm 0.9}$	$77.1_{\pm 0.7}$
Opus ft	AWESOME ft w/ co	$76.1_{\pm 1.2}$	$73.4_{\pm 0.6}$	$74.7_{\pm 0.9}$	$77.0_{\pm 0.7}$
Opus ft	AWESOME pt+ft w/o co	$76.1_{\pm 1.2}$	$73.5_{\pm 0.6}$	$74.8_{\pm 0.8}$	$77.0_{\pm 0.7}$
Opus ft	AWESOME pt+ft w/ co	$74.1_{\pm 1.1}$	$73.4_{\pm 0.6}$	$73.8_{\pm 0.8}$	$76.3_{\pm 0.6}$

Table 25: Full results for the `translate-test` approach in French with multilingual language models.

translation	aligner	precision	recall	micro-f1	macro-f1
		distilmBERT			
FAIR	FastAlign	$37.6_{\pm 2.2}$	$36.5_{\pm 1.1}$	$37.0_{\pm 1.5}$	$24.4_{\pm 1.2}$
FAIR	AWESOME w/o co	$66.8_{\pm 2.7}$	$63.7_{\pm 1.4}$	$65.2_{\pm 1.9}$	$49.1_{\pm 1.6}$
FAIR	AWESOME w/ co	$66.9_{\pm 2.7}$	$63.9_{\pm 1.6}$	$65.3_{\pm 1.9}$	$49.3_{\pm 1.7}$
FAIR	AWESOME ft w/o co	$68.6_{\pm 2.9}$	$63.7_{\pm 1.4}$	$66.0_{\pm 1.9}$	$49.6_{\pm 1.6}$
FAIR	AWESOME ft w/ co	$68.8_{\pm 2.8}$	$64.5_{\pm 1.4}$	$66.6_{\pm 1.9}$	$50.2_{\pm 1.6}$
FAIR	AWESOME pt+ft w/o co	$67.2_{\pm 2.8}$	$62.9_{\pm 1.4}$	$64.9_{\pm 1.9}$	$48.6_{\pm 1.6}$
FAIR	AWESOME pt+ft w/ co	$67.3_{\pm 2.9}$	$63.9_{\pm 1.6}$	$65.5_{\pm 2.0}$	$49.4_{\pm 1.7}$
FAIR ft	FastAlign	$29.1_{\pm 0.6}$	$31.1_{\pm 0.8}$	$30.1_{\pm 0.6}$	$18.3_{\pm 0.6}$
FAIR ft	AWESOME w/o co	$69.8_{\pm 2.6}$	$65.4_{\pm 1.4}$	$67.5_{\pm 1.8}$	$50.9_{\pm 3.8}$
FAIR ft	AWESOME w/ co	$68.9_{\pm 2.5}$	$66.4_{\pm 1.4}$	$67.6_{\pm 1.7}$	$51.4_{\pm 3.8}$
FAIR ft	AWESOME ft w/o co	$69.5_{\pm 2.3}$	$65.4_{\pm 1.4}$	$67.4_{\pm 1.7}$	$50.8_{\pm 3.8}$
FAIR ft	AWESOME ft w/ co	$69.7_{\pm 2.3}$	$65.5_{\pm 1.5}$	$67.5_{\pm 1.6}$	$51.0_{\pm 3.8}$
FAIR ft	AWESOME pt+ft w/o co	$\mathbf{69.9}_{\pm 2.4}$	$\mathbf{66.7}_{\pm 1.6}$	$\mathbf{68.3}_{\pm 1.8}$	$\mathbf{51.9}_{\pm 3.7}$
FAIR ft	AWESOME pt+ft w/ co	$69.2_{\pm 2.1}$	$66.2_{\pm 1.4}$	$67.6_{\pm 1.6}$	$51.3_{\pm 3.8}$
		XLM-R Base			
FAIR	FastAlign	$37.0_{\pm 1.5}$	$41.2_{\pm 0.8}$	$39.0_{\pm 0.8}$	$28.4_{\pm 0.5}$
FAIR	AWESOME w/o co	$71.1_{\pm 0.7}$	$72.3_{\pm 0.8}$	$71.7_{\pm 0.3}$	$56.4_{\pm 0.6}$
FAIR	AWESOME w/ co	$71.1_{\pm 0.7}$	$72.3_{\pm 0.8}$	$71.7_{\pm 0.3}$	$56.4_{\pm 0.6}$
FAIR	AWESOME ft w/o co	$72.0_{\pm 0.7}$	$71.4_{\pm 0.8}$	$71.7_{\pm 0.3}$	$56.1_{\pm 0.7}$
FAIR	AWESOME ft w/ co	$\mathbf{72.3}_{\pm 0.7}$	$72.3_{\pm 0.8}$	$72.3_{\pm 0.3}$	$\mathbf{56.7}_{\pm 0.7}$
FAIR	AWESOME pt+ft w/o co	$70.6_{\pm 0.7}$	$70.6_{\pm 0.8}$	$70.6_{\pm 0.3}$	$55.2_{\pm 0.6}$
FAIR	AWESOME pt+ft w/ co	$70.8_{\pm 0.7}$	$71.4_{\pm 0.8}$	$71.1_{\pm 0.3}$	$55.8_{\pm 0.6}$
FAIR ft	FastAlign	$29.7_{\pm 0.7}$	$35.1_{\pm 1.1}$	$32.2_{\pm 0.7}$	$22.0_{\pm 0.8}$
FAIR ft	AWESOME w/o co	$71.0_{\pm 1.2}$	$71.9_{\pm 2.0}$	$71.4_{\pm 1.2}$	$54.3_{\pm 5.2}$
FAIR ft	AWESOME w/ co	$70.4_{\pm 0.8}$	$72.9_{\pm 1.6}$	$71.6_{\pm 0.6}$	$54.8_{\pm 5.0}$
FAIR ft	AWESOME ft w/o co	$71.7_{\pm 1.2}$	$72.4_{\pm 1.6}$	$72.1_{\pm 0.8}$	$55.1_{\pm 5.3}$
FAIR ft	AWESOME ft w/ co	$72.1_{\pm 1.2}$	$72.4_{\pm 1.6}$	$72.3_{\pm 0.7}$	$55.2_{\pm 5.2}$
FAIR ft	AWESOME pt+ft w/o co	$72.0_{\pm 1.0}$	$\mathbf{73.4}_{\pm 1.7}$	$\mathbf{72.7}_{\pm 0.8}$	$55.5_{\pm 5.0}$
FAIR ft	AWESOME pt+ft w/ co	$71.7_{\pm 1.1}$	$72.8_{\pm 2.0}$	$72.2_{\pm 1.0}$	$55.3_{\pm 5.5}$
		XLM-R Large			
FAIR	FastAlign	$39.9_{\pm 1.7}$	$41.4_{\pm 0.7}$	$40.6_{\pm 1.0}$	$28.7_{\pm 1.2}$
FAIR	AWESOME w/o co	$76.4_{\pm 1.6}$	$72.5_{\pm 0.9}$	$74.4_{\pm 1.0}$	$49.2_{\pm 0.4}$
FAIR	AWESOME w/ co	$76.4_{\pm 1.6}$	$72.5_{\pm 0.9}$	$74.4_{\pm 1.0}$	$49.2_{\pm 0.4}$
FAIR	AWESOME ft w/o co	$77.5_{\pm 1.7}$	$71.6_{\pm 0.9}$	$74.5_{\pm 1.0}$	$49.0_{\pm 0.5}$
FAIR	AWESOME ft w/ co	$77.7_{\pm 1.7}$	$72.5_{\pm 0.9}$	$75.0_{\pm 1.0}$	$49.6_{\pm 0.5}$
FAIR	AWESOME pt+ft w/o co	$75.9_{\pm 1.7}$	$70.8_{\pm 0.9}$	$73.3_{\pm 1.0}$	$47.9_{\pm 0.4}$
FAIR	AWESOME pt+ft w/ co	$76.1_{\pm 1.6}$	$71.6_{\pm 0.9}$	$73.8_{\pm 1.0}$	$48.6_{\pm 0.4}$
FAIR ft	FastAlign	$30.6_{\pm 1.5}$	$34.0_{\pm 0.9}$	$32.2_{\pm 1.2}$	$20.9_{\pm 0.9}$
FAIR ft	AWESOME w/o co	$76.6_{\pm 3.5}$	$72.7_{\pm 2.1}$	$74.6_{\pm 2.7}$	$49.4_{\pm 1.8}$
FAIR ft	AWESOME w/ co	$76.5_{\pm 3.7}$	$\mathbf{74.4}_{\pm 2.1}$	$75.4_{\pm 2.8}$	$50.7_{\pm 1.9}$
FAIR ft	AWESOME ft w/o co	$77.8_{\pm 3.6}$	$73.3_{\pm 2.2}$	$75.5_{\pm 2.7}$	$50.1_{\pm 1.9}$
FAIR ft	AWESOME ft w/ co	$\mathbf{78.3}_{\pm 3.7}$	$73.3_{\pm 2.2}$	$75.7_{\pm 2.8}$	$50.2_{\pm 1.9}$
FAIR ft	AWESOME pt+ft w/o co	$77.9_{\pm 3.8}$	$\mathbf{74.4}_{\pm 2.1}$	$\mathbf{76.1}_{\pm 2.8}$	$\mathbf{51.0}_{\pm 1.9}$
FAIR ft	AWESOME pt+ft w/ co	$77.5_{\pm 3.8}$	$73.5_{\pm 2.1}$	$75.5_{\pm 2.8}$	$50.2_{\pm 1.9}$

Table 26: Full results for the `translate-test` approach in German with multilingual language models.

translation	aligner	precision	recall	micro-f1	macro-f1
Opus	FastAlign	$68.5_{\pm1.2}$	$69.1_{\pm1.2}$	$68.8_{\pm1.0}$	$71.2_{\pm1.0}$
Opus	AWESOME w/o co	$75.2_{\pm1.4}$	$\mathbf{71.8}_{\pm1.3}$	$\mathbf{73.5}_{\pm1.2}$	$\mathbf{75.8}_{\pm1.2}$
Opus	AWESOME w/ co	$74.8_{\pm1.5}$	$\mathbf{71.8}_{\pm1.4}$	$73.3_{\pm1.3}$	$75.6_{\pm1.2}$
Opus	AWESOME ft w/o co	$74.7_{\pm1.4}$	$\mathbf{71.8}_{\pm1.3}$	$73.2_{\pm1.2}$	$75.5_{\pm1.2}$
Opus	AWESOME ft w/ co	$74.1_{\pm1.4}$	$71.5_{\pm1.3}$	$72.8_{\pm1.2}$	$75.1_{\pm1.2}$
Opus	AWESOME pt+ft w/o co	$74.7_{\pm1.4}$	$\mathbf{71.8}_{\pm1.3}$	$73.2_{\pm1.2}$	$75.5_{\pm1.2}$
Opus	AWESOME pt+ft w/ co	$73.3_{\pm1.4}$	$71.5_{\pm1.3}$	$72.4_{\pm1.2}$	$74.9_{\pm1.2}$
Opus ft	FastAlign	$67.9_{\pm1.3}$	$69.3_{\pm1.5}$	$68.6_{\pm1.2}$	$70.6_{\pm1.2}$
Opus ft	AWESOME w/o co	$\mathbf{75.4}_{\pm1.3}$	$71.4_{\pm1.6}$	$73.3_{\pm1.3}$	$\mathbf{75.8}_{\pm1.3}$
Opus ft	AWESOME w/ co	$74.1_{\pm1.4}$	$71.4_{\pm1.7}$	$72.7_{\pm1.4}$	$75.4_{\pm1.4}$
Opus ft	AWESOME ft w/o co	$74.7_{\pm1.4}$	$71.4_{\pm1.5}$	$73.0_{\pm1.3}$	$75.4_{\pm1.3}$
Opus ft	AWESOME ft w/ co	$74.3_{\pm1.4}$	$71.5_{\pm1.6}$	$72.9_{\pm1.4}$	$75.3_{\pm1.4}$
Opus ft	AWESOME pt+ft w/o co	$74.3_{\pm1.3}$	$71.5_{\pm1.6}$	$72.9_{\pm1.3}$	$75.4_{\pm1.3}$
Opus ft	AWESOME pt+ft w/ co	$72.3_{\pm1.3}$	$71.5_{\pm1.6}$	$71.9_{\pm1.3}$	$74.7_{\pm1.3}$

Table 27: Results of `translate-test` in French with PubMedBERT.

translation	aligner	precision	recall	micro-f1	macro-f1
FAIR	FastAlign	$40.5_{\pm2.0}$	$41.8_{\pm1.0}$	$41.1_{\pm1.4}$	$29.2_{\pm1.2}$
FAIR	AWESOME w/o co	$73.6_{\pm2.7}$	$71.9_{\pm1.1}$	$72.7_{\pm1.6}$	$55.1_{\pm3.9}$
FAIR	AWESOME w/ co	$73.6_{\pm2.7}$	$71.9_{\pm1.1}$	$72.7_{\pm1.6}$	$55.1_{\pm3.9}$
FAIR	AWESOME ft w/o co	$74.7_{\pm2.7}$	$71.1_{\pm1.1}$	$72.8_{\pm1.6}$	$54.9_{\pm3.8}$
FAIR	AWESOME ft w/ co	$\mathbf{74.8}_{\pm2.9}$	$71.9_{\pm1.1}$	$\mathbf{73.3}_{\pm1.7}$	$\mathbf{55.4}_{\pm3.9}$
FAIR	AWESOME pt+ft w/o co	$73.2_{\pm2.8}$	$70.3_{\pm1.1}$	$71.7_{\pm1.6}$	$53.9_{\pm3.8}$
FAIR	AWESOME pt+ft w/ co	$73.4_{\pm2.8}$	$71.1_{\pm1.1}$	$72.2_{\pm1.6}$	$54.5_{\pm3.9}$
FAIR ft	FastAlign	$29.7_{\pm1.8}$	$34.3_{\pm1.7}$	$31.8_{\pm1.6}$	$21.7_{\pm2.6}$
FAIR ft	AWESOME w/o co	$72.6_{\pm0.7}$	$70.8_{\pm2.3}$	$71.6_{\pm1.5}$	$51.9_{\pm4.0}$
FAIR ft	AWESOME w/ co	$71.3_{\pm0.9}$	$71.9_{\pm2.0}$	$71.6_{\pm1.3}$	$52.4_{\pm3.6}$
FAIR ft	AWESOME ft w/o co	$73.4_{\pm1.0}$	$71.1_{\pm2.5}$	$72.2_{\pm1.7}$	$52.7_{\pm4.5}$
FAIR ft	AWESOME ft w/ co	$73.4_{\pm1.0}$	$71.1_{\pm2.5}$	$72.2_{\pm1.7}$	$52.7_{\pm4.5}$
FAIR ft	AWESOME pt+ft w/o co	$72.9_{\pm0.9}$	$\mathbf{72.3}_{\pm2.3}$	$72.6_{\pm1.5}$	$53.1_{\pm3.9}$
FAIR ft	AWESOME pt+ft w/ co	$72.9_{\pm0.8}$	$71.8_{\pm2.5}$	$72.3_{\pm1.6}$	$53.0_{\pm4.5}$

Table 28: Results of `translate-test` in German with PubMedBERT.

model	Drug			Strength			Frequency			Duration			Dosage			Form			global	
	p	r	f1	p	r	f1	p	r	f1	p	r	f1	p	r	f1	p	r	f1	macro	micro
distilmBERT																				
CLT	52.2	82.4	63.8	**79.1**	**92.2**	**85.1**	**63.1**	**58.2**	**60.5**	87.6	88.4	88.0	67.6	67.4	67.5	55.3	38.5	44.4	68.2	65.9
+ realigned	67.8	80.9	73.8	78.5	85.9	82.0	50.3	37.9	43.2	74.7	73.5	74.1	67.9	63.2	65.4	70.9	61.3	65.8	67.4	66.4
Opus - FastAlign	**69.9**	77.0	73.2	76.3	84.3	80.1	48.7	44.7	46.6	**94.2**	**90.7**	**92.4**	64.1	70.5	67.1	67.4	62.2	64.6	**70.7**	**68.3**
Opus - AWESOME pt	68.1	82.7	74.7	77.1	89.0	82.5	40.5	39.2	39.8	93.8	**90.7**	92.2	67.0	70.8	68.8	69.7	59.8	64.2	70.4	67.9
Opus - AWESOME ft	68.0	**83.3**	**74.8**	74.4	87.1	80.1	33.7	33.2	33.4	93.8	**90.7**	92.2	66.6	70.5	68.5	68.2	60.0	63.7	68.8	66.2
Opus - AWESOME pt+fr	65.9	82.4	73.2	76.4	89.4	82.3	39.4	38.9	39.1	92.9	**90.7**	91.8	**68.4**	**72.1**	**70.2**	70.1	60.4	64.8	70.2	67.8
Opus ft - FastAlign	68.4	80.6	74.0	74.3	85.1	79.3	45.0	44.7	44.8	91.2	**90.7**	90.9	59.7	68.2	63.6	71.0	**64.3**	**67.4**	70.0	67.7
Opus ft - AWESOME pt	66.3	81.8	73.2	77.0	91.8	83.7	37.1	36.3	36.7	92.9	**90.7**	91.8	62.4	68.4	65.3	69.8	59.1	64.0	69.1	66.5
Opus ft - AWESOME ft	65.5	83.0	73.2	78.5	90.6	84.0	34.2	32.9	33.5	93.3	**90.7**	92.0	61.0	69.5	65.0	70.4	59.1	64.1	68.6	65.9
Opus ft - AWESOME pt+fr	65.0	82.7	72.8	77.6	90.2	83.4	37.3	37.9	37.6	93.3	**90.7**	92.0	59.8	69.5	64.3	**71.1**	58.5	64.1	69.0	66.2
XLM-R Base																				
CLT	83.8	83.0	83.4	87.7	97.6	92.4	**73.0**	**75.3**	**74.0**	88.8	87.9	88.3	77.4	76.6	77.0	70.9	69.5	70.2	**80.9**	**79.1**
+ realigned	**83.9**	82.8	**83.3**	87.2	98.0	92.3	66.9	69.7	68.2	90.7	90.1	90.4	71.7	67.8	69.7	71.8	66.9	69.3	78.9	76.7
Opus - FastAlign	81.2	82.7	82.0	83.2	94.9	88.6	66.9	72.1	69.4	89.6	88.4	89.0	78.5	77.6	78.1	73.9	70.3	72.0	79.8	78.2
Opus - AWESOME pt	80.5	82.4	81.4	85.3	95.3	90.0	55.3	61.6	58.1	**92.0**	**90.2**	**91.1**	78.3	**78.7**	78.4	75.2	69.7	72.2	78.6	76.4
Opus - AWESOME ft	80.5	82.4	81.4	84.9	94.9	89.6	51.0	56.1	53.4	91.5	89.8	90.6	73.1	74.2	73.6	**75.3**	69.7	72.4	76.8	74.6
Opus - AWESOME pt+fr	81.1	81.8	81.4	85.1	95.7	90.1	52.1	57.9	54.8	91.4	89.3	90.4	**79.9**	**78.7**	**79.3**	73.2	71.2	72.2	78.0	75.8
Opus ft - FastAlign	83.7	82.7	83.2	84.2	96.1	89.7	70.7	73.4	72.0	90.5	87.9	89.2	75.0	75.0	75.0	72.7	**72.3**	**72.5**	80.3	78.6
Opus ft - AWESOME pt	81.2	**83.3**	82.2	87.3	**98.4**	92.5	47.4	53.4	50.2	90.2	89.3	89.7	74.0	74.7	74.3	73.0	69.9	71.4	76.7	74.2
Opus ft - AWESOME ft	81.0	82.4	81.7	**89.0**	98.0	**93.3**	46.3	52.9	49.4	90.2	89.8	90.0	74.8	76.6	75.7	74.8	69.7	72.1	77.0	74.5
Opus ft - AWESOME pt+fr	80.7	82.4	81.5	87.0	97.3	91.9	54.9	61.6	58.0	89.8	89.3	89.5	78.9	**78.7**	78.8	72.5	69.9	71.2	78.5	76.3
XLM-R Large																				
CLT	80.3	81.5	80.9	**90.5**	97.3	**93.8**	66.0	68.9	67.3	93.2	**89.3**	91.2	76.1	75.3	75.7	75.8	67.5	71.3	80.0	77.9
+ realigned	**85.5**	82.1	83.7	90.2	97.3	93.6	**66.6**	67.6	67.1	92.7	88.8	90.7	**80.2**	77.4	**78.7**	74.2	68.0	70.9	**80.8**	**78.8**
Opus - FastAlign	82.0	80.3	81.2	87.6	96.5	91.8	65.7	69.7	**67.6**	92.3	88.4	90.3	76.5	76.1	76.3	78.7	66.2	71.9	79.8	78.0
Opus - AWESOME pt	80.8	82.7	81.7	89.8	96.1	92.8	54.8	60.0	57.3	93.2	88.8	91.0	76.7	75.5	76.0	80.3	68.0	73.5	78.7	76.5
Opus - AWESOME ft	82.7	81.2	81.9	88.1	95.7	91.7	43.9	49.5	46.5	92.7	88.4	90.5	76.8	76.8	76.8	**80.4**	67.5	73.4	76.8	74.2
Opus - AWESOME pt+fr	83.8	81.2	82.4	88.8	96.5	92.5	49.0	56.6	52.5	93.2	88.8	91.0	78.3	**77.6**	78.0	79.2	65.4	71.6	78.0	75.4
Opus ft - FastAlign	84.6	80.6	82.5	79.0	88.2	83.3	61.5	**72.1**	66.3	91.8	88.4	90.1	77.3	74.7	76.0	75.4	68.6	71.8	78.3	76.6
Opus ft - AWESOME pt	83.2	**85.1**	**84.1**	89.9	96.9	93.2	51.5	58.2	54.6	94.2	88.8	91.4	74.8	76.3	75.5	79.4	70.3	**74.6**	78.9	76.5
Opus ft - AWESOME ft	84.4	81.8	83.0	89.9	**97.6**	93.6	45.4	51.1	48.1	93.2	**89.3**	91.2	74.3	76.6	75.4	75.6	**71.4**	73.4	77.4	74.8
Opus ft - AWESOME pt+fr	82.4	80.9	81.6	89.6	**97.6**	93.4	51.4	56.1	53.6	**94.6**	**89.3**	**91.9**	74.0	76.3	75.1	75.1	70.5	72.6	78.1	75.5

Table 29: Comparison class by class for `translate-train` and CLT on MedNERF with multilingual language models.

G Examples from the different datasets

EXAMPLES FROM N2C2
The patient's agitation was managed with **nightly**$_{\text{Frequency}}$ **haldol**$_{\text{Drug}}$ with **as needed**$_{\text{Frequency}}$ **haldol**$_{\text{Drug}}$ as well.
Improvement in clinical status was noted overnight and his **morphine**$_{\text{Drug}}$ drip was discontinued.
- hold all **antihypertensives**$_{\text{Drug}}$; plan to add back slowly at reduced doses and varying schedule - rule out MI - **bolus**$_{\text{Dosage}}$ **NS**$_{\text{Drug}}$ to maintain MAP > 60 with caution given ESRD and oliguric.
folic acid$_{\text{Drug}}$ **1 mg**$_{\text{Strength}}$ **Tablet**$_{\text{Form}}$ Sig: **One (1)**$_{\text{Dosage}}$ **Tablet**$_{\text{Form}}$ PO **DAILY (Daily)**$_{\text{Frequency}}$.
Iron$_{\text{Drug}}$ **50 mg**$_{\text{Strength}}$ **Tablet**$_{\text{Form}}$ Sustained Release Sig: **One (1)**$_{\text{Dosage}}$ **Tablet Sustained Release**$_{\text{Form}}$ PO **once a day**$_{\text{Frequency}}$.
EXAMPLES FROM GERNERMED TEST SET
Das **Eplerenon**$_{\text{Drug}}$ ist wegen Ihrer Herzinsuffizienz. Da können wir jetzt auf **50 mg**$_{\text{Strength}}$ p.o. 1-0-0 augmentieren. *Eplerenon is for your heart failure. We can now augment to 50mg p.o. 1-0-0.*
Wegen der COPD-Exazerbation wurde **Terbutalin**$_{\text{Drug}}$ **0,25 mg**$_{\text{Strength}}$ dem Patienten appliziert. Hierfür wurde der subkutane Weg gewählt. *Because of the COPD exacerbation, terbutaline 0.25 mg was administered to the patient. The subcutaneous route was chosen for this.*
Bei Vorhofflimmern ist neben **Betablockern**$_{\text{Drug}}$ auch die Gabe von **Magnesium**$_{\text{Drug}}$ p.o. sinnvoll. Hierfür würden wir mit **300 mg**$_{\text{Strength}}$ **einmal täglich**$_{\text{Frequency}}$ starten. Sofern möglich, ist eine Einnahme **mittags**$_{\text{Frequency}}$ (ca. 12 Uhr) zu bevorzugen. *In atrial fibrillation, in addition to beta-blocks, the administration of magnesium p.o. is also useful. For this we would start with 300 mg once a day. If possible, it is preferable to take it at noon (around 12 o'clock).*
Zur Optimierung der Herzinsuffizienztherapie wurde die Dosis von **Sacubitril / Valsartan**$_{\text{Drug}}$ auf **97 / 103 mg**$_{\text{Strength}}$ in **Tablettenform**$_{\text{Form}}$ mit Einnahme **am Morgen und am Abend**$_{\text{Frequency}}$ erweitert. *To optimize heart failure therapy, the dose of sacubitril / valsartan was extended to 97 / 103 mg in tablet form with intake in the morning and evening.*
Bei bekannter koronarer Herzerkrankung sollte **lebenslang**$_{\text{Duration}}$ **Acetylsalicylsäure**$_{\text{Drug}}$ **100 mg**$_{\text{Strength}}$ **morgens täglich**$_{\text{Frequency}}$ in oraler Applikation eingenommen werden. *In cases of known coronary artery disease, acetylsalicylic acid 100mg should be taken orally daily in the morning as a lifelong treatment.*
EXAMPLES FROM MEDNERF
TRAMADOL / PARACETAMOL$_{\text{Drug}}$ **37,5mg / 325mg**$_{\text{Strength}}$ *TRAMADO/PARACETAMOL 37,5mg/325mg*
AMLODIPINE$_{\text{Drug}}$ **5 mg**$_{\text{Strength}}$; **cpr**$_{\text{Form}}$ **1**$_{\text{Dosage}}$ **comprimé**$_{\text{Form}}$ **matin**$_{\text{Frequency}}$ **1**$_{\text{Dosage}}$ **comprimé**$_{\text{Form}}$ **soir**$_{\text{Frequency}}$ *AMLODIPINE 5mg; tab 1 tablet in the mording 1 tablet in the evening*
DOLIPRANETABS$_{\text{Drug}}$ **1000 MG**$_{\text{Strength}}$ **CPR PELL**$_{\text{Form}}$ PLQ / 8 (**Paracétamol**$_{\text{Drug}}$ **1.000 mg**$_{\text{Strength}}$ **comprimé**$_{\text{Form}}$) *DOLIPRANETABS 1000mg TAB PLQ / 8 (Paracetamol 1,000mg tablet)*
ACIDE ACETYLSALICYLIQUE$_{\text{Drug}}$ (**sel de lysine**$_{\text{Drug}}$) **75 mg**$_{\text{Strength}}$ **pdre p sol buv sach**$_{\text{Form}}$ (**KARDEGIC**$_{\text{Drug}}$) *ACETYLSALICYLIC ACID (lysine salts) 75mg oral powder for suspension (KARDEGIC)*
1$_{\text{Dosage}}$ **sachet**$_{\text{Form}}$ **matin midi et soir**$_{\text{Frequency}}$ si besoin *1 packet in the morning, at noon, and in the evening, if needed*

UMLS-KGI-BERT: Data-Centric Knowledge Integration in Transformers for Biomedical Entity Recognition

Aidan Mannion[1,2], **Thierry Chevalier**[3], **Didier Schwab**[1], **Lorraine Goeuriot**[1]

[1]Univ. Grenoble Alpes, CNRS, LIG, Grenoble, France

[2]EPOS SAS, Issy-les-Moulineaux, France

[3]UFR de Médecine Univ. Grenoble Alpes, La Tronche, France

Abstract

Pre-trained transformer language models (LMs) have in recent years become the dominant paradigm in applied NLP. These models have achieved state-of-the-art performance on tasks such as information extraction, question answering, sentiment analysis, document classification and many others. In the biomedical domain, significant progress has been made in adapting this paradigm to NLP tasks that require the integration of domain-specific knowledge as well as statistical modelling of language. In particular, research in this area has focused on the question of how best to construct LMs that take into account not only the patterns of token distribution in medical text, but also the wealth of structured information contained in terminology resources such as the UMLS. This work contributes a data-centric paradigm for enriching the language representations of biomedical transformer-encoder LMs by extracting text sequences from the UMLS. This allows for graph-based learning objectives to be combined with masked-language pre-training. Preliminary results from experiments in the extension of pre-trained LMs as well as training from scratch show that this framework improves downstream performance on multiple biomedical and clinical Named Entity Recognition (NER) tasks. All pre-trained models, data processing pipelines and evaluation scripts will be made publicly available.

1 Introduction

In recent times, transformer language models (Vaswani et al., 2017) have become the most popular and effective sequence modelling framework in almost all areas of applied Natural Language Processing. Unsupervised pre-training on large quantities of text allows transformers to capture rich semantic and syntactic patterns that can be transferred to many specialised language processing objectives. As such, transformer models that use the transfer learning paradigm whereby the model is trained in an unsupervised manner on a large text corpus and then fine-tuned on a downstream supervised-learning task have achieved state-of-the-art results across a wide range of general and domain-specific applications.

The proliferation of textual data in the biomedical domain (Electronic Health Records (EHRs), clinical documents, pharmaceutical specifications, etc) has precipitated the broad adoption of deep learning & NLP techniques for information extraction and processing (Li et al., 2021; Tiwari et al., 2020; Dubois et al., 2017). Moreover, it has been shown that language models are capable of encoding clinical knowledge to a certain extent (Singhal et al., 2022). Biomedical and clinical NLP, however, is widely recognised to present particular challenges that do not apply to the same extent in other domains, in particular the need to incorporate structured domain knowledge into text encodings (Chang et al., 2020). In order for neural language modelling to be reliable in a discipline as highly specialised as medicine, there is a more acute need for models to learn directly from domain-specific terminologies, as opposed to relying solely on corpus-based learning. Thus, a significant amount of research effort in the medical NLP community has been directed towards the question of how best to inject information from knowledge graphs (KGs) into LMs (He et al., 2022; Naseem et al., 2022; Li et al., 2020). However, a generalisable, widely-accepted approach to this technique that can be easily transferred across different problem settings, models and training corpora has yet to emerge. In addition, research into knowledge graph integration in NLP in the biomedical domain has tended to focus on English-language corpora; the utility and transferability of these techniques for other languages, for which less textual resources are available, as well as for multilingual models, remains therefore an under-explored area.

This paper aims to contribute to the resolution of

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 312–322

these issues by proposing a general framework for training BERT encoders (Devlin et al., 2019) using the UMLS (Unified Medical Language System, Bodenreider (2004)) alongside free-text corpora.

The main contributions of this work are as follows:

- We propose a data-centric method for formulating the KG-based learning objectives of triple classification and entity/link prediction in the language modelling paradigm, and implement a framework for training transformers using the UMLS knowledge base in parallel with masked-language pre-training.

- Pre-training on the UMLS alongside the European Clinical Case Corpus (Minard et al., 2021; Magnini et al., 2020), we show that this method brings improvements to pre-trained models across a range of biomedical entity recognition tasks in three different languages, as well as functioning as a competitive pre-training strategy that requires much less training data in comparison to state-of-the-art transformer models. We release the monolingual and multilingual model weights trained in this way, UMLS-KGI-BERT, as open-source resources for the clinical NLP research community.

- Based on this work, we release the Python library `bertify_umls`, built mainly on the `transformers` and `pandas` libraries, which allows researchers to create custom text datasets and effectively use the UMLS knowledge base as a training corpus for BERT-style LMs.

2 Related Work

2.1 Pre-trained LMs for Medical Applications

In general, the standard methodology for adapting neural text encoders to the biomedical domain has been to take a model that has been pre-trained on general-domain text corpora and continue this unsupervised pre-training on a medical corpus (Alrowili and Shanker, 2021; Lee et al., 2020; Alsentzer et al., 2019). However, recent work has suggested that, given enough training data, it is preferable to pre-train these models on large domain-specific corpora only, without starting from a general-domain checkpoint (Gu et al., 2021; Rasmy et al., 2021). In this work we explore both approaches, extending existing biomedical and general-domain models as well as training BERT models from scratch on our own generated datasets.

2.2 Knowledge-enhanced LMs

Techniques for the incorporation of knowledge graph structure into BERT models can, broadly speaking, be divided into three categories, each focusing on one of the three fundamental components of a machine learning system, i.e. 1) the training data, 2) the model architecture and 3) the objective function to be optimised. The first type of approach prioritises the augmentation of BERT's input data with information extracted from a knowledge graph. This extra information can be numerical, e.g. precomputed graph embeddings (Jeong et al., 2019) or textual, e.g. KG triples linked to input sentences (Liu et al., 2019).

The second type of approach focuses on adapting the architecture of BERT so that its language representations become fused with knowledge graph embeddings (KGEs) (Wang et al., 2021; Peters et al., 2019; Zhang et al., 2019). Knowledge graph fusion techniques such as these have been shown to be beneficial on certain English-language medical NLP tasks (Meng et al., 2021; Roy and Pan, 2021).

Thirdly, the self-supervised pre-training objective of BERT models can be augmented using the kind of knowledge graph reasoning tasks used to build KGE models. This approach is more commonly used for knowledge graph completion (Kim et al., 2020; Yao et al., 2019) but has also been shown to be an effective strategy in the biomedical NLP domain (Hao et al., 2020).

As previously mentioned, given that the medical domain is particularly exacting in terms of requirements for the use of structured facts, the exploration of ways in which ontological knowledge can be integrated into automated text processing is a very active area of research (Khosla et al., 2020; Mondal et al., 2019). In particular, there have been multiple successful efforts to integrate the UMLS knowledge graph into BERT models, notably UmlsBERT (Michalopoulos et al., 2021), which proposes a data-augmentation technique allowing for concept and semantic type information to be linked to input text, and SapBERT (Liu et al., 2021b,a), which introduced a self-alignment strategy for learning from UMLS synonym pairs via a multi-similarity (MS) loss function to force related concepts closer to one another in BERT's repre-

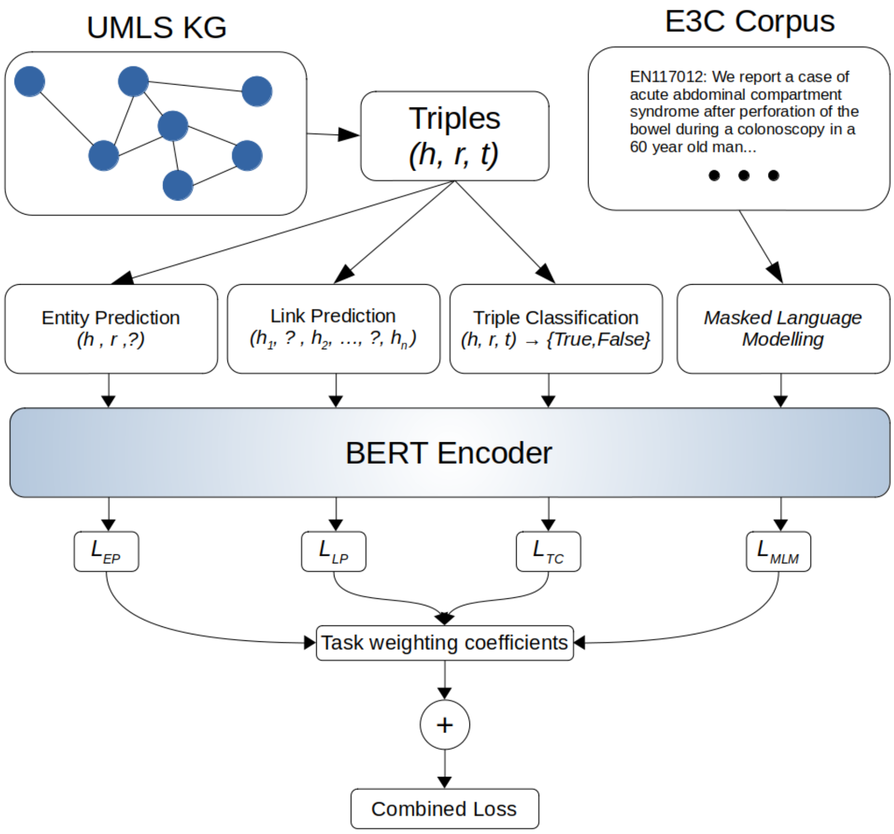


Figure 1: Overview of the UMLS-KGI pre-training process.

sentation space. Yuan et al. (2022) build on this strategy by applying MS loss to relation triples. In contrast, in this work we show that information from the UMLS can be incorporated into BERT models in a simpler way, using only cross-entropy classification loss, while also balancing this training process with standard masked-language BERT pre-training.

Recent general overviews of the landscape of AI research have highlighted the importance of data-centric approaches to building models (Zha et al., 2023; Hamid, 2022; Jakubik et al., 2022) and in light of these trends this work focuses on types 1) and 3) of knowledge base integration described above, i.e. on improving the performance of standard model architectures by constructing high-quality datasets that can be integrated into the self-supervised language modelling paradigm by modifying the BERT objective function. The motivation for this kind of approach is also to provide a pre-training framework that is more widely transferable and does not rely on any particular transformer-encoder architecture.

3 Methodology

In this work, we experiment with training BERT language models with three knowledge graph reasoning tasks derived from the UMLS, in addition to the standard masked-language modelling objective: entity prediction, link prediction and triple classification.

3.1 Dataset Construction

Formally, we consider the UMLS KG in the standard fashion, as a directed graph $G = (C, E, R)$ where C is the set of all medical concepts in the KG, E the set of all edges or relations that link these concepts to one another, and R the set of possible relation types, i.e. the labels r for each $e \in E$. The training sequences are thus generated from the KG dataset of ordered triples (h, r, t) where $(h, r) \in C \times C$ and $r \in R$. As a compendium of multiple different sources of taxonomic biomedical information, the UMLS metathesaurus contains multiple levels of granularity at which meaning representation can be analysed. We consider three such levels of granularity in our work:

- Terms - string descriptors for conceptual enti-

Table 1: Pre-training corpora sizes used in the experiments.

	Triple Classification	Entity Prediction	Paths	E3C corpus (num. documents)	Total Training Examples	Memory Footprint
French	200K	100K	64,208	25,740	389,948	604MB
Spanish	200K	100K	100K	1,876	401,876	162MB
English	200K	100K	100K	9,779	409,779	174MB
Total	600K	300K	264,208	37,395	1,201,603	940MB

ties

- CUIs (Concept Unique Identifiers) - the basic unit of meaning representation for the nodes in the knowledge graph, i.e. the elements of the set C.
- Semantic Groups - these are groupings of concepts that can be considered to define the type of entity a concept represents; e.g. anatomical structure, chemical, disorder etc.

Each concept (CUI) can be associated with multiple terms and multiple semantic groups. Thus, given that the entities h and t that make up the knowledge graph triples are represented as CUIs, in order to represent them as input text sequences for BERT models, we use the "preferred term" strings associated with the concepts h and t, except in the case of synonym relations where we randomly select another of the terms associated with the concept in question to associate with t. We also introduce a set of special tokens to represent the relation types R, of which there are seven (parent, child, synonymy, allowed qualifier, qualified by, broader, narrower). Concretely, the tokenization function for BERT models forms text classification sequences from triples in the following way;

$$\begin{aligned}\text{Tokenize}(h, r, t) = &\texttt{[CLS]}w_1^h \cdots w_m^h \\ &\texttt{[REL]}w_1^t \cdots w_n^t\texttt{[SEP]} \quad (1)\end{aligned}$$

where the w_i represent the token sequences corresponding to the strings h and t, `[CLS]` and `[SEP]` are BERT's standard classification and sequence-separation tokens as defined by Devlin et al. (2019), and `[REL]` is one of the relation tokens. For link prediction, we construct a dataset of variable-length paths through the KG by iteratively selecting a list of triples $(h_1, r_1, t_1), \ldots, (h_n, r_n, t_n)$ where $h_{i+1} = t_i$ to form a path $p = (h_1, r_1, h_2, \ldots, r_n, t_n)$.

Entity Prediction The entity classification task can be trivially integrated into the masked-language objective of BERT, by masking the tokens associated with the concept t.

Link Prediction We formulate link prediction as a narrow masked-language task by masking the relation tokens in the path dataset with another *hidden relation* token, for which the model is trained to fill in one of six relation types - as the triple classification and entity prediction tasks already have the partial goal of improving the model's capability to associate synonymous terms with each other, we exclude synonym relations from the path dataset.

Triple Classification Following the work of Hao et al. (2020), the triple classification objective is formulated as a binary classification problem where the model is tasked with classifying triples as true or false. In order to generate training examples of false triples, we use two different negative sampling strategies. Firstly, to provide directly contrastive examples for existing relations, we sample triples (h, r, t) where h and t belong to different semantic groups and construct corresponding false triples with the same relation type and semantic group categories, i.e. $(\hat{h}, r, \hat{t}) \notin G$ where $\hat{h}$ and $\hat{t}$ are of the same semantic group as h and t respectively. Secondly, to provide contrastive examples for relation types, we sample triples for which h and t are of the same semantic group, and form the negative training example by changing the relation type r. To ensure balance, the triple classification datasets used in this work are made up of 50% positive examples (real triples from the KG), 25% examples generated by the first negative sampling method and the rest by the second.

We perform stratified sampling on the base knowledge graph according to semantic groups, i.e. we ensure that the proportional representation of each semantic group in the knowledge-base triples for each language is maintained in the training datasets.

Mixed Objective Function In order to train BERT models using the UMLS-based reasoning tasks described above alongside the masked-language objective, each training example is augmented with an indicator label that tells the model which loss function to apply to the sequence in question. The overall loss function is then calculated as

$$\mathcal{L} = \mathcal{L}_{\text{MLM}} + \alpha_1 \mathcal{L}_{\text{EP}} + \alpha_2 \mathcal{L}_{\text{LP}} + \alpha_3 \mathcal{L}_{\text{TC}} \quad (2)$$

where the α_i are scalar task-weighting coefficients and $\mathcal{L}_{\text{MLM}}$, $\mathcal{L}_{\text{EP}}$, $\mathcal{L}_{\text{LP}}$, and $\mathcal{L}_{\text{TC}}$ correspond to the loss values for masked language modelling, entity prediction, link prediction and triple classification respectively. We use the standard cross-entropy classification loss for all tasks.

4 Experiments

For the evaluation of the approach described in the previous section, we restrict our attention in this paper to NER tasks. Where possible, we use the datasets and training-evaluation-test splits that are publicly available via the Huggingface datasets library[1].

4.1 KG-integrated pre-training

Pre-training corpora As a resource for masked-language pre-training, we utilise the European Clinical Case Corpus (E3C) version 2.0.0[2], a freely-available multilingual corpus of clinical narratives. We evaluate our method in three different languages; English, French and Spanish. These languages were chosen as they are the three most well-represented languages in the metathesaurus for which we have access to pre-trained clinical BERT models for comparison. The sizes of the combined UMLS-E3C datasets used are shown in Table 1.

For each language, we compare the performance of 1) a transformer model trained from scratch on each monolingual dataset (KGI-BERT$_{EN,FR,ES}$) against 2) a multilingual version of the same model trained on all three datasets (KGI-BERT$_m$), 3) a pre-trained monolingual biomedical model and 4) the same pre-trained model with supplementary training on the corresponding monolingual UMLS-E3C dataset.

The UMLS-KGI models were trained for 64 epochs on each dataset, using the PyTorch implementation of the weighted ADAM optimizer (Loshchilov and Hutter, 2019) with default parameters. We use a maximal sequence length of 256 for the masked-language modelling sequences, an effective batch size of 1500 and a triangular learning rate schedule peaking at 7.5×10^{-4}. To take into account the varying sizes of the components of the pre-training dataset we set the values of the coefficients of the loss function such that they are inversely proportional to the number of documents available:

$$\alpha_i = \frac{\sum_{j=0, j\neq i}^{3} n_j}{2\sum_{k=0}^{3} n_k}$$

where the n_k correspond to the number of documents in the training set for each UMLS-based task. In this way, the E3C masked-language loss has the same weighting as the UMLS-based task losses.

Pre-trained models For supplementary training, we make use of what are, to the best of our knowledge, the overall best-performing biomedical BERT models of their size (pre-trained using masked-language tasks only) for each language, according to baseline experiments on the NER tasks.

For French, we use DrBERT (Labrak et al., 2023), for Spanish the RoBERTa-based biomedical model released by Carrino et al. (2021), which we refer to as BioRoBERTa-ES, and for English PubMedBERT (Gu et al., 2021). For training from scratch, we use the DistilBERT model configuration (Sanh et al., 2019) with 12 encoder layers and 12 attention heads.

4.2 Evaluation corpora

We evaluate these models on nine different clinical entity recognition tasks; four in French, two in Spanish and three in English. In order to ensure a fair comparison between models and evaluate more directly the knowledge transfer capabilities of the pre-trained models, we restrict ourselves to a *one-shot* setting for all tasks, i.e. the model is given a single pass over the training data before being evaluated on the test set. For all fine-tuning runs, we use an effective batch size of 4 (we found that very frequent optimizer updates give better results in for few-shot learning), learning rate 2×10^{-5} and weight decay of 0.01.

CAS/ESSAIS CAS (Grabar et al., 2018) and ESSAIS (Dalloux et al., 2021) are corpora of clinical cases in French for which a subset is annotated with part-of-speech tags as well as semantic biomedical annotations (UMLS concepts, negation, and

[1] https://huggingface.co/datasets

[2] https://live.european-language-grid.eu/catalogue/corpus/7618

Table 2: Results on the French-language NER tasks. **Bold:** best result, underlined: next best.

	CAS-POS			CAS-SG			QUAERO-MEDLINE			ESSAI-POS		
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
DrBERT-4GB	90.94	91.59	90.84	65.86	64.89	62.20	68.65	69.38	66.66	94.83	95.08	94.69
+ UMLS-KGI	**93.15**	**93.22**	**92.84**	70.82	**69.98**	67.14	71.59	72.37	69.90	94.92	94.76	94.59
KGI-BERT$_{FR}$	88.55	88.40	87.82	**71.57**	66.90	65.79	71.78	**72.93**	70.75	**95.46**	**95.40**	**95.18**
KGI-BERT$_{m}$	90.87	90.58	90.16	71.14	69.81	**67.28**	**72.04**	72.89	**70.96**	94.88	94.84	94.55

Table 3: Results on the English-language NER tasks.

	NCBI-Disease			BioRED-NER			JNLPBA04		
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
PubMedBERT	93.81	94.26	93.53	**84.76**	85.33	83.35	81.57	82.59	81.13
+ UMLS-KGI	**94.65**	**95.11**	**94.46**	84.28	**85.92**	**83.64**	**85.75**	**86.04**	**85.15**
KGI-BERT$_{EN}$	89.33	89.43	88.99	82.98	85.89	82.99	81.82	82.90	81.47
KGI-BERT$_{m}$	89.40	90.04	89.16	82.67	84.63	81.97	81.24	82.47	82.02

uncertainty). We evaluate our models on the two corresponding medical POS-tagging tasks, CAS-POS and ESSAI-POS, as well as formulating a semantic-group token classification task using the CAS corpus annotations (CAS-SG).

QUAERO The QUAERO French Medical Corpus (Névéol et al., 2014) is a corpus of biomedical documents from EMEA and Medline annotated with UMLS concepts to facilitate entity recognition and document classification tasks. The NER evaluation task we make use of here, QUAERO-MEDLINE, involves semantic group identification in the Medline documents.

PharmaCoNER (Gonzalez-Agirre et al., 2019) Designed for the automated recognition of pharmacological substances, compounds and proteins in Spanish-language clinical documents, this is a manually annotated subset of the Spanish Clinical Case Corpus (SPACCC (Intxaurrondo, 2018)).

MEDDOCAN Similarly to PharmaCoNER, the MEDDOCAN corpus (Marimon et al., 2019) is an annotated subset of SPACCC, in this case with semantic entity types relevant to clinical document anonymisation, i.e. words and expressions constituting Personal Health Information (PHI).

NCBI-Disease (Doğan et al., 2014) The NCBI disease corpus is made up of PubMed abstracts with annotated disease mentions. In this work, we restrict our attention to token classification at the mention level.

BioRED (Luo et al., 2022) This corpus is designed for biomedical relation extraction and entity recognition; we focus on the latter in this work. This task can be considered a more semantically general version of the NCBI disease recognition task, in that the BioRED corpus consists of PubMed abstracts annotated with a diverse range of entity types including genes, proteins and chemicals.

JNLPBA04 NER Dataset (Collier and Kim, 2004) Developed in the context of a biomedical entity recognition shared task, this corpus consists of Medline documents annotated with mentions of DNA, RNA, proteins, cell types and cell lines.

We report the macro-averaged precision, recall and F1-score for each task. Results for the French, English and Spanish tasks can be seen in Tables 2, 3, and 4 respectively. We find that the best-performing models are in general the pre-trained checkpoints for which training has been extended via knowledge graph integration. This is unsurprising given that these are the models that have undergone the most domain-specific pre-training among all variants. It is important to highlight, moreover, the fact that the KGI-BERT variants are competitive with the pre-trained baselines for many tasks, despite being trained on less data. The largest improvements brought about by the UMLS-KGI training strategy can be seen in the French and Spanish tasks, suggesting that this technique will be more beneficial for lower-resource languages for which there is more room for improvement with respect to existing models.

The number of documents and target label classes for each evaluation task is show in Table 5.

Table 4: Results on the Spanish-language NER tasks. **Bold:** best result, underlined: next best.

	PharmaCoNER			MEDDOCAN		
Model	**P**	**R**	**F1**	**P**	**R**	**F1**
BioRoberta-ES	81.11	81.99	80.41	91.41	93.15	91.84
+ UMLS-KGI	83.52	84.30	83.90	**93.65**	**95.32**	**91.99**
KGI-BERT$_{ES}$	79.95	80.14	78.11	92.28	92.93	92.17
KGI-BERT$_{m}$	**85.05**	**85.95**	**85.49**	92.32	92.65	91.98

Table 5: Number of documents and target classes in the NER evaluation datasets

Dataset	Train	Dev	Test	N. Classes
CAS-POS	2,652	569	569	31
CAS-SG	167	54	54	15
QUAERO-MEDLINE	788	790	787	11
ESSAI-POS	5,072	1,088	1,087	34
NCBI-Disease	5,433	924	941	3
BioRED-NER	387	98	97	7
JNLPBA04	16,619	1,927	3,856	11
PharmaCoNER	500	250	250	5
MEDDOCAN	500	250	250	22

4.3 Ablation Experiments

In order to measure the relative effect of the three KG-derived pre-training tasks on downstream performance, we perform ablation experiments with the continually pre-trained models. This involved comparing the downstream performance on the NER tasks of different versions of the UMLS-extended models, each with one of the three KG-based pre-training tasks excluded from the pre-training process. For ablation, we use identical experimental settings to those described previously, except with 32 pre-training epochs rather than 64.

In general, the ablation results, for which the macro F1 scores are shown in Table 6, suggest that the majority of the benefits in terms of NER performance are brought about by the link prediction task, although there are not enough statistically significant differences among the results to fully justify this conclusion.

It is clear also that certain tasks tend to add unhelpful noise to the model with respect to some tasks, in particular the ESSAI-POS task in French and the MEDDOCAN task in Spanish. This may be due to the nature of these entity recognition tasks being more linked to general semantic patterns (i.e. parts-of-speech and identifying information) such that the addition of biomedical knowledge to the models does not improve their representation of the relevant concepts.

5 Conclusions and Future Work

This paper introduces UMLS-KGI, a framework for training BERT models using knowledge graphs requiring highly minimal adjustments to the standard language modelling paradigm. We show the potential of this method to increase the performance of BERT models on various NER tasks. The results presented in this paper suggest that for clinical NER tasks, high-quality small-scale datasets derived from structured information, alongside alongside relatively small clinical text corpora, can be as effective as large-scale corpora for pre-training BERT models. We make our models and data-processing pipelines freely available online.

Future work in this direction will involve the incorporation of more diverse graph-based reasoning tasks in the pre-training strategy with more fine-grained representation of relation types, as well as intrinsic evaluation of the UMLS-KGI-BERT language representations via embedding visualisation and interpretability studies.

Limitations

The work presented in this paper is subject to a number of limitations which will be addressed in future work. Firstly, we evaluate UMLS-KGI-BERT on a very narrow range of tasks limited to token classification - a broader range of information extraction and reasoning tasks would be necessary for a more complete picture of the utility of our pre-training methods. In addition, we only train models for mid-to-high-resource languages; to properly validate the applicability of this approach, in particular the lessening of the need to rely on large training corpora, it will be necessary to train and evaluate such models in more low-resource settings.

Table 6: Macro-F1 scores for the ablation experiments.

		Dataset			
Base Model	**KG Tasks**	**CAS-POS**	**CAS-SG**	**QUAERO-MEDLINE**	**ESSAI-POS**
DrBERT-4GB	-	90.84	62.20	66.66	**94.69**
	EP+LP	91.59	64.85	66.08	94.62
	EP+TC	90.86	62.11	66.75	94.88
	TC+LP	92.01	65.98	66.89	94.41
	all	**92.04**	**66.22**	**67.15**	94.50
		NCBI-Disease	**BioRED-NER**	**JNLPBA04**	
PubMedBERT	-	93.53	83.35	81.13	
	EP+LP	93.24	82.40	81.25	
	EP+TC	93.37	83.09	82.66	
	TC+LP	**94.13**	83.38	84.30	
	all	94.11	**83.45**	**84.36**	
		PharmaCoNER	**MEDDOCAN**		
BioRoberta-ES	-	81.11	91.84		
	EP+LP	81.12	**91.86**		
	EP+TC	82.40	91.80		
	TC+LP	83.22	91.71		
	all	**83.46**	91.77		

References

Sultan Alrowili and Vijay Shanker. 2021. BioM-Transformers: Building Large Biomedical Language Models with BERT, ALBERT and ELECTRA. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 221–227, Online. Association for Computational Linguistics.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

O. Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. PubMed PMID: 14681409; PubMed Central PMCID: PMC308795.

Casimiro Pio Carrino, Jordi Armengol-Estapé, Asier Gutiérrez-Fandiño, Joan Llop-Palao, Marc Pàmies, Aitor Gonzalez-Agirre, and Marta Villegas. 2021. Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario. *CoRR*, abs/2109.03570. ArXiv preprint 2109.03570.

David Chang, Ivana Balažević, Carl Allen, Daniel Chawla, Cynthia Brandt, and Andrew Taylor. 2020. Benchmark and Best Practices for Biomedical Knowledge Graph Embeddings. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 167–176, Online. Association for Computational Linguistics.

Nigel Collier and Jin-Dong Kim. 2004. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 73–78, Geneva, Switzerland. COLING.

Clément Dalloux, Vincent Claveau, Natalia Grabar, Lucas Emanuel Silva Oliveira, Claudia Maria Cabral Moro, Yohan Bonescki Gumiel, and Deborah Ribeiro Carvalho. 2021. Supervised learning for the detection of negation and of its scope in French and Brazilian Portuguese biomedical corpora. *Natural Language Engineering*, 27(2):181–201.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Sébastien Dubois, Nathanael Romano, Kenneth Jung, Nigam Haresh Shah, and David C. Kale. 2017. The effectiveness of transfer learning in electronic health records data. In *International Conference on Learning Representations*.

Aitor Gonzalez-Agirre, Montserrat Marimon, Ander Intxaurrondo, Obdulia Rabal, Marta Villegas, and Martin Krallinger. 2019. Pharmaconer: Pharmacological substances, compounds and proteins named entity recognition track. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 1–10, Hong Kong, China. Association for Computational Linguistics.

Natalia Grabar, Vincent Claveau, and Clément Dalloux. 2018. CAS: French Corpus with Clinical Cases. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 122–128, Brussels, Belgium. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthcare*, 3(1). Place: New York, NY, USA Publisher: Association for Computing Machinery.

Oussama H. Hamid. 2022. From model-centric to data-centric AI: A paradigm shift or rather a complementary approach? In *2022 8th International Conference on Information Technology Trends (ITT)*, pages 196–199.

Boran Hao, Henghui Zhu, and Ioannis Paschalidis. 2020. Enhancing Clinical BERT Embedding using a Biomedical Knowledge Base. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 657–661, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yong He, Cheng Wang, Shun Zhang, Nan Li, Zhaorong Li, and Zhenyu Zeng. 2022. KG-MTT-BERT: Knowledge graph enhanced BERT for multi-type medical text classification. ArXiv preprint 2210.03970.

Ander Intxaurrondo. 2018. SPACCC. Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).

Johannes Jakubik, Michael Vössing, Niklas Kühl, Jannis Walk, and Gerhard Satzger. 2022. Data-centric artificial intelligence. ArXiv preprint 2212.11854.

Chanwoo Jeong, Sion Jang, Hyuna Shin, Eunjeong Lucy Park, and Sungchul Choi. 2019. A context-aware citation recommendation model with bert and graph convolutional networks. *Scientometrics*, pages 1–16.

Sopan Khosla, Shikhar Vashishth, Jill Fain Lehman, and Carolyn Rose. 2020. MedFilter: Improving Extraction of Task-relevant Utterances through Integration of Discourse Structure and Ontological Knowledge. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7781–7797, Online. Association for Computational Linguistics.

Bosung Kim, Taesuk Hong, Youngjoong Ko, and Jungyun Seo. 2020. Multi-Task Learning for Knowledge Graph Completion with Pre-trained Language Models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1737–1743, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023. DrBERT: A robust pre-trained model in French for biomedical and clinical domains. ArXiv preprint 2304.00958.

J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. Ho So, and J. Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics 2020*.

Irene Li, Jessica Pan, Jeremy Goldwasser, Neha Verma, Wai Pan Wong, Muhammed Yavuz Nuzumlalı, Benjamin Rosand, Yixin Li, Matthew Zhang, David Chang, R. Andrew Taylor, Harlan M. Krumholz, and Dragomir Radev. 2021. Neural natural language processing for unstructured data in electronic health records: a review. ArXiv preprint 2107.02975.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. 2020. BEHRT: Transformer for Electronic Health Records. *Scientific Reports*, 10(1):7155. Number: 1 Publisher: Nature Publishing Group.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021a. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021b. Learning domain-specialised representations for cross-lingual biomedical entity linking. In *Proceedings of ACL-IJCNLP 2021*, pages 565–574.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2019. K-BERT: Enabling Language Representation with Knowledge Graph. In *AAAI Conference on Artificial Intelligence*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. ArXiv:1711.05101.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N. Arighi, and Zhiyong Lu. 2022. Biored: A comprehensive biomedical relation extraction dataset. *CoRR*, abs/2204.04263.

B. Magnini, B. Altuna, A. Lavelli, M. Speranza, and R. Zanoli. 2020. The e3c project: Collection and annotation of a multilingual corpus of clinical cases.

Montserrat Marimon, Aitor Gonzalez-Agirre, Ander Intxaurrondo, Heidy Rodriguez, Jose Lopez Martin, Marta Villegas, and Martin Krallinger. 2019. Automatic de-identification of medical texts in spanish: the meddocan track, corpus, guidelines, methods and evaluation of results. In *IberLEF@ SEPLN*, pages 618–638.

Zaiqiao Meng, Fangyu Liu, Thomas Clark, Ehsan Shareghi, and Nigel Collier. 2021. Mixture-of-Partitions: Infusing Large Biomedical Knowledge Graphs into BERT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4672–4681, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. 2021. Umls-BERT: Clinical domain knowledge augmentation of contextual embeddings using the Unified Medical Language System Metathesaurus. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1744–1753, Online. Association for Computational Linguistics.

Anne-Lyse Minard, Roberto Zanoli, Begoña Altuna, Manuela Speranza, Bernardo Magnini, and Alberto Lavelli. 2021. European clinical case corpus. Bruno Kessler Foundation.

Ishani Mondal, Sukannya Purkayastha, Sudeshna Sarkar, Pawan Goyal, Jitesh Pillai, Amitava Bhattacharyya, and Mahanandeeshwar Gattu. 2019. Medical entity linking using triplet network. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 95–100, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Usman Naseem, Ajay Bandi, Shaina Raza, Junaid Rashid, and Bharathi Raja Chakravarthi. 2022. Incorporating Medical Knowledge to Transformer-based Language Models for Medical Dialogue Generation. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 110–115, Dublin, Ireland. Association for Computational Linguistics.

Aurélie Névéol, Cyril Grouin, Jeremy Leixa, Sophie Rosset, and Pierre Zweigenbaum. 2014. The QUAERO French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization. In *Proc of BioTextMining Work*, pages 24–30.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 43–54, Hong Kong, China. Association for Computational Linguistics.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine*, 4(1):86.

Arpita Roy and Shimei Pan. 2021. Incorporating medical knowledge in BERT for clinical relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5357–5366, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large Language Models Encode Clinical Knowledge. ArXiv:2212.13138 [cs].

Premanand Tiwari, Kathryn L. Colborn, Derek E. Smith, Fuyong Xing, Debashis Ghosh, and Michael A. Rosenberg. 2020. Assessment of a Machine Learning Model Applied to Harmonized Electronic Health Record Data for the Prediction of Incident Atrial Fibrillation. *JAMA Network Open*, 3(1). Publisher: American Medical Association.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation. *Transactions of the Association for Computational Linguistics*, 9:176–194. Place: Cambridge, MA Publisher: MIT Press.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. KG-BERT: BERT for knowledge graph completion. ArXiv preprint 1909.03193.

Zheng Yuan, Zhengyun Zhao, Haixia Sun, Jiao Li, Fei Wang, and Sheng Yu. 2022. CODER: Knowledge-infused cross-lingual medical term embedding for term normalization. *Journal of biomedical informatics*, 126:103983. Place: United States.

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. 2023. Data-centric artificial intelligence: A survey.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced Language Representation with Informative

Table 7: Size of the UMLS dataset from which the KG-based pre-training corpus was sampled.

Language	Terms	CUIs	Relations
English	3,912,195	2,245,468	17,121,829
Spanish	303,978	118,061	437,578
French	202,963	171,060	669,006
Total	4,419,136	2,534,589	18,228,413

Entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

A Dataset Statistics

A.1 UMLS Knowledge Graph

We use the 2022AB release of the UMLS knowledge graph, which contains 8,751,471 concepts defined by 3,711,072 unique identifiers (CUIs), and 25,369,590 relations. Restricting our attention to semantic types related to human biology and medicine, we end up with the base dataset outlined in Table 7.

B Supplementary Experimental Details

Pre-trained Checkpoints We use the following pre-trained model weights downloaded from the HuggingFace model hub as baseline models;

- DrBERT: `Dr-BERT/DrBERT-4GB`
- PubMedBERT: `microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext`
- BioRoBERTa-ES: `PlanTL-GOB-ES/roberta-base-biomedical-clinical-es`

Model Hyperparameters The hyperparameter settings used for the pre-training on the UMLS-based dataset are shown in Table 8. The pre-training process used a linear learning rate schedule with warmup, where the learning rate increases from zero over the warmup period until it reaches the specified before decaying linearly over the rest of the training steps. In the interest of minimising the energy consumption of our experiments, we carried out very minimal hyperparameter search, leaving most parameters at their default values. The experiments were run using Python 3.8.15, with PyTorch version 2.0.0 and CUDA 11.8, along with the `transformers` library version 4.27.4.

Table 8: Hyperparameter settings for pre-training the UMLS-KGI models.

Parameter	Value
Sequence Length	256
Learning rate	0.00075
Learning rate warmup steps	10,770
Batch size	15
Gradient accumulation steps	100
MLM probability	0.15

Hardware specifications The pre-training experiments were run on four Nvidia Tesla V100 GPUs with 32GB of RAM, while the fine-tuning experiments were run on an RTX 2080 Ti with 11GB of RAM.

WangLab at MEDIQA-Chat 2023: Clinical Note Generation from Doctor-Patient Conversations using Large Language Models

John Giorgi[1,2,3*] **Augustin Toma**[1,3,4*] **Ronald Xie**[1,2,3,4*]
Sondra S. Chen[1,5] **Kevin R. An**[1,6] **Grace X. Zheng**[1,5] **Bo Wang**[1,3,4*]

[1]University of Toronto [2]Terrence Donnelly Centre for Cellular and Biomolecular Research
[3]Vector Institute for AI [4]University Health Network [5]Sunnybrook Health Sciences Centre
[6]Department of Cardiac Surgery, University of Toronto

{john.giorgi, augustin.toma, ronald.xie, bowang.wang}@mail.utoronto.ca

Abstract

This paper describes our submission to the MEDIQA-Chat 2023 shared task for automatic clinical note generation from doctor-patient conversations. We report results for two approaches: the first fine-tunes a pre-trained language model (PLM) on the shared task data, and the second uses few-shot in-context learning (ICL) with a large language model (LLM). Both achieve high performance as measured by automatic metrics (e.g. ROUGE, BERTScore) and ranked second and first, respectively, of all submissions to the shared task. Expert human scrutiny indicates that notes generated via the ICL-based approach with GPT-4 are preferred about as often as human-written notes, making it a promising path toward automated note generation from doctor-patient conversations.[1]

1 Introduction

The growing burden of clinical documentation has emerged as a critical issue in healthcare, increasing job dissatisfaction and burnout rates among clinicians and negatively impacting patient experiences (Friedberg et al., 2013; Babbott et al., 2014; Arndt et al., 2017). On the other hand, timely and accurate documentation of patient encounters is critical for safe, effective care and communication between specialists. Therefore, interest in assisting clinicians by automatically generating consultation notes is mounting (Finley et al., 2018; Enarvi et al., 2020; Molenaar et al., 2020; Knoll et al., 2022).

To further encourage research on automatic clinical note generation from doctor-patient conversations, the MEDIQA-Chat Dialogue2Note shared task was proposed (Ben Abacha et al., 2023). Here, we describe our submission to subtask B: the generation of full clinical notes from doctor-patient dialogues. We explored two approaches; the first fine-tunes a pre-trained language model (PLM, §3.1), while the second uses few-shot in-context learning (ICL, §3.2). Both achieve high performance as measured by automatic natural language generation metrics (§4) and ranked second and first, respectively, of all submissions to the shared task. In a human evaluation with three expert physicians, notes generated via the ICL-based approach with GPT-4 were preferred about as often as human-written notes (§4.3).

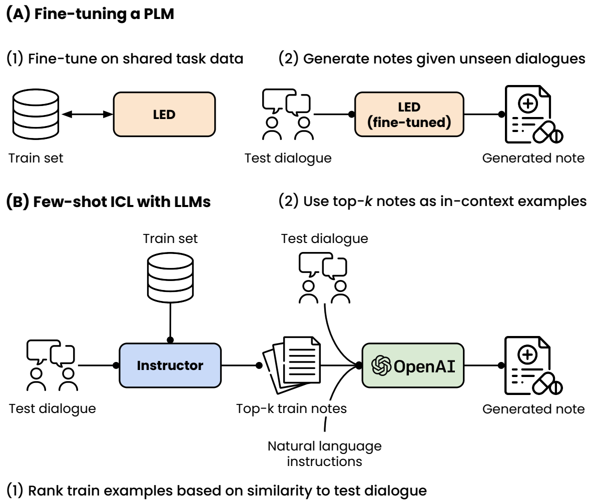


Figure 1: (**A**) Fine-tuning a pre-trained language model (PLM), Longformer-Encoder-Decoder (LED, Beltagy et al. 2020). (**B**) In-context learning (ICL) with large language models (LLMs). We rank train examples based on their similarity to the test dialogue using Instructor (Su et al., 2022a). Notes of the top-k most similar examples are then used as in-context examples to form a prompt alongside natural language instructions and fed to GPT-4 (OpenAI, 2023) to generate the clinical note.

2 Shared Task and Dataset

MEDIQA-Chat 2023 proposed two shared tasks:

1. **Dialogue2Note Summarization**: Given a conversation between a doctor and patient, the task is to produce a clinical note summarizing the conversation with one or more note sec-

[*]Core contributors. See author contributions
[1]https://github.com/bowang-lab/MEDIQA-Chat-2023

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 323–334
July 14, 2023 ©2023 Association for Computational Linguistics

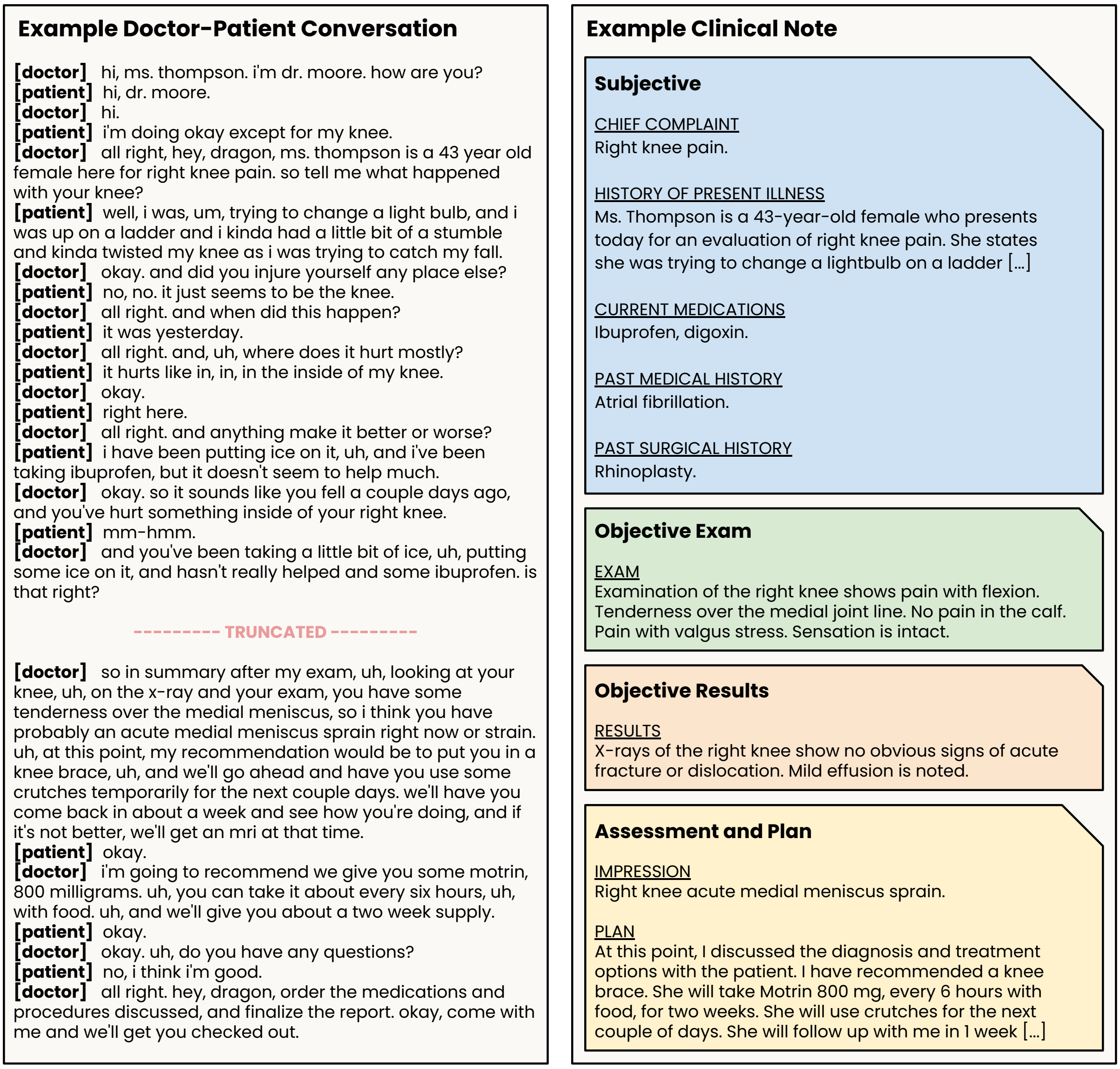


Figure 2: Example of a paired doctor-patient conversation and clinical note from the subtask B validation set. Dialogue has been lightly cleaned for legibility (e.g. remove trailing white space). Parts of the dialogue and note have been truncated. During evaluation, sections are grouped under one of four categories: "Subjective", "Objective Exam", "Objective Results", and "Assessment and Plan" (see §2.1 for details).

tions (e.g. Assessment, Past Medical History).

2. **Note2Dialogue Generation**: Given a clinical note, the task is to generate a synthetic doctor-patient conversation related to the information described in the note.

We focused on Dialogue2Note, which is divided into two subtasks. In subtask 'A' (Ben Abacha et al., 2023), the goal is to generate *specific sections* of a note given partial doctor-patient dialogues. In subtask 'B' (Yim et al., 2023), the goal is *full* note generation from complete dialogues. The remainder of the paper focuses on subtask B; see Appendix A for our approach to subtask A, which also ranks first of all submissions to the shared task.

2.1 Task definition

Each of the k examples consist of a doctor-patient dialogue, $D = d_1, \ldots, d_k$ and a corresponding clinical note, $N = n_1, \ldots, n_k$. The aim is to automatically generate a note n_i given a dialogue d_i. Each note comprises one or more sections, such as "Chief Complaint", and "Family history". During evaluation, sections are grouped under one of four categories: "Subjective", "Objective Exam", "Objective Results", and "Assessment and Plan".[2] See Figure 2 for an example doctor-patient conversation and clinical note pair.

[2]See here for the mapping

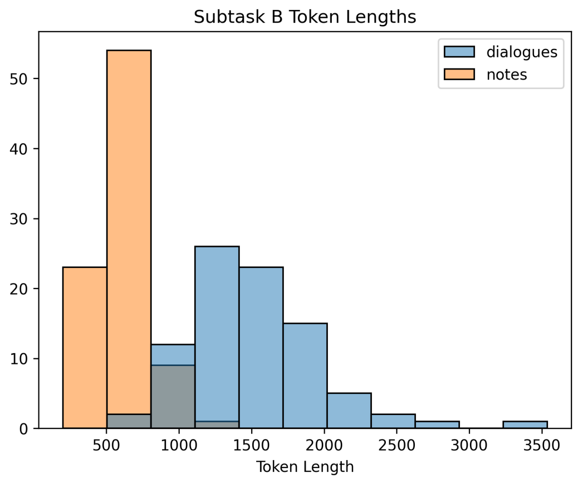


Figure 3: Histogram of token lengths for subtask B train and validation sets. Dialogues and notes were tokenized with `tiktoken` using the "gpt-4" encoding.

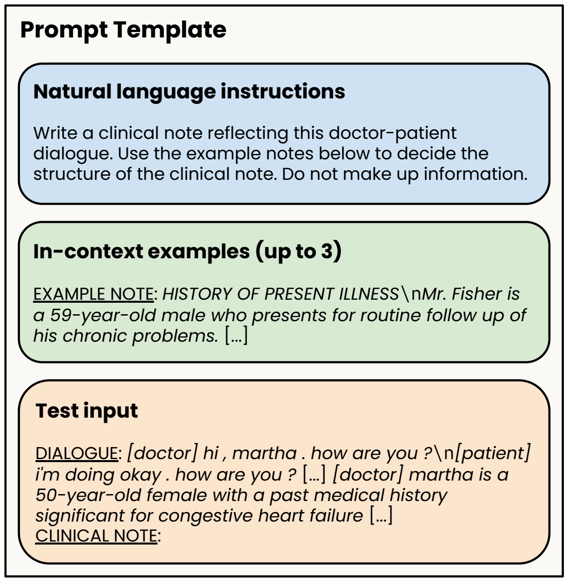


Figure 4: Prompt template for our in-context learning (ICL) based approach. Each prompt includes natural language instructions, up to 3 in-context examples, and an unseen doctor-patient dialogue as input.

2.2 Dataset

The dataset comprises 67 train and 20 validation examples, featuring transcribed dialogues from doctor-patient encounters and the resulting clinician-written notes. Each example is labelled with the 'dataset source', indicating the dialogue transcription system used to produce the note.

3 Approach

We take two high-performant approaches to the shared task. In the first, we fine-tune a pre-trained language model (PLM) on the provided training set (§3.1). In the second, we use in-context learning (ICL) with a large language model (LLM, §3.2).

3.1 Fine-tuning pre-trained language models

As a first approach, we fine-tune a PLM on the training set following a canonical, sequence-to-sequence training process (Figure 1 A; see Appendix C for details). Given the length of input dialogues (Figure 3), we elected to use Longformer-Encoder-Decoder (LED, Beltagy et al. 2020), which has a maximum input size of 16,384 tokens. We begin fine-tuning from a LED$_{\text{LARGE}}$ checkpoint tuned on the PubMed summarization dataset (Cohan et al., 2018), which performed best in preliminary experiments.[3] The model was fine-tuned using HuggingFace Transformers (Wolf et al., 2020) on a single NVIDIA A100-40GB GPU. Hyperparameters were lightly tuned on the validation set.[4]

3.2 In-context learning with LLMs

As a second approach, we attempt subtask B with ICL. We chose GPT-4 (OpenAI, 2023)[5] as the LLM and designed a simple prompt, which included natural language instructions and in-context examples (Figure 4). We limited the prompt size to 6192 tokens — allowing for 2000 output tokens, as the model's maximum token size is 8,192 — and used as many in-context examples as would fit within this token limit, up to a maximum of 3. We set the `temperature` parameter to 0.2 and left all other hyperparmeters of the OpenAI API at their defaults.

Natural language instructions During preliminary experiments, we found that GPT-4 was not overly sensitive to the exact phrasing of the natural language instructions in the prompt. We, therefore, elected to use short, simple instructions (Figure 4).

In-context example selection Each in-context example is a note from the train set. To select the notes, we first embed the dialogues of each training example and the input dialogue. Train dialogues are then ranked based on cosine similarity to the input dialogue; notes of the resulting top-k training examples are selected as the in-context examples (see Figure 1, B). Dialogues were embedded using Instructor (Su et al., 2022a), a text encoder that

[3] `https://huggingface.co/patrickvonplaten/led-large-16384-pubmed`

[4] See Appendix B.1 for details

[5] Specifically, the 03/14/2023 snapshot, "`gpt-4-0314`"

supports natural language instructions.[6] Lastly, we restricted in-context examples to be of the same 'dataset source' (see §2.2) as the input dialogue, hypothesizing that this may improve performance.[7]

3.3 Evaluation

Models are evaluated with the official evaluation script[8] on the validation set (as test notes are not provided). Generated notes are evaluated against the provided ground truth notes with ROUGE (Lin, 2004), BERTScore (Zhang et al., 2020) and BLEURT (Sellam et al., 2020). We report performance as the arithmetic mean of ROUGE-1 F1, BERTScore F1 and BLEURT-20 (Pu et al., 2021).

4 Results

4.1 Fine-tuning pre-trained language models

We present the results of fine-tuning LED in Table 1. Due to the non-determinism of the LED implementation,[9] we report the mean results of three training runs. Unsurprisingly, we find that scaling the model size from LED_{BASE} (12 layers, ∼162M parameters) to $\text{LED}_{\text{LARGE}}$ (24 layers, ∼460M parameters) leads to sizable gains in performance. Performance further improves by initializing the model with a checkpoint fine-tuned on the PubMed summarization dataset ($\text{LED}_{\text{LARGE-PubMed}}$). This is likely because (1) Dialouge2Note resembles a summarization task, and (2) text from PubMed is more similar to clinical text than is the general domain text used to pre-train LED.[10] Our submission to the shared task using this approach ranked second overall, outperforming the next-best submission by 2.7 average score; a difference comparable to the improvement in performance we see by doubling model size (see LED_{BASE} vs. $\text{LED}_{\text{LARGE}}$, Table 1).

4.2 In-context learning with LLMs

We present the results of ICL with GPT-4 in Table 2. We note several interesting trends in order of magnitude of impact. First, *selecting in-context examples based on the similarity of dialogues has a strong positive impact*, typically improving average score by 4 or more. *Using only notes as in-context examples, as opposed to dialogue-note pairs, also has a positive impact*, typically improving average score by ∼1. Surprisingly, *increasing the number of in-context examples had a marginal effect on performance*. Together these results suggest that the in-context examples' primary benefit is providing guidance with regard to the expected note structure, style and length. Finally, *filtering in-context examples to be of the same 'dataset source' as the input dialogue has a negligible impact on performance*.

Table 1: Fine-tuning LED. Mean and standard deviation (SD) of three training runs is shown. Scaling model size and pre-training on a related task improve performance. **Bold**: best scores.

Model	ROUGE-1 F1	BERTScore F1	BLEURT	Avg.
LED_{BASE}	$57.0_{0.4}$	$67.3_{0.1}$	$36.9_{0.0}$	53.8
$\text{LED}_{\text{LARGE}}$	$59.8_{0.2}$	$70.0_{0.6}$	$41.1_{0.8}$	57.0
$\text{LED}_{\text{LARGE-PubMed}}$	$\mathbf{61.7_{0.4}}$	$\mathbf{70.7_{0.2}}$	$\mathbf{41.5_{0.6}}$	**57.9**

The best strategy out-performs LED by almost 3 average score (60.8 vs. 57.9, see Table 1 & Table 2) and achieves first place of all submissions to the shared task, out-performing the runner up by > 9 average score. We conclude that (1) few-shot ICL with GPT-4, using as little as one example, is a performant approach for note generation from doctor-patient conversations, and (2) using the notes of semantically similar dialogue-note pairs is a strong strategy for selecting the in-context examples.

4.3 Human evaluation

Automatic evaluation metrics like ROUGE, BERTScore and BLEURT are imperfect and may not correlate with aspects of human judgment.[11] Therefore, we conducted an expert human evaluation to validate our results. To make annotation feasible, we conducted it on the validation set (20 examples) using the best performing fine-tuned model: $\text{LED}_{\text{LARGE-PubMed}}$ (Table 1), and best performing ICL-based approach: 3-shot, similar, note-only examples filtered by dataset type (Table 2).

Three senior resident physicians[12] were shown a ground truth note, a note generated by the fine-tuned model, and a note generated by the ICL-based approach for each example (presented in random order as clinical note 'A', 'B' and 'C')

[6] We used the following instructions: "Represent the Medicine dialogue for clustering: {dialogue}"

[7] Manual review revealed that dataset source was predictive of note structure & style; likely because it indicates which clinician or electronic health record system produced the note

[8] https://github.com/abachaa/MEDIQA-Chat-2023/blob/main/scripts/evaluate_summarization.py

[9] https://github.com/huggingface/transformers/issues/12482

[10] LED is initialized from BART (Lewis et al., 2020), which was pre-trained on a combination of text from Wikipedia and BooksCorpus (Zhu et al., 2015)

[11] See §6 for an extended discussion

[12] The three annotators are a subset of the authors who did not interact with the model or model outputs before annotation

Table 2: ICL with GPT-4. Mean of ROUGE-1 F1, BERTScore F1 and BLEURT for three runs is shown. Selecting in-context examples based on similarity to input dialogue improves performance. Dialogue-note pairs as in-context examples (omitting 3-shot results due to token length limits) underperforms notes only. Filtering in-context examples to be of the same 'dataset source' as the input dialogue has little effect. **Bold**: best scores. SD < 0.1 in all cases.

	Unfiltered				Filtered by dataset source			
Example selection strategy	0-shot	1-shot	2-shot	3-shot	0-shot	1-shot	2-shot	3-shot
Dialogue-note pairs as in-context examples								
random	52.2	54.5	53.9	–	–	54.8	54.5	–
similar dialogues	–	59.4	59.4	–	–	60.1	60.3	–
Notes only as in-context examples								
random	–	56.3	56.7	56.7	–	56.3	56.5	56.7
similar dialogues	–	60.7	60.6	60.4	–	**60.8**	60.4	**60.8**

Table 3: Human evaluation. Three physicians selected their preference from human written ground-truth notes (GT), notes produced by the fine-tuned model (FT) and notes produced by in-context learning (ICL). Win rate is % of cases where note was preferred, excluding ties.

	Preferred			Ties		Win rate (%)		
Physician	GT	FT	ICL	FT/ICL	All	GT	FT	ICL
1	9	1	4	2	4	64	7	29
2	5	0	14	0	1	26	0	74
3	9	0	6	0	5	60	0	40
Total	23	1	24	2	10	48	2	50

and asked to select which note(s) they preferred, given a dialogue and some simple instructions:

Instructions: Please asses the **clinical notes A**, **B** and **C** relative to the **provided doctor-patient dialogue**. For each set of notes, you should select which note you prefer ('A', 'B', or 'C'). If you have approximately equal preference for two notes, select ('A/B', 'B/C', or 'C/A'). If you have no preference, select 'A/B/C'. A 'good' note should contain *all* **critical**, *most* **non-critical** and *very little* **irrelevant** information mentioned in a dialogue:

- **Critical**: Items medico-legally required to document the diagnosis and treatment decisions whose absence or incorrectness may lead to wrong diagnosis and treatment later on, e.g. the symptom "cough" in a suspected chest infection consultation. This is the key information a note needs to capture correctly in order to not mislead clinicians.

- **Non-critical**: Items that should be documented in a complete note but whose absence will not affect future treatment or diagnosis, e.g. "who the patient lives with" in a consultation about chest infection.

- **Irrelevant**: Medically irrelevant information covered in the consultation, e.g. the pet of a patient with a suspected chest infection just died.

The definitions of critical, non-critical and irrelevant information are taken from previous work on human evaluation of generated clinical notes (Moramarco et al., 2022; Savkov et al., 2022).

In short, notes generated by ICL are strongly preferred over notes generated by the fine-tuned model and, on average, *slightly* preferred over the human-written notes (Table 3), validating the high performance reported by the automatic metrics. We note, however, that inter-annotator agreement is low and speculate why this might be in §6.

5 Related Work

Automated note generation from doctor-patient conversations has received increasing attention in recent years (Finley et al., 2018; Enarvi et al., 2020; Molenaar et al., 2020; Knoll et al., 2022). Different methods have been proposed, such as extractive-abstractive approaches (Joshi et al., 2020; Krishna et al., 2021; Su et al., 2022b) and fine-tuning PLMs (Zhang et al. 2021, similar to our approach in §3.1). Others have focused on curating data for training and benchmarking (Papadopoulos Korfiatis et al., 2022), including the use of LLMs to produce synthetic data (Chintagunta et al., 2021). Lastly, there have been efforts to improve the evaluation of generated clinical notes, both with automatic metrics (Moramarco et al., 2022) and human evaluation (Savkov et al., 2022). While recent literature has commented on the *potential* of ICL for note generation (Lee et al., 2023), our work is among the first to evaluate this approach rigorously.

6 Conclusion

We present our submission to the MEDIQA-Chat shared task for clinical note generation from doctor-patient dialogues. We evaluated a fine-tuning-based approach with LED and an ICL-based approach with GPT-4, ranking second and first, respectively, among all submissions. Human evaluation with three physicians revealed that notes produced by GPT-4 via ICL were strongly preferred over notes produced by LED and, on average, slightly preferred over human-written notes. We conclude that ICL is a promising path toward clinical note generation from doctor-patient conversations.

Limitations

Evaluation of generated text is difficult Evaluating automatically generated text, including clinical notes, is generally hard due to the inherently subjective nature of many aspects of output quality. Automatic evaluation metrics such as ROUGE and BERTScore are imperfect (Deutsch et al., 2022) and may not correlate with aspects of expert judgment. However, they are frequently used to evaluate model-generated clinical notes and do correlate with certain aspects of quality (Moramarco et al., 2022). To further validate our findings, we also conducted a human evaluation with three expert physicians (§4.3). As noted previously (Savkov et al., 2022), even human evaluation of clinical notes is far from perfect; inter-annotator agreement is generally low, likely because physicians have differing opinions on the importance of each patient statement and whether it should be included in a consultation note. We also found low inter-annotator agreement in our human evaluation and speculate this is partially due to differences in specialties among the physicians. Physicians 1 and 3, both from family medicine, had high agreement with each other but low agreement with physician 2 (cardiac surgery, see Table 3). Investigating better automatic metrics and best practices for evaluating clinical notes (and generated text more broadly) is an active field of research. We hope to integrate novel and performant metrics in future work.

Data privacy While our GPT-4 based solution achieves the best performance, it is not compliant with data protection regulations such as HIPAA; although Azure does advertise a HIPAA-compliant option.[13] From a privacy perspective, locally deploying a model such as LED may be preferred; however, our results suggest that more work is needed for this approach to reach acceptable performance (see Table 3). In either case, when implementing automated clinical note-generation systems, healthcare providers and developers should ensure that the whole system — including text-to-speech, data transmission & storage, and model inference — adheres to privacy and security requirements to maintain trust and prevent privacy violations in the clinical setting.

Ethics Statement

Developing an automated system for clinical note generation from doctor-patient conversations raises several ethical considerations. First, informed consent is crucial: patients must be made aware of their recording, and data ownership must be prioritized. Equitable access is also important; the system must be usable for patients from diverse backgrounds, including those with disabilities, limited technical literacy, or language barriers. Addressing issues of data bias and fairness are necessary to avoid unfair treatment or misdiagnosis for certain patient groups. The system must implement robust security measures to protect patient data from unauthorized access or breaches. Establishing clear lines of accountability for errors or harms arising from using an automated system for note generation is paramount. Disclosure of known limitations or potential risks associated with using the system is essential to maintain trust in the patient-physician relationship. Finally, ongoing evaluations are necessary to ensure that system performance does not degrade and negatively impact the quality of care.

Acknowledgements

This research was enabled in part by support provided by the Digital Research Alliance of Canada (alliancecan.ca) and Compute Ontario (www.computeontario.ca). We thank all internal and external reviewers for their thoughtful feedback, which improved earlier drafts of this manuscript.

Author Contributions

John Giorgi (JG), Augustin Toma (AT) and Ronald Xie (RX) led the project in general, including

[13] https://azure.microsoft.com/en-us/products/cognitive-services/openai-service#security

data cleaning and processing, model implementation, and running experiments. JG wrote the initial manuscript and designed the human evaluation with feedback from AT and RX. AT and RX recruited Sondra S. Chen, Kevin R. An and Grace X. Zheng, who served as expert physicians in the human evaluation. Bo Wang provided high-level feedback and advice.

References

Brian G. Arndt, John W. Beasley, M. Watkinson, Jonathan L Temte, Wen-Jan Tuan, Christine A. Sinsky, and Valerie J. Gilchrist. 2017. Tethered to the ehr: Primary care physician workload assessment using ehr event log data and time-motion observations. *The Annals of Family Medicine*, 15:419–426.

Stewart F. Babbott, Linda Baier Manwell, Roger L. Brown, Enid N. H. Montague, Eric S. Williams, Mark D. Schwartz, Erik P. Hess, and Mark Linzer. 2014. Electronic medical records and physician stress in primary care: results from the memo study. *Journal of the American Medical Informatics Association : JAMIA*, 21 e1:e100–6.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *ArXiv*, abs/2004.05150.

Asma Ben Abacha, Wen wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen. 2023. Overview of the mediqa-chat 2023 shared tasks on the summarization and generation of doctor-patient conversations. In *ACL-ClinicalNLP 2023*.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2283–2294, Dubrovnik, Croatia. Association for Computational Linguistics.

Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2021. Medically aware GPT-3 as a data generator for medical dialogue summarization. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 66–76, Online. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. Re-examining system-level correlations of automatic summarization evaluation metrics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 6038–6052, Seattle, United States. Association for Computational Linguistics.

Seppo Enarvi, Marilisa Amoia, Miguel Del-Agua Teba, Brian Delaney, Frank Diehl, Stefan Hahn, Kristina Harris, Liam McGrath, Yue Pan, Joel Pinto, Luca Rubini, Miguel Ruiz, Gagandeep Singh, Fabian Stemmer, Weiyi Sun, Paul Vozila, Thomas Lin, and Ranjani Ramamurthy. 2020. Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, pages 22–30, Online. Association for Computational Linguistics.

Gregory Finley, Erik Edwards, Amanda Robinson, Michael Brenndoerfer, Najmeh Sadoughi, James Fone, Nico Axtmann, Mark Miller, and David Suendermann-Oeft. 2018. An automated medical scribe for documenting clinical encounters. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 11–15, New Orleans, Louisiana. Association for Computational Linguistics.

Mark W. Friedberg, Peggy G. Chen, Kristin R Van Busum, Frances Aunon, Chau Pham, John P. Caloyeras, Soeren Mattke, Emma Pitchforth, Denise D. Quigley, Robert Henry Brook, Francis J. Crosson, and Michael A. Tutty. 2013. Factors affecting physician professional satisfaction and their implications for patient care, health systems, and health policy. *Rand health quarterly*, 3 4:1.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *ArXiv preprint*, abs/1211.3711.

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3755–3763, Online. Association for Computational Linguistics.

Tom Knoll, Francesco Moramarco, Alex Papadopoulos Korfiatis, Rachel Young, Claudia Ruffini, Mark Perera, Christian Perstl, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. User-driven research of medical note generation software. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 385–394, Seattle, United States. Association for Computational Linguistics.

Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary C. Lipton. 2021. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4958–4972, Online. Association for Computational Linguistics.

Peter Lee, Sébastien Bubeck, and Joseph Petro. 2023. Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine. *The New England journal of medicine*, 388 13:1233–1239.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Sabine Molenaar, Lientje Maas, Verónica Burriel, Fabiano Dalpiaz, and Sjaak Brinkkemper. 2020. Medical dialogue summarization for automated reporting in healthcare. *Advanced Information Systems Engineering Workshops*, 382:76–88.

Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Damir Juric, Jack Flann, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. Human evaluation and correlation with automatic metrics in consultation note generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5739–5754, Dublin, Ireland. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. PriMock57: A dataset of primary care mock consultations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 588–598, Dublin, Ireland. Association for Computational Linguistics.

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 751–762, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aleksandar Savkov, Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Anya Belz, and Ehud Reiter. 2022. Consultation checklists: Standardising the human evaluation of medical note generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 111–120, Abu Dhabi, UAE. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2022a. One embedder, any task: Instruction-finetuned text embeddings. *ArXiv*, abs/2212.09741.

Jing Su, Longxiang Zhang, Hamidreza Hassanzadeh, and Thomas Schaaf. 2022b. Extract and abstract with bart for clinical notes from doctor-patient conversations. In *Interspeech*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. The aci demo corpus: An open dataset for benchmarking the state-of-the-art for automatic note generation from doctor-patient conversations. *Submitted to Nature Scientific Data*.

Longxiang Zhang, Renato Negrinho, Arindam Ghosh, Vasudevan Jagannathan, Hamid Reza Hassanzadeh, Thomas Schaaf, and Matthew R. Gormley. 2021. Leveraging pretrained models for automatic summarization of doctor-patient conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3693–3712, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 19–27. IEEE Computer Society.

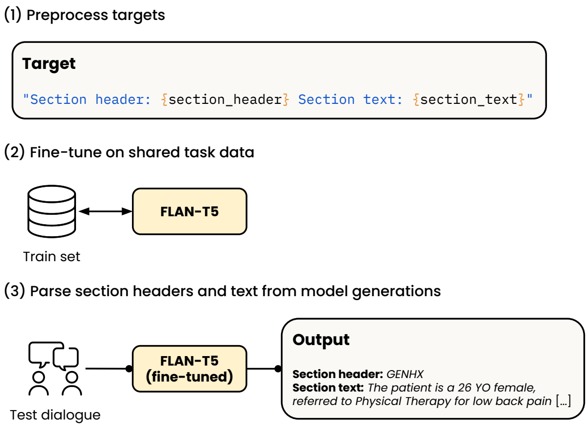


Figure 5: Fine-tuning FLAN-T5 (Chung et al., 2022) for subtask A. Before training, targets are preprocessed as "Section header: `{section_header}` Section text: `{section_text}`". After decoding, the section header and text are parsed using regular expressions.

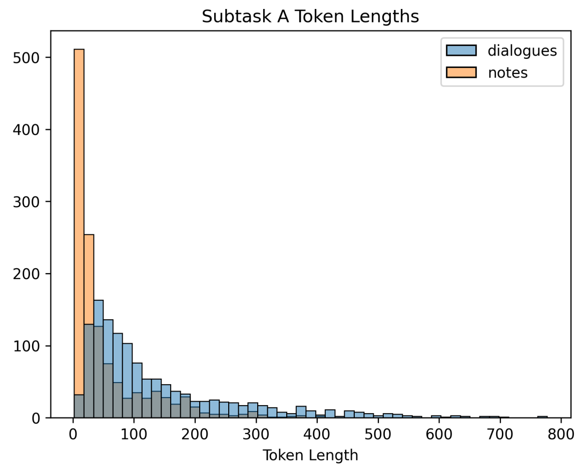


Figure 6: Histogram of token lengths for subtask A train and validation sets. Dialogues and notes were tokenized with HuggingFace Tokenizers using "`google/flan-t5-large`". Lengths greater than the 99th-percentile are omitted to make the plot legible.

A Subtask A

In subtask A of the Dialogue2Note Summarization shared task, given a partial doctor-patient dialogue, the goals are to: (1) predict the appropriate section header, e.g. "PASTMEDICALHX" and (2) generate that specific section of a note. We approached this task by fine-tuning a PLM on the provided training set, following a canonical, sequence-to-sequence training process (see Appendix C for details). In preliminary experiments, we found that the instruction-tuned FLAN-T5 (Chung et al., 2022) performed particularly well at this task.

We hypothesized that jointly learning to predict the section header and generate the section text would improve overall performance. To do this, we preprocessed the training set so the targets were of the form: "Section header: `{section_header}` Section text: `{section_text}`". After decoding, the section header and text were parsed using regular expressions and evaluated separately (Figure 5). Section header prediction was evaluated as the fraction of predicted headers that match the ground truth (accuracy), and section text was evaluated similarly to subtask B (see §3.3). In cases where the model output an invalid section header,[14] we replaced it with "GENHX" (general history), which tends to summarize the contents of the other sections. The model was fine-tuned on a single NVIDIA A100-40GB GPU. Hyperparameters were lightly tuned on the validation set (Table 4).

We present the results of our approach on the validation set in Table 5. Similar to subtask B (see §4.1), we find, perhaps unsurprisingly, that scaling the model size from FLAN-$T5_{BASE}$ (24 layers, ~250M parameters) to FLAN-$T5_{LARGE}$ (48 layers, ~780M parameters) leads to large improvements in performance. Performance is further improved by jointly learning to predict section headers and generate note sections. Our submission to the shared task based on this approach tied for first on section header prediction (78% accuracy), and ranked first for note section generation (average ROUGE-1, BERTScore and BLEURT F1-score of 57.9).

[14]In practice, we found that the fine-tuned model rarely, if ever, generates invalid section headers

B Subtask B

B.1 Hyperparameter tuning of LED

We lightly tuned the hyperparameters of $LED_{LARGE\text{-}PubMed}$ on the subtask B validation set against the average ROUGE-1 F1, BERTScore F1 and BLEURT-20 scores. The best hyperparameters obtained are given in Table 6. We used the same hyperparameters when fine-tuning LED_{BASE} and LED_{LARGE} in §4.1.

B.2 Post processing LEDs outputs

In practice, we found that the fine-tuned LED model sometimes produces invalid section headers; notably, this problem did not occur with the ICL-based approach using GPT-4. Therefore, we lightly post-processed LEDs outputs using a simple script that identifies section headers produced by the model not in the ground truth set and uses fuzzy

Table 4: Hyperparameters used with FLAN-T5 on the Dialogue2Note subtask A

Hyperparameter	Value	Comment
`max_source_length`	1024	truncate input sequences to this max length
`max_target_length`	512	truncate output sequences to this max length
`source_prefix`	*"Summarize the following patient-doctor dialogue. Include all medically relevant information, including family history, diagnosis, past medical (and surgical) history, immunizations, lab results and known allergies. You should first predict the most relevant clinical note section header and then summarize the dialogue. Dialogue:"*	instruction text prepended to all inputs
`train_batch_size`	8	batch size during training
`eval_batch_size`	12	batch size during inference
`learning_rate`	1e-4	learning rate during training
`optimizer`	AdamW (Loshchilov and Hutter, 2019)	optimizer used during training
`num_train_epochs`	20	total number of training epochs
`warmup_ratio`	0.1	proportion of training steps to linearly increase the learning rate to `learning_rate`
`lr_scheduler`	linear with warmup	learning rate linearly increased during first `warmup_ratio` fraction of train steps and linearly decreased to 0 afterwords
`weight_decay`	0.01	not applied to bias & `LayerNorm` weights
`label_smoothing`	0.1	label smoothing factor used during training
`bf16`	true	whether to use BF16 during training
`num_beams`	2	beam size used during beam search decoding

Table 5: Fine-tuning FLAN-T5. Accuracy of predicted section headers and score of generated note sections is shown. Jointly learning to predict section headers and generate notes improve performance. **Bold**: best scores.

		Note generation			
Model	Header prediction (%)	ROUGE-1 F1	BERTScore F1	BLEURT	Avg.
Random header	8.0	–	–	–	–
Majority header	22.0	–	–	–	–
FLAN-T5$_{\text{BASE}}$	71.0	40.1	70.5	52.7	54.5
FLAN-T5$_{\text{LARGE}}$	**79.0**	**49.8**	**74.5**	**58.0**	**60.8**
$\hookrightarrow$ *w/o header prediction*	–	48.0	74.3	57.6	59.9

string matching[15] to replace them with the closest valid header. For example, in one run, this process converted the (incorrect) predicted section header "HISTORY OF PRESENT" to the nearest valid header "HISTORY OF PRESENT ILLNESS".

C Fine-tuning Seq2Seq Models

When training the sequence-to-sequence (seq2seq) models for both subtask A (Appendix A) and B (§3.1), we followed a canonical supervised fine-tuning (SFT) process. We start with a pre-trained, encoder-decoder transformer-based language model (Vaswani et al., 2017). First, the encoder maps each token in the input to a contextual embedding. Then, the autoregressive decoder generates an output, token-by-token, attending to the outputs of the encoder at each timestep. Decoding proceeds until a special "end-of-sequence" token (e.g. `</s>`) is generated, or a maximum number of tokens have been generated. Formally, X is the *input* sequence, which in our case is a doctor-patient dialogue, and Y is the corresponding *output* sequence of length T, in our case a clinical note. We model the conditional probability:

$$p(Y|X) = \prod_{t=1}^{T} p(y_t|X, y_{<t}) \tag{1}$$

During training, we optimize over the model parameters θ the sequence cross-entropy loss:

$$\ell(\theta) = -\sum_{t=1}^{T} \log p(y_t|X, y_{<t}; \theta) \tag{2}$$

[15] We used `https://github.com/seatgeek/thefuzz`

Table 6: Hyperparameters used with Longformer-Encoder-Decoder (LED) on the Dialogue2Note subtask B

Hyperparameter	Value	Comment
max_source_length	4096	truncate input sequences to this max length
max_target_length	1024	truncate output sequences to this max length
source_prefix	*"Summarize the following patient-doctor dialogue. Include all medically relevant information, including family history, diagnosis, past medical (and surgical) history, immunizations, lab results and known allergies. Dialogue:"*	instruction text prepended to all inputs
train_batch_size	8	batch size during training
eval_batch_size	6	batch size during inference
learning_rate	3e-5	learning rate during training
optimizer	AdamW (Loshchilov and Hutter, 2019)	optimizer used during training
num_train_epochs	50	total number of training epochs
warmup_ratio	0.1	proportion of training steps to linearly increase the learning rate to learning_rate
lr_scheduler	linear with warmup	learning rate linearly increased during first warmup_ratio fraction of train steps and linearly decreased to 0 afterwords
weight_decay	0.01	not applied to bias & LayerNorm weights
label_smoothing	0.1	label smoothing factor used during training
fp16	true	whether to use FP16 during training
num_beams	4	beam size used during beam search decoding
min_length	100	min length of generated sequences
max_length	1024	max length of generated sequences
length_penalty	2.0	values > 0 promote longer output sequences
no_repeat_ngram	3	ngrams of this size can only occur once

maximizing the log-likelihood of the training data. As is common, we use *teacher forcing* during training, feeding previous ground truth inputs to the decoder when predicting the next token in the sequence. During inference, we generate the output using beam search (Graves, 2012). Beams are ranked by mean token log probability after applying a length penalty. Models are fine-tuned using the HuggingFace Transformers library.[16]

[16] https://github.com/huggingface/transformers/blob/main/examples/pytorch/summarization/run_summarization.py

Automatic Coding at Scale: Design and Deployment of a Nationwide System for Normalizing Referrals in the Chilean Public Healthcare System

Fabián Villena
Center for Mathematical Modeling
& Department of Computer Sciences
University of Chile
fabian.villena@uchile.cl

Matías Rojas
Center for Mathematical Modeling
University of Chile
matias.rojas.g@ug.uchile.cl

Felipe Arias
Center for Mathematical Modeling
University of Chile
felipe.arias.t@ug.uchile.cl

Jorge Pacheco
Dept. of Statistics and Health Information
Chilean Ministry of Health
jorge.pacheco@minsal.cl

Paulina Vera
Dept. of Statistics and Health Information
Chilean Ministry of Health
paulina.vera@minsal.cl

Jocelyn Dunstan
Dept. Computer Science & IMC
Pontifical Catholic University of Chile
jdunstan@uc.cl

Abstract

The disease coding task involves assigning a unique identifier from a controlled vocabulary to each disease mentioned in a clinical document. This task is relevant since it allows information extraction from unstructured data to perform, for example, epidemiological studies about the incidence and prevalence of diseases in a determined context. However, the manual coding process is subject to errors as it requires medical personnel to be competent in coding rules and terminology. In addition, this process consumes a lot of time and energy, which could be allocated to more clinically relevant tasks. These difficulties can be addressed by developing computational systems that automatically assign codes to diseases. In this way, we propose a two-step system for automatically coding diseases in referrals from the Chilean public healthcare system. Specifically, our model uses a state-of-the-art NER model for recognizing disease mentions and a search engine system based on Elasticsearch for assigning the most relevant codes associated with these disease mentions. The system's performance was evaluated on referrals manually coded by clinical experts. Our system obtained a MAP score of 0.63 for the subcategory level and 0.83 for the category level, close to the best-performing models in the literature. This system could be a support tool for health professionals, optimizing the coding and management process. Finally, to guarantee reproducibility, we publicly release the code of our models and experiments.

1 Introduction

The clinical text represents a significant proportion of patient's health records, commonly found in a non-structured format. These texts have particular challenges due to the extensive use of abbreviations, the variability of clinical language across medical specialties, and its restricted availability for privacy reasons (Dalianis, 2018). Due to the complexity of its analysis, this data is commonly discarded in projects that seek to support clinical decision-making (Kong, 2019).

Clinical coding involves mapping medical texts into codes using a controlled vocabulary consistent across different departments, hospitals, or even countries (Dong et al., 2022). The World Health Organization maintains an open, controlled vocabulary called the International Classification of Diseases (ICD), which is used in almost every country. Currently, the most widely used revision is the tenth (ICD-10) (World Health Organization, 2015), and they are developing its eleventh revision, which will include not only diseases (World Health Organization, 2023).

Regarding the Chilean public health system, the ICD-10 terminology is used for coding hospital discharges (morbidity coding by each healthcare provider) and deaths (mortality coding by the Ministry of Health). Having patients' data normalized using these controlled vocabularies enables the ability to summarize information automatically and

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 335–343
July 14, 2023 ©2023 Association for Computational Linguistics

not deal with the noisiness of free-text data. The already-digested information from the normalized data empowers data analysts who are not experts in NLP to add more complex information into their workflows.

The Waiting Time Management System (SIGTE, in Spanish) contains electronic records of referrals from the Chilean Waiting List, which is the system that manages the high demand existent for consultation by specialists (Ministerio de Salud de Chile, 2011). This data provided by 29 health services contain information about the medical diagnoses of patients but is not standardized (Báez et al., 2022).

As of November 2022, SIGTE recorded 25,374,491 waiting list referrals, of which 18,716,629 correspond to "new specialty referrals" and are associated with patient pathologies. Of these referrals, approximately 5,760,750 (30.7 %) have an ICD-10 code. This calculation was performed by searching for a regular expression formatted as an ICD-10 code in the free-text diagnosis fields.

Clinical experts perform the disease coding task manually, which is not optimal for several reasons. Firstly, since this process is subject to errors, medical personnel must have significant competence in coding rules and a thorough knowledge of specialized terminologies, such as ICD, which also get updated frequently. In other words, expert coding staff must be familiar with the clinical field, analytical and focused, and have fundamental skills for inspecting and analyzing highly specialized texts. In addition, manual coding is time-consuming (Yan et al., 2022), which could be optimized by a support system, and this time could be used for other tasks relevant to clinical decision-making.

These difficulties can be efficiently addressed using computational systems capable of automatically performing the coding task using NLP. Currently, most automatic coding systems are based on an end-to-end architecture based on deep learning techniques. Although these systems have boosted the performance of several coding tasks, they cannot incorporate context-specific rules, such as code priority, medical assumptions, code definition, and synonyms.

In this work, we developed an automated disease coding system, thus being able to code the entire historical waiting list in Chile, identifying a total of 18,716,629 referrals. Our system is based on two steps; first, the automatic extraction of diseases is addressed using a state-of-the-art NER model, and then, using a search engine, the most probable code for each disease found is identified. Finally, we explored the potential applications derived from this system and studied in more depth the most frequent diseases in the country today.

2 Related Work

The disease coding task involves transforming clinical texts, commonly written by physicians in a non-structured format, into codes following medical terminologies. This is not an easy task since a medical ontology such as ICD in Spanish has 14,668 codes, an example of extreme multi-label classification (Barros et al., 2022). We have identified two major groups of computational methods proposed to solve this task; rule-based coding and neural network-based coding.

2.1 Rule-based Models

This approach involves designing hand-crafted rules to represent and simulate the flow that clinical experts follow when assigning codes. Most of the studies are based on using regular expressions and keywords to transform diseases found in the text into their respective codes. However, these methods are not feasible since manually capturing all the relations between texts and codes is time-consuming and complex.

Different approaches based on machine learning have been proposed to address this issue. In this way, features extracted from statistical models such as decision trees and support vector machines, among others, are incorporated into the manual rules (Stanfill et al., 2010; Teng et al., 2022; Farkas and Szarvas, 2008). Another method is to create a list of synonyms of the original text to calculate a word distance with respect to the code descriptions of the terminology. Despite their disadvantages, these methods have yielded high results in the literature, effectively supporting manual coding performed by humans (Zhou et al., 2020).

2.2 Models based on neural networks

Deep learning-based methods have significantly improved the disease coding task in recent years. The advantage of using these models is that the healthcare-specific domain knowledge is no longer needed for the manual development of complex rules. In contrast, these methods can automatically build features powerful enough to capture the relationships between clinical texts and their respective codes.

Most proposed systems are based on posing the problem as a multi-label text classification task (Karimi et al., 2017; Mullenbach et al., 2018; Yu et al., 2019; Cao et al., 2020). Thus, the algorithm's input is text, while the output can be one or more codes associated with diseases. Unlike traditional text classification problems, this problem is considered extreme since the number of possible labels increases to thousands (depending on the terminology).

The main disadvantage of this approach is that manual coding requires incorporating context-specific rules, such as code priority, medical assumptions, code definition, and synonyms, among other types of information, to improve system performance. In the case of deep learning, this is not considered since the systems are commonly created using an end-to-end approach, meaning that no human knowledge is involved when creating the features or making the predictions.

To solve the previous problem, we followed another approach used in the literature, which consists of mixing the previous ideas using two sequential steps; the first one uses deep learning algorithms, while the second allows us to incorporate medical knowledge into the computational system. Firstly, we used a Named Entity Recognition model for automatically recognizing sequences of words in the text which are associated with diseases. Then, each disease found is associated with its most likely ICD-10 code, a task better known as Entity Linking (Kraljevic et al., 2021; Wiegreffe et al., 2019). Nowadays, the most commonly used methods for solving the NER task are based on deep neural networks such as transformers-based models or recurrent neural networks, while a frequent technique for assigning codes is to use distance algorithms or search engines to compare the diseases found with the code descriptions of the terminology.

2.3 Commercial Systems

A handful of commercial products offer information extraction from clinical data, including automatic coding. These products usually are delivered as services and offered by leading cloud providers such as Amazon Web Services with Amazon Comprehend Medical[1], Google Cloud with Google Cloud Healthcare Data Engine[2] and Microsoft Azure with Azure Cognitive Service for Language[3]. The problem with these services is that they do not offer automatic coding for languages other than English.

Data privacy concerns may arise from using this third-party software to extract patients' information. Some healthcare providers may prohibit sending data to systems outside the primary source due to potential cybersecurity issues.

3 Data and Methods

The Chilean Waiting List is characteristic of the the public healthcare system. This list arises due to the high demand for medical care and the limited capacity of the public health system to meet it. Entry on the waiting list begins when a patient goes to primary care or secondary care physician to treat pathology. The patient has two possible paths: if the pathology is included in the "Garantías Explícitas en Salud" (GES) program, the patient enters a process where his or her health problem is assured a maximum waiting time for medical attention. If the GES program does not cover the pathology, the referral is classified in one of these five options: New Specialty Consultations (CNE), Follow-up Consultations (CCE), Diagnostic Procedures (Proc), Surgical Intervention (IQ) and Complex Surgical Intervention (IQC). In any of these alternatives, the patient is placed on a waiting list and must wait a variable amount of time to receive medical attention from a specialist.

The Chilean Waiting List comprises 25,374,491 referrals, divided into five categories: 18,716,629 correspond to CNE type referrals, 4,391,257 to Proc type referrals, 2,222,545 to IQ type referrals, 39,266 to CCE type referrals, and finally, 4,794 to IQC type referrals. In particular, this work will focus on CNE-type referrals.

Within the Chilean Waiting database, 73 attributes are separated into two main types of sets. The first set corresponds to the attributes associated with the person (date of birth, sex, national identifier). In contrast, the second set corresponds to the administrative information associated with the referral given to the person (date of admission, date of discharge, the benefit provided, specialty, diagnostic suspicion, and diagnostic confirmation).

[1] https://aws.amazon.com/comprehend/medical/

[2] https://cloud.google.com/healthcare

[3] https://azure.microsoft.com/en-in/products/cognitive-services/language-service

For the analysis of the diagnoses present in the referrals, two free-text attributes representing medical diagnoses are considered: diagnostic suspicion and diagnostic confirmation. Table 1 shows the frequency of referrals according to medical specialty, while Table 2 shows corpus statistics of the texts analyzed.

Specialty	Referrals	Relative Freq. (%)
Ophthalmology	3,352,203	17.91
Otorhinolaryngology	1,270,563	6.79
Traumatology	1,066,814	5.70
Gynecology	991,166	5.30
General Surgery	982,500	0.05
Dermatology	762,758	4.08
Internal Medicine	703,844	3.76
Endodontics	662,607	3.54
Removable Prosthesis	652,604	3.49
Urology	605,425	3.23

Table 1: Top 10 specialties with the highest presence on the Chilean Waiting List for a medical appointment.

We used 10,000 referrals from the historical Chilean Waiting List to train the NER module for disease recognition. As detailed in (Báez et al., 2020; Báez et al., 2022), these referrals were previously consolidated by a team of clinical experts, thus constituting the so-called Chilean Waiting list corpus. In addition, we performed rounds of evaluation of the NER performance, identifying diseases that the model could not identify. Thus, these diseases were incorporated as new examples of the model training process.

4 Proposed System

To code the narratives, we first used a NER model to automatically recognize sequences of words in the text associated with diseases. Then, each disease found is associated with its most likely ICD-10 code through a search engine. Figure 1 shows an overview of our proposed system.

4.1 NER Model

As shown in Figure 1, the input of our system is the referral written by the physician in an unstructured format. These texts are used as input for the automatic disease recognition model. In particular, this NER model is based on the work proposed in (Rojas et al., 2022a), where a simple but highly effective architecture for medical entity recognition is introduced. This model, named Multiple LSTM-CRF (MLC), is a deep neural network system composed of three main modules, emphasizing the impact of using domain-specific contextualized embeddings.

The first layer of the MLC approach, the "stacked embedding layer", transforms the texts associated with the diagnoses into a vector representation using character-level contextual embeddings and static word embeddings, both trained in the clinical domain. Then, in the encoding layer, a recurrent neural network is used to obtain long-distance dependencies between words in the sentence, thus obtaining a better context to improve the previous layer's representations. Finally, the classification layer assigns the most probable label to each word in the diagnosis using the CRF algorithm, identifying which parts of the text correspond to the beginning and end of a disease.

Regarding the experimental setup, the disease model was trained to 150 epochs using an SGD optimizer with mini-batches of size 32 and a learning rate of 0.1. As mentioned, to encode sentences, we used two types of representations; a 300-dimensional word embedding model trained on the Chilean Waiting List corpus[4] and character-level contextualized embeddings retrieved from the Clinical Flair model (Rojas et al., 2022b). To implement the model and perform our experiments, we used the Flair framework, widely used by the NLP research community (Akbik et al., 2019).

4.2 Search Engine

The output of the NER step is a list containing all the diseases mentioned in the referral. This second module aims to assign an ICD-10 code to each disease found, which can be used later for clinical decisions or management. The assignment of the ICD-10 code is done through a search engine tool based on Elasticsearch[5], an open-source search and analytics engine. This system can assign similarities between the mention of the disease and each of the codes of the ICD-10 tabular list.

Unlike the algorithms of distance comparison between words, this search engine has an index that contains each of the ICD-10 diseases represented through a series of synonymous sentences extracted from different sources of information, simulating

[4] `https://zenodo.org/record/3924799`

[5] Registered trademark of Elasticsearch B.V. Available at `https://www.elastic.co/elasticsearch/`

Specialty	Number of tokens (std)	Number of sentences (std)	Tokens per sentence
Infectology	28.59 (48.04)	1.50 (1.61)	18.94
Medical Oncology	20.09 (42.12)	1.15 (0.66)	17.40
Diabetology	19.03 (33.07)	1.33 (1.31)	14.22
Pediatric Rheumatology	15.05 (30.95)	1.19 (0.75)	12.61
Oral Pathology	14.54 (24.51)	1.17 (0.69)	12.34
Neonatology	12.92 (31.31)	1.10 (0.55)	11.64
Pediatric Hemato-Oncology	12.70 (26.74)	1.16 (0.75)	10.94
Pediatric Plastic Surgery	17.51 (16.33)	1.25 (0.51)	10.88
Pediatric Gynecology	13.01 (22.61)	1.22 (0.82)	10.61
Pediatric Traumatology	12.27 (18.39)	1.16 (0.74)	10.55

Table 2: Top 10 specialties with the highest number of tokens per sentence on average in Chilean Waiting List.

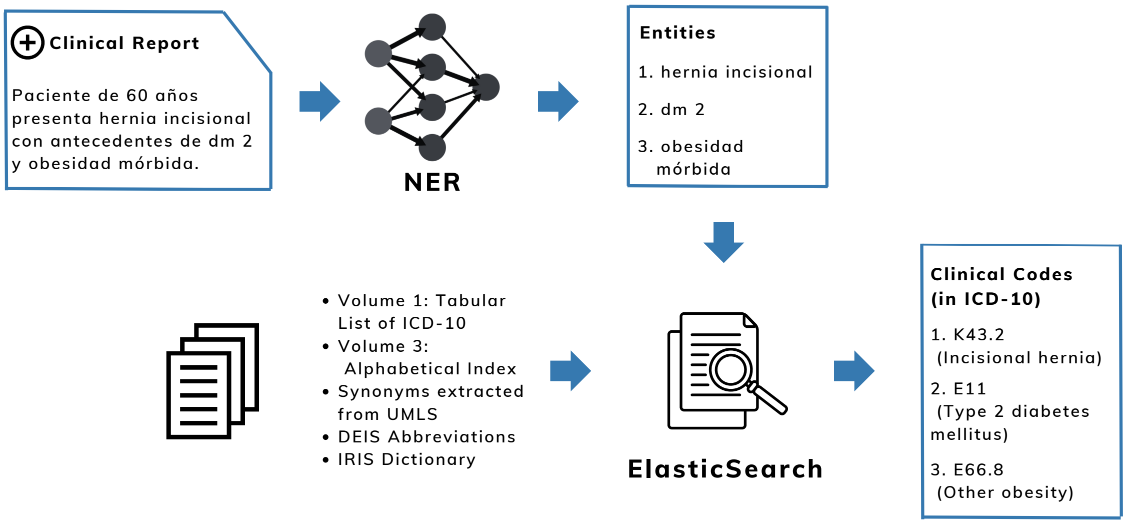


Figure 1: Overview of the proposed disease coding system.
TRANSLATION Clinical report: A 60-year-old patient presented with an incisional hernia with a history of dm 2 and morbid obesity. Entities: 1. incisional hernia, 2. dm 2, 3. morbid obesity.

in a better way the process followed by clinical experts to determine the code of a disease.

For example, in the index, the code "K02.2" contains the canonical code description "Caries of cementum" and multiple synonymous definitions, such as "Cement caries" and "Root caries". This is important as disease mentions found in unstructured diagnoses are rarely equivalent to the exact definition.

The sources of information used for the extraction of synonymous disease definitions were as follows:

Tabular list of ICD-10 terminology: This is the basis of the index, which tells us which codes we will assign to the disease mentions.

Alphabetical index of ICD-10 terminology: The guide for the manual assignment of codes to diseases and was obtained using the "web scraping" technique from the website of the Spanish Ministry of Health [6].

IRIS dictionary: It maps natural language sentences to an ICD-10 code. This dictionary was built from the mortality coding rounds con-

[6] https://eciemaps.mscbs.gob.es/ecieMaps/browser/index_10_2008.html

ducted in the Chilean Department of Statistics and Health Information.

UMLS: Spanish definitions from multiple vocabularies were extracted from the metatresaurus database.

DEIS abbreviations: Manually constructed list of abbreviations and their expansions.

4.3 Experiments

In our experiments, we measure how well the predictions made by the model fit compared to the decisions made by clinical experts. In this way, a subset of the referrals described in Section 3 was selected to be manually coded by a team of two clinical coders. The manual annotation process and system validation steps are provided below.

4.3.1 Manual coding

The clinical experts carried out the annotation process using Excel software. For this purpose, a file containing a unique identifier for each referral, the associated diagnostic suspicion, and a blank column for the actual coding was provided to the coders. This way, the expert coders identified disease codes in 1,188 clinical narratives from the Chilean Waiting List for a new specialty.

It is important to mention that in this process, codes were identified at the referral level, not at the entity level; therefore, it is not possible to determine the performance of the NER model in this experiment. In future work, specialized software such as INCEpTION, could be used, as proposed in the work of (Báez et al., 2020). This software would make it possible to identify which parts of the text refer to diseases. On the other hand, only diseases were coded, but future research could extend it to new entity types, such as clinical procedures or clinical findings.

4.3.2 Metric

The Mean Average Precision (MAP) metric is used to evaluate the performance of our coding system. This metric is widely used in works that address the same automatic coding task. This metric is defined as follows:

$$AveP = \frac{\sum(P(k) \cdot rel(k))}{\text{number of relevant documents}}, \quad (1)$$

where P(k) represents the precision at position k, and rel(k) is an indicator function equal to 1 if the element in rank k is a relevant document and 0 otherwise.

The MAP is computed using the Python implementation of the TREC evaluation tool, `trectools`, by (Palotti et al., 2019), where an adaptation was applied, in which the coded diagnoses have to be ordered based on a ranking, which for this work is considered the order in which the mention was found and subsequently the code was assigned.

5 Results

Orthodontics (3.28)	Obstetrics (1.01)
Endodontics (2.52)	Plastic Surgery (0.93)
Oral Rehab. (1.26)	Pediatrics (0.84)
Nephrology (1.26)	P. Physical Med. (0.84)
P. Dentistry (1.26)	Dental Operatory (0.34)
Psychiatry (1.18)	Maxillo. Radiol. (0.34)
P. Urology (1.01)	STI (0.084)

P. = Pediatric, Maxillo. Radiol. = Maxillofacial Radiology, STI = Sexually transmitted infections.

Table 3: Specialties with a perfect MAP score. The relative frequency (in percentage) of referrals in the dataset is in parentheses.

5.1 Coding Performance

The ICD-10 consists of a solitary coded catalog composed of categories with three characters, each of which can be additionally subdivided into as many as ten subcategories of four characters.

We computed the MAP metric over the test set at the category (e.g. K02) and subcategory (e.g. K02.2) levels. We achieved a MAP of 0.83 for the category and 0.63 for the subcategory level.

To underline the difficulty of achieving outstanding results in coding, we analyzed the results obtained only by clinical experts. The expert coders achieved an agreement MAP of 0.75 for subcategory and 0.83 for category level. Several reasons, such as the subjectivity in clinical judgment, the complexity of coding guidelines, the evolving nature of medicine, the time pressure and workload, personal bias, and lack of standardization, could explain the low agreement score.

6 Error Analysis

To better understand the errors made by our coding system, we performed a granular analysis of the scores obtained among the different specialties in the corpus. Tables 3 and 4 show the top 14 best and

Specialty	MAP at Category Level	Relative freq. in %
Neurology	0.68	2.10
Immunology	0.67	0.84
Geriatrics	0.60	0.93
Pediatric Gastroenterology	0.58	0.84
Cardiothoracic surgery	0.53	0.84
Radiation therapy	0.50	0.84
Pediatric Family Medicine	0.45	0.84
Hematology	0.42	1.18
Diabetology	0.35	1.01
Pediatric Traumatology	0.29	0.93

Table 4: Top 10 worst scores according to the specialties.

10 worst scores according to the specialties. We noted that in the top 14 best specialties the diagnostic suspicions registered in the referral were written straightforwardly and were specific diagnoses, such as "lipoma", "caries", and "nephrolithiasis", avoiding other clinical information like comorbidity, medication intake, or some other medical history. Furthermore, it can be noted that half of these referrals are related to dental diagnosis.

On the other hand, the top 10 worst specialties share in common that most of the diagnoses are very unspecific, with the incorporation of non-medical information such as the patient's phone number, patient's address, physician's name, the specialty the patient is referred to and information about comorbidity. Besides, several referrals are without a diagnosis but with the text "unspecific consultation" or "other".

7 Model Deployment and Use Cases

Due to internal regulations, we could not send patients' data to third-party systems such as cloud providers or academic supercomputing clusters (Ministerio Secretaría Regional de la Presidencia, 1999). For this reason, we deployed the whole coding system on-premise on a bare metal machine with a GPU compute module (NVIDIA RTX A4000[7]) to process the coding requests from the whole department efficiently.

The complete automatic coding system was deployed as a pair of microservices running inside containers to ease portability. One container hosts the NER module and exposes an API as a web service listening to disease-mention detection requests. The other container consists of the recommended implementation of the Elasticsearch software, which also exposes its API as a web service listening to mention-coding requests.

To code the waiting list and schedule recurrent coding when new data arrives, we used the KNIME[8] software, a visual-programming data mining platform. We chose this software because of its ease of use for non-expert developers. The workflow starts with the raw waiting list, which is first passed through the NER module to detect disease mentions, and then each mention is sent to the coding module to assign the most relevant code.

The automatic coding result from the workflow mentioned above is persisted on a table inside a database that stores each disease mention for each referral along with the predicted code from the system.

8 Conclusions

In this work, we created a nationwide system to improve the management of the Chilean public healthcare system. Specifically, we addressed the challenge of creating an automated system to code the diseases present in the Chilean Waiting List referrals. We developed and validated a model based on two steps: a NER model to recognize disease mentions and a search engine based on Elasticsearch to assign the codes to each disease. This mapping system was enriched with several terminology resources used in real life by manual coders to assign codes, thus partially simulating

[7]The compute module has 16 GB of GPU memory and 6.144 CUDA cores. More information at https://www.nvidia.com/en-us/design-visualization/rtx-a4000/

[8]Registered trademark of KNIME GmbH. Available at https://www.knime.com/

the pipeline followed by these professionals when solving this task.

The system allowed us to assign codes to 18,716,629 referrals, thus demonstrating its efficiency and effectiveness. The performance obtained in our experiments was 0.83 according to the MAP score, which is close to the most advanced systems currently in the coding task. The model was deployed into production in the Department of Health Statistics and Information Systems of the Ministry of Health of Chile.

The use of this system could be an important support for the management of waiting lists. In addition, since 75% of the Chilean population is in the public healthcare system, the analysis of the new specialty consultations can be used for epidemiological studies, such as the one done on the incidence of psoriasis (Lecaros et al., 2021).

Acknowledgements

This work was funded by ANID Chile: Basal Funds for Center of Excellence FB210005 (CMM); Millennium Science Initiative Program ICN17_002 (IMFD) and ICN2021_004 (iHealth), Fondecyt grant 11201250, and National Doctoral Scholarship 21220200. We also acknowledge Daily Piedra and Marcela Carmona for their work on annotating and coding the test dataset.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Pablo Báez, Antonia Paz Arancibia, Matías Ignacio Chaparro, Tomás Bucarey, Fredy Núñez, and Jocelyn Dunstan. 2022. Procesamiento de lenguaje natural para texto clínico en español: el caso de las listas de espera en chile. *Revista Médica Clínica Las Condes*, 33(6):576–582.

Pablo Báez, Felipe Bravo-Marquez, Jocelyn Dunstan, Matías Rojas, and Fabián Villena. 2022. Automatic extraction of nested entities in clinical referrals in spanish. *ACM Trans. Comput. Healthcare*, 3(3).

Pablo Báez, Fabián Villena, Matías Rojas, Manuel Durán, and Jocelyn Dunstan. 2020. The Chilean waiting list corpus: a new resource for clinical named entity recognition in Spanish. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 291–300, Online. Association for Computational Linguistics.

Jose Barros, Matías Rojas, Jocelyn Dunstan, and Andres Abeliuk. 2022. Divide and conquer: An extreme multi-label classification approach for coding diseases and procedures in spanish. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 138–147.

Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong. 2020. HyperCore: Hyperbolic and co-graph representation for automatic ICD coding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3105–3114, Online. Association for Computational Linguistics.

Hercules Dalianis. 2018. *Clinical Text Mining*, first edition. Springer International Publishing.

Hang Dong, Matúš Falis, William Whiteley, Beatrice Alex, Joshua Matterson, Shaoxiong Ji, Jiaoyan Chen, and Honghan Wu. 2022. Automated clinical coding: what, why, and where we are? *npj Digital Medicine*, 5(1).

Richárd Farkas and György Szarvas. 2008. Automatic construction of rule-based ICD-9-CM coding systems. *BMC Bioinformatics*, 9(S3).

Sarvnaz Karimi, Xiang Dai, Hamed Hassanzadeh, and Anthony Nguyen. 2017. Automatic diagnosis coding of radiology reports: A comparison of deep learning and conventional classification methods. In *BioNLP 2017*, pages 328–332, Vancouver, Canada,. Association for Computational Linguistics.

Hyoun-Joong Kong. 2019. Managing unstructured big data in healthcare system. *Healthcare Informatics Research*, 25:1.

Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A. Folarin, Angus Roberts, Rebecca Bendayan, Mark P. Richardson, Robert Stewart, Anoop D. Shah, Wai Keong Wong, Zina Ibrahim, James T. Teo, and Richard J.B. Dobson. 2021. Multi-domain clinical natural language processing with medcat: The medical concept annotation toolkit. *Artificial Intelligence in Medicine*, 117:102083.

C Lecaros, J Dunstan, F Villena, DM Ashcroft, R Parisi, CEM Griffiths, S Härtel, JT Maul, and C De la Cruz. 2021. The incidence of psoriasis in chile: an analysis of the national waiting list repository. *Clinical and Experimental Dermatology*, 46(7):1262–1269.

Ministerio de Salud de Chile. 2011. Norma técnica parael registro de las listas de espera.

Ministerio Secretaría Regional de la Presidencia. 1999. Ley 19628 sobre protección de la vida privada.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana. Association for Computational Linguistics.

João Palotti, Harrisen Scells, and Guido Zuccon. 2019. Trectools: An open-source python library for information retrieval practitioners involved in trec-like campaigns. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 1325–1328, New York, NY, USA. Association for Computing Machinery.

Matias Rojas, Felipe Bravo-Marquez, and Jocelyn Dunstan. 2022a. Simple yet powerful: An overlooked architecture for nested named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2108–2117, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Matías Rojas, Jocelyn Dunstan, and Fabián Villena. 2022b. Clinical flair: A pre-trained language model for Spanish clinical natural language processing. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 87–92, Seattle, WA. Association for Computational Linguistics.

Mary Stanfill, Margaret Williams, Susan Fenton, Robert Jenders, and William Hersh. 2010. A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association : JAMIA*, 17:646–51.

Fei Teng, Yiming Liu, Tianrui Li, Yi Zhang, Shuangqing Li, and Yue Zhao. 2022. A review on deep neural networks for icd coding. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1.

Sarah Wiegreffe, Edward Choi, Sherry Yan, Jimeng Sun, and Jacob Eisenstein. 2019. Clinical concept extraction for document-level coding. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 261–272, Florence, Italy. Association for Computational Linguistics.

World Health Organization. 2015. *International statistical classification of diseases and related health problems*, 10th revision, fifth edition, 2016 edition. World Health Organization.

World Health Organization. 2023. *International statistical classification of diseases and related health problems*, 11th revision edition. World Health Organization.

Chenwei Yan, Xiangling Fu, Xien Liu, Yuanqiu Zhang, Yue Gao, Ji Wu, and Qiang Li. 2022. A survey of automated international classification of diseases coding: development, challenges, and applications. *Intelligent Medicine*, 2(3):161–173.

Ying Yu, Min Li, Liangliang Liu, Zhihui Fei, Fang-Xiang Wu, and Jianxin Wang. 2019. Automatic icd code assignment of chinese clinical notes based on multilayer attention birnn. *Journal of Biomedical Informatics*, 91:103114.

Lingling Zhou, Cheng Cheng, Dong Ou, and Hao Huang. 2020. Construction of a semi-automatic icd-10 coding system. *BMC Medical Informatics and Decision Making*, 20.

Building blocks for complex tasks: Robust generative event extraction for radiology reports under domain shifts

Sitong Zhou[♠] **Meliha Yetisgen**[◇] **Mari Ostendorf**[♠]
[♠]University of Washington, Electrical & Computer Engineering
[◇]University of Washington, Biomedical and Health Informatics
{sitongz,melihay,ostendor}@uw.edu

Abstract

This paper explores methods for extracting information from radiology reports that generalize across exam modalities to reduce requirements for annotated data. We demonstrate that multi-pass T5-based text-to-text generative models exhibit better generalization across exam modalities compared to approaches that employ BERT-based task-specific classification layers. We then develop methods that reduce the inference cost of the model, making large-scale corpus processing more feasible for clinical applications. Specifically, we introduce a generative technique that decomposes complex tasks into smaller subtask blocks, which improves a single-pass model when combined with multitask training. In addition, we leverage target-domain contexts during inference to enhance domain adaptation, enabling use of smaller models. Analyses offer insights into the benefits of different cost reduction strategies.

1 Introduction

Radiology reports contain a diverse and rich set of clinical abnormalities documented by radiologists during their interpretation of the images. Automatic extraction of radiological findings would enable a wide range of secondary use applications to support diagnosis, triage, outcomes prediction, and clinical research (Lau et al., 2020). We adopt an event-based schema to capture both indications, the reason for radiology exams, and abnormal findings documented in radiology reports. We use an annotated a corpus of reports from three distinct radiology examination modalities (Lybarger et al., 2022): Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), and Computed Tomography (CT). Each event consists of a trigger, words that indicate a particular indication or finding (e.g., lesion), and a set of attributes (assertion, anatomy, characteristics, size, size trend, size count) that describe this indication or finding. Manual annotation of radiology reports is costly, therefore we hope models can generalize across different exam modalities. In this work, we define each modality in our annotated corpus as a domain and study cross-domain generalization among different modalities for the task of event extraction. Event extraction can be conceptualized as a series of subtasks, which include entity detection (trigger and attribute spans), relation detection (between triggers and attributes), and entity normalization (fine-grained labels on spans). In our experiments, we focus on trigger detection and anatomy attribute extraction with normalized labels.

To enhance generalization capabilities, some studies employ generative models and formulate tasks as question answering and using texts to represent both inputs and outputs (Raffel et al., 2020; Xie et al., 2022), as opposed to allowing the model to solely learn task intent from training data (Eberts and Ulges, 2019; Lybarger et al., 2023).

The exceptional performance of generative models often rely on large model size; however, in real-time inference for processing large-scale clinical notes, reducing inference costs is crucial. To address this need, for task inference, we want to reduce the number of decoding passes and employ smaller models. Due to the high inference costs, there is a desire to merge these subtasks and decode them in a single step. However, the generative approach has been reported to perform better on solving subtasks individually but worsen when combined, a phenomenon referred to as the compositionality gap (Press et al., 2022). This gap can be exacerbated under domain shifts when models learn subtasks jointly, as interdependence of subtasks may vary across domains.

While large language models (LLMs) mitigate the compositionality gap using reasoning steps (Wei et al., 2022; Press et al., 2022) to solve complex questions by decomposing them into smaller ones, there is limited work on reasoning for highly specialized domains (such as medical event extrac-

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 344–357
July 14, 2023 ©2023 Association for Computational Linguistics

tion) or with smaller models. In this paper, we reduce the compositionality gap for smaller models through formatting of complex tasks into easier subtasks as blocks. This approach teaches models how to solve individual subtasks independently and how to assemble them for solving more complex tasks.

The generative model enables seamless integration of supplementary contexts into the prompt, which compensates for the knowledge gap to larger models and reduces inference costs. To aid in domain adaptation, we extract target domain contexts that are likely to be helpful for the task, instead of retrieving similar contexts for general purpose. Specifically, to assist with anatomy normalization tasks, we employ an unsupervised extractor to acquire pertinent contexts that likely contain anatomical information from the same document and/or unannotated text from the same domain. This process can either disambiguate the original single-sentence input or provide anatomy-related hints that the model can utilize. To avoid introducing source-domain-specific reliance on the contexts, we incorporate the contexts only at the inference stage.

In our experiments, we first study domain shift for extracting radiology finding events and observe that cross-domain performance decline is more pronounced for knowledge-intensive anatomy normalization tasks, while detecting entity spans exhibits relatively stable performance. We demonstrate that building subtask blocks and assembling them as sequences to solve complex tasks can reduce the compositionality gap in smaller models. We show that incorporating target-domain contexts in domain adaptation can compensate for reduced model sizes, enabling good performance with smaller models.

2 Task

2.1 Event extraction for radiology findings

Our event scheme includes three event types: i) *Indication* is the reason for the imaging (e.g. motor vehicle accident or cancer staging); ii) *Lesion* captures lesions uncovered by the exam (e.g. mass or tumor); and iii) *Medical Problem* characterizes non-lesion abnormalities (e.g. fracture or hernia). Each finding event is characterized by an event trigger and set of attributes (assertion, anatomy, characteristics, size, size-trend, count). In this work, we focus only on extracting events with normalized anatomical information and investigate cross-domain generalization for different examination modalities. Figure 1 presents a *Lesion* event example. The event extraction process can be broken down into four subtasks: (1) Trigger span extraction (e.g., "density"), (2) Trigger type classification (e.g., "density" - Lesion), (3) Anatomy span extraction (e.g., "left lobe of liver" associated with the trigger "density"), and (4) Anatomy normalization to parent-child anatomy categories (e.g., "left lobe of liver" - Parent: Hepato-Biliary, Child - Liver). See Appendix A for the full list of hierarchical parent-child anatomy categories.

We evaluate event extraction performance using the F1 metrics by Lybarger et al. (2021). Our assessment of the trigger extraction is based on the span overlap and the event type match with respect to the gold standard labels. The anatomy extraction is first assessed at the span level. A correct anatomy prediction is associated with a correct predicted trigger and anatomy span overlap with the gold standard labels. Additionally, we evaluate anatomy extraction based on the normalization level, irrespective of their spans. A match between the predicted anatomy entity and the gold label indicates that the trigger is matched, and the normalized anatomy category is equal.

2.2 Domain shifts across radiology modalities

Our research investigates cross-domain generalization among three distinct radiology examination modalities: MRI, PET, and CT. These exam modalities are performed for different reasons with different technologies and the resulting radiology reports differ in terms of level of details as well as anatomy distribution. While CT and MRI scans allow radiologists to view structures inside the body, a PET scan, on the other hand, captures how tissues in the body work on the cellular level and shows unusual activity. MRI scans very frequently involve neurological exams. The most common use of PET scans is to diagnose or monitor certain cancer types. In our experiments, we define each modality as a domain. We use PET as the target domain, and train on three domains separately to evaluate both in-domain and cross-domain scenarios.

3 Method

3.1 Generative event extraction with T5

In order to improve the model's generalization capabilities over BERT-based alternatives (Lybarger et al., 2023; Eberts and Ulges, 2019), we struc-

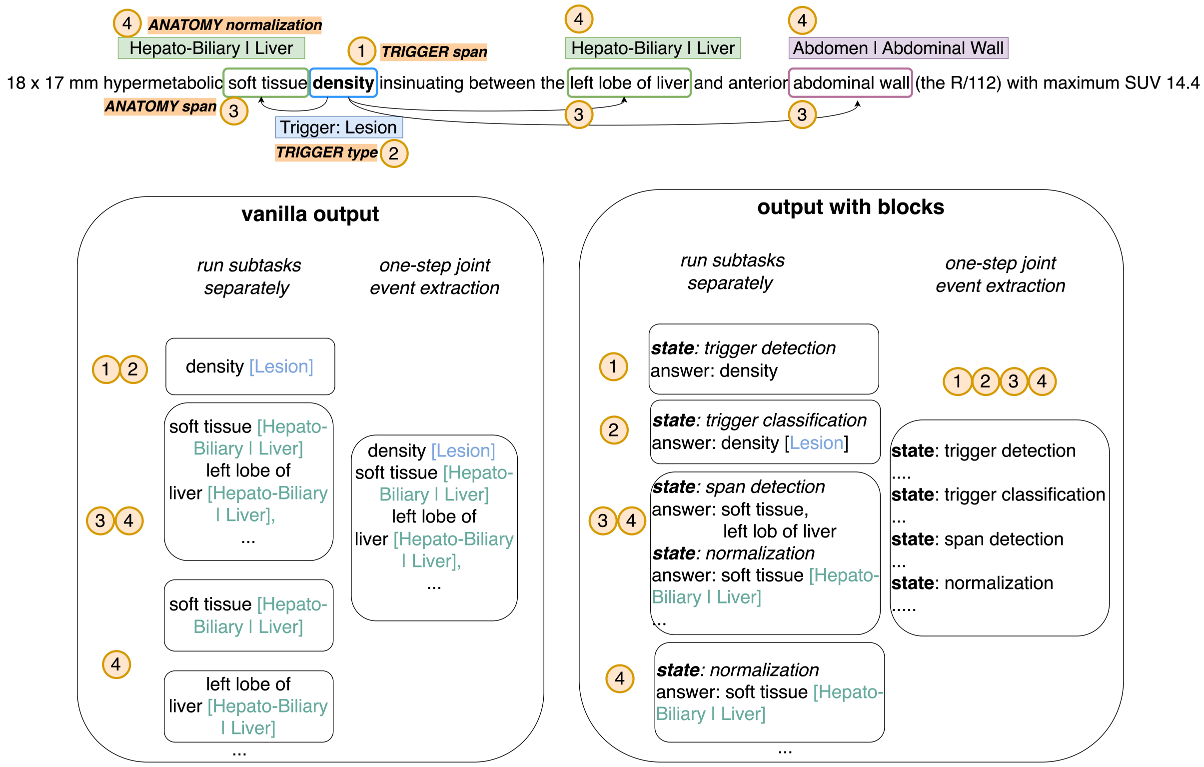


Figure 1: Representations of anatomical information in radiology reports, with the event-based annotation at the top and two generative model output formats to multi-step and one-step processing. The left-hand side shows the **vanilla format** and the right-hand side shows the **building block format**.

ture our event extraction task in a unified question-answering (QA) format (Xie et al., 2022; Raffel et al., 2020). With the generative approach, the model leverages the semantic meaning of prompts for specifying subtasks and associated categorization labels. Based on experiments with in-context learning (Hu et al., 2022), we expect this to be beneficial for domain-mismatches in class label distributions, e.g. where infrequent classes in the source domain are frequent in the target domain. Furthermore, the text-to-text format offers the flexibility to incorporate additional contexts to facilitate tasks, as discussed in Section 3.3.

The input prompt comprises: (1) an input sentence from clinical notes to extract events from, (2) a question that describes the task or subtask, and (3) an ontology that provides textual labels for classification tasks and hierarchical relationships if multi-level granularities are required. The output is a word sequence that specifies the extracted information (the answer). Two alternative output formats are discussed in the next section; example input-output pairs for both are in Appendix B.

Event extraction can be seen as a multi-hop question-answering process, involving a series of subtasks for successful completion. We use a pipeline approach to address the event extraction subtasks in different steps, where each step in the pipeline consists of a specialized generative model trained for one or more of the subtask types. Three different architectures are explored:

Three-step approach: This involves a first step for detecting trigger spans and trigger types, followed by a second step for identifying the anatomy associated with each detected trigger, and a third step for normalizing each identified anatomical entity at parent and child levels individually.

Two-step approach: This involves a first step for detecting trigger spans and trigger types, followed by a second step for identifying and normalizing the anatomy associated with each detected trigger.[1]

One-step approach: we address all subtasks, which may be associated with multiple entities, in a single pass per input sentence. This method results in longer output lengths compared to the individual steps of previous two approaches.

The one-step approach substantially reduces in-

[1] Both the 2-step and 3-step approaches use the same second step, predicting anatomy spans and their normalized values. The three-step approach drops the normalized values from its second step.

ference costs compared to other two multi-step approaches. However, we find that it negatively impacts model performance due to the longer output and the compositionality gap. The performance loss is mostly recovered by changing the output format (as described next) together with a multi-task training strategy. Specifically, we train the model on both the complete task and the decomposed subtasks. This allows the model to perform subtasks independently and assemble subtask sequences for complex tasks. During inference, we decode in a single step to minimize costs.

Our work builds on generative models, specifically the clinical version of the T5 models (Lu et al., 2022), which are pre-trained on medical articles and clinical notes. This choice leverages their strengths in comprehending clinical text styles and medical knowledge.

3.2 Output formats

We explore two different output formats as illustrated in Figure 1, with subtask answers provided in sequence when there are multiple subtasks.

The baselines leverage a standard output format (referred to here as the **vanilla format**), which specifies the answer for a subtask with an extracted span followed by the entity label in brackets "[]". When multiple entities are detected, they are generated in sequence.

The vanilla format can be used with the one-step approach, but the resulting output can be very long when multiple triggers and/or entities are detected. The lack of distinction between types of spans in the output makes it harder for the language model to learn the subtask structure. To address this problem, we introduce a state-augmented prompt (referred to as the **building block** format), in which each subtask is associated with a state (as in a finite-state transducer) and explicitly named. Our approach is motivated by the work on chain-of-thought LLMs (Wei et al., 2022; Press et al., 2022), which use natural language reasoning in the generated outputs to address the compositionality gap. However, it differs in that we do not use natural language reasoning, but rather more of a programming-like description. In addition, the finite-state framework is amenable to meulti-task training, which is particularly important for the block approach.

3.3 Using target-domain contexts in prompts

A single input sentence may not provide enough information for a model to complete a task, as additional details may be needed for disambiguation or to supplement missing knowledge in pre-trained language models. Fortunately, the text format of the input allows for the seamless integration of additional contexts from the target domain during inference to aid in the task and infuse helpful domain-specific bias, even if the models were not trained for reading contexts.

The desired contexts should be relevant to the input sentence and contain helpful task information. We utilize two types of contexts: document-level and domain-level contexts to help anatomy normalization subtasks. Document-level contexts include adjacent sentences before and after the input, automatically extracted section headers[2] and exam type metadata associated with the same clinical note. The document-level contexts are likely to describe relevant anatomical parts, as section headers and exam types often summarize anatomical information. Domain-level contexts are retrieved from the unlabeled target-domain corpus. We search for the most similar sentence with the greatest lexical overlap degree, using the search algorithm BM25 (Trotman et al., 2014).[3] When the search pool is large, the top-ranked retrieved context sentence likely describes a similar anatomy part as the queried input sentence. To reduce computational costs and ensure that the retrieved sentences contain useful anatomical information, we pre-filter the target corpus to limit the search scope to sentences containing common anatomy terms listed from anatomy normalization categories and high-frequency auto-extracted section headers, reducing the number by 74%. More context-retrieval details are in Appendix D.

We add contexts only during decoding (and not in training) to prevent the model from relying too much on source-domain contexts. In the input prompts, exam type, section headers and prior sentences are placed before input sentences, following their natural orders. Other contexts are inserted between the input sentences and task ontology.[4] We test this approach in a separate anatomy nor-

[2]We extract section headers as the beginning of the last previous sentence containing ':'

[3]We implement the BM25 algorithm using `https://github.com/dorianbrown/rank_bm25`

[4]The full T5 input template is described in Table 9 from Appendix B

malization run after the one-step building block model. This process combines building block output format with target domain context integration. The reason for not directly adding it to a one-step process is that introducing contexts to inputs can potentially corrupt span detection, as the model may extract spans from the context rather than exclusively from the input sentence.

4 Experiments

4.1 Radiology datasets across exam modalities

Data split	Note Count	Sent Count
CT (train)	143	3707
MRI (train)	144	3551
PET (train)	142	5184
PET (valid)	20	758
PET (test)	40	1481
PET (unlabeled)	1471	50000

Table 1: Dataset statistics for the three radiology examination modalities: CT, MRI, and PET. We explore in-domain and cross-domain training, evaluating on PET.

We use an annotated corpus containing radiology notes about CT, MRI, and PET imaging exams; statistics are given in Table 1. The anatomy normalization labels are grouped into sublevels according to the SNOMED CT concepts. Notes in the test and validation sets are all doubly annotated. The inter-rater agreement for Trigger is 0.73 F1.

Variations in anatomy distribution across imaging modalities can cause domain discrepancies. PET has the most balanced distribution among parent-level anatomy categories, followed by CT. However, MRI has a heavily skewed distribution, with 62% of trigger-associated anatomy entities being neurological among 16 parent-level categories. See Appendix A for anatomy distribution details.

To enhance domain-specific context retrieval and boost the chances of retrieving helpful contexts, we expand the search pool by sampling 50,000 unlabeled PET report sentences from the same distribution as in the annotated reports (Lybarger et al., 2022), with a minimum of three tokens.

4.2 Implementation

In the non-generative baseline, we adopt the mSpERT model (Lybarger et al., 2023) for hierarchical multi-label entity and relation extraction. Entities are extracted as spans. We initialize with Bio-Clinical BERT (Alsentzer et al., 2019).

For the T5 model using both vanilla output formats and the subtask block formats, we initialize with ClinicalT5 (Lu et al., 2022),

For all models, the best checkpoint is chosen after 15 training epochs based on the validation performance on the target domain. For T5 models with multitask training on subtask blocks, which involves a higher number of training steps, we evaluate the model on the validation set after every 0.5 epoch approximately. For methods that do not involve multitask training, we evaluate the model on the validation set per epoch.

We implement multitask training on subtask blocks for MRI and PET, using the auxiliary tasks, as described in Section 3.1, including trigger span detection, trigger classification, joint anatomy span detection and normalization, and anatomy normalization. For the CT-PET transfer scenario, we add an additional anatomy span detection auxiliary task, as we observe that more aggressive learning is needed for anatomy span detection in the CT domain. Detailed information about hyperparameters can be found in Appendix E.

5 Results

Table 2 shows the trigger and anatomy detection results for mSpERT compared to different context-independent T5-base alternatives. For the in-domain condition, all T5 approaches outperform the mSpERT model for the three anatomy-related metrics. The results for trigger detection are mixed, but fairly similar for all. The best performance overall is obtained using the 2-step vanilla output T5 model. For the cross-domain scenarios, all models suffer degradation in performance compared to the in-domain condition, with the greatest performance drop for the normalized anatomy categories, particularly for the MRI-PET condition which has the greatest mismatch in anatomy distribution. The performance loss is greatest for the mSpERT model, with a 44% relative reduction in F1 scores for normalized anatomy (at both parent and child levels) for the MRI-PET case. In contrast, the relative loss on the parent and child levels for the T5 models is 24-29%. For both within and across-domain scenarios, the building block technique improves the 1-step results for all categories, but particularly for the more difficult anatomy normalization tasks. As described later in Section 6.2, the 1-step

Table 2: F1 scores (%) for: non-generative mSpERT (Lybarger et al., 2023), generative vanilla T5 models with both multi-step pipeline and one-step joint approaches, and our proposed one-step T5 model using the building block technique. All models adopt the T5-base architecture and are initialized with ClinicalT5 (Lu et al., 2022). **Best overall** scores are in bold, and <u>best one-step</u> scores are underlined.

Entity	mSpERT	T5-base 3-step (vanilla)	T5-base 2-step (vanilla)	T5-base 1-step (vanilla)	T5-base 1-step (blocks)
			PET → PET		
Trigger	82.4	81.9	81.9	82.1	**82.6**
Anatomy Span	65.8	**67.6**	**67.6**	66.0	<u>66.1</u>
Anatomy Parent	61.9	64.7	**64.9**	63.3	<u>63.5</u>
Anatomy Child	59.6	62.1	**62.3**	59.7	<u>60.7</u>
			MRI → PET		
Trigger	75.6	76.6	76.6	76.4	<u>**77.8**</u>
Anatomy Span	59.9	60.9	60.9	59.2	<u>**61.1**</u>
Anatomy Parent	34.7	**48.6**	47.1	44.9	<u>48.3</u>
Anatomy Child	33.5	44.6	44.0	41.2	<u>**44.8**</u>
			CT → PET		
Trigger	75.7	76.1	76.1	74.0	<u>**76.6**</u>
Anatomy Span	59.7	**61.4**	**61.4**	56.3	<u>59.8</u>
Anatomy Parent	53.2	**55.8**	54.8	50.8	<u>55.0</u>
Anatomy Child	47.5	**53.3**	51.8	48.1	<u>51.2</u>

approach is sensitive to the compositionality gap, which is ameliorated by the block approach. For the cross-domain scenarios, the best overall results are obtained with the 3-step approach for the CT-PET condition and with the 1-step block approach for the MRI-PET condition (greater mismatch). An additional advantage of the 1-step approach is the lower latency associated with using only one decoding pass.

As described earlier, target-domain contexts are added to prompts during a second step of T5 decoding to help anatomy normalization, after the 1-step subtask block decoding with T5-base. Table 3 shows results for all different types of contexts, as well as using either T5-large or T5-base in the second step without context. Without context, the T5-base and T5-large models give similar results for in-domain and CT-PET cross-domain conditions, but T5-large improves results for the MRI-PET condition. (Note that T5-large is only used in the last step; a bigger benefit could be observed if used in both steps.) All types of context are useful for the two domain-shift cases, but there is little or no benefit for the in-domain case. Of the different types of context, automatically retrieved similar sentences from unlabeled target-domain data provide the greatest benefit in the mismatched scenarios. Combining all contexts provides a small additional benefit, except for the anatomy parent in the MRI-PET case. Anecdotally, we observe that same-document contexts are useful for disambiguation, while hints for challenging examples are more likely collected from a large domain-level corpus rather than just the same document. (For examples, see Appendix F.)

Table 4 provides information on the relative cost of the different T5 models. The multi-pass models have higher latency (average passes/sample) in that passes are necessarily sequential. (Note that samples with no findings or no anatomy identified in the first pass do not require additional passes.) The number of tokens per sample is an indicator of cost. The 1-step model with blocks has a higher cost than the 2-step approach because of the additional tokens introduced by the state-augmented prompt, but the cost is still lower than the 3-step approach. The use of context adds additional cost.

Table 3: F1 scores (%) for T5 anatomy classification models with and without contexts. Results with context involve a first pass with the 1-step T5-base building blocks method, the same as "T5-base one-step (blocks)" in Table 2, followed by another pass that normalizes the anatomy spans that are previously detected by the 1-step T5-base (block) model. We normalize with the model used in the last step of the 3-step (vanilla) pipeline, optionally augmented with contexts in the prompts. We also add the T5-large normalization model without context to compare with the larger-scale counterpart.

Normalization model	**T5-large**	**T5-base**	**T5-base**	**T5-base**	**T5-base**	**T5-base**
Context	**n/a**	**n/a**	**adjacent sentences**	**metadata & header**	**BM25 retrieval**	**all combined**
	PET → PET, Trigger: 82.6, Anatomy Span: 66.1					
Anatomy Parent	63.6	**63.9**	63.8	63.7	63.8	63.7
Anatomy Child	60.9	60.9	61.0	**61.1**	60.3	60.4
	MRI → PET, Trigger: 77.8, Anatomy Span: 61.1					
Anatomy Parent	51.2	50.8	52.1	51.6	**53.8**	53.5
Anatomy Child	48.6	45.4	47.1	46.6	48.3	**48.8**
	CT → PET, Trigger: 77.8, Anatomy Span: 59.8					
Anatomy Parent	54.1	54.2	55.5	55.0	55.5	**55.9**
Anatomy Child	51.2	51.2	52.2	51.6	52.6	**53.0**

Table 4: Average number of decoding passes per sample (indicating relative decoding time) and tokens per sample (indicating relative cost) of one-step and multi-step approaches for testing on the PET domain. The token counts per sample are the average of the sum of input and output token counts, which is used for proportionality pricing LLM usage by ChatGPT. The context method uses all context combined in another normalization step as in Table 3.

Method	**passes/ sample**	**tokens/ sample**
3-step (vanilla)	2.5	355
2-step (vanilla)	1.7	199
1-step (block) + context	1.7	450
1-step (block)	1	245

Table 5: F1 scores (%) for the cross-domain MRI-PET condition using 1-step T5-base models, comparing: vanilla output format, building block format but no multitask training, and building block format with multitask training.

Entity	**vanilla**	**blocks, no multitask**	**blocks, multitask**
Trigger	76.4	76.0	**77.8**
Anatomy	59.2	57.1	**61.1**
Parent	44.9	38.6	**48.3**
Child	41.2	36.9	**44.8**

6 Analysis

In this section, we analyze results to better understand performance improvements associated with the subtask block format and retrieved context in prompts.

6.1 Multitask training for subtask blocks

To understand the contributing factors for the subtask block method's effectiveness, we examine whether the output format encodes helpful structural task information, or multitask training on individual subtasks predominantly drives performance. We conduct an additional experiment using the same subtask block output format, but without the multitask training for individual blocks. We use MRI as the source domain, because it suffers the most cross-domain performance drop. The results in Table 5 show a substantial drop in the model's performance in the absence of multi-task training, as compared to both the multi-task version and the baseline output format. This performance degradation may be attributed to increased decoding lengths.

6.2 Predictions for multiple anatomy parents

In addition to differences in the anatomy parent class distribution across domains, the three examination modalities also differ in how frequently sentences with multiple anatomy entities involve multiple parent classes. As shown in Table 6, 57% of the sentences with multiple anatomy entities in the target domain (PET) have multiple parents, whereas the percentage is much lower for the other domains (only 12% for MRI). When using the vanilla method, models trained on a domain with few instances of multiple parents will tend to predict the same parent class for each entity, as shown by the lower frequency of prediction in the table. The use of subtask blocks together with multitask training substantially improves the model's ability to identify multiple parent types when there are multiple anatomy entities. In all domains, roughly 20% of sentences have multiple anatomy entities, so this leads to overall performance improvement.

Table 6: Relative frequency (%) of sentences with multiple anatomy entities that have different parents, comparing frequencies as predicted by different models to the frequencies based on gold annotations for training data. The gold relative frequency on the PET test data is 55%.

Domain	Training	Vanilla	Blocks
PET	57	53	**56**
MRI	12	29	**46**
CT	33	45	**52**

6.3 Target domain retrieval filtering

Table 7: Normalized anatomy F1 score (%) for the MRI-PET condition, comparing approaches for using target-domain context retrieved using BM25: no context, unfiltered retrieval, and filtering the retrieval corpus to anatomy informative sentences.

Entity	no context	unfiltered contexts	filtered contexts
Parent	50.8	52.7	**53.8**
Child	45.4	47.4	**48.3**
Trigger: 77.8, Anatomy: 61.1			

To reduce the search costs, we filter the unlabeled target domain data to include only sentences with anatomy terms before running retrieval with BM25. To understand the impacts on performance, we run experiments on unfiltered data, again focusing on the MRI data where domain differences are greatest. Table 7 shows that filtering for anatomy not only reduces costs but also gives a small improvement in results for identifying normalized categories.

7 Related work

7.1 Event extraction methods

Event extraction research has predominantly depended on BERT-based (Devlin et al., 2019; Alsentzer et al., 2019) models, where the extraction subtasks are performed by classifiers utilizing the language model layer representations (Eberts and Ulges, 2019; Zhong and Chen, 2021; Lybarger et al., 2023). They often yield satisfactory results when training on sufficient in-domain training data. For example, when training and testing on CT scan reports, normalizing anatomical terms can result in an F1 score of 79% for nine major body parts and 73% for 41 sub-body parts (Lybarger et al., 2021). Recently, there has been growing interest in adopting generative approaches (Raffel et al., 2020; Brown et al., 2020) for information extraction, which incorporates task descriptions and auxiliary context information to enhance performance (Xie et al., 2022). Many efforts (Lu et al., 2022; Phan et al., 2021; Lehman et al., 2023; Luo et al., 2022) support exploration of clinical tasks through pre-training generative models for biomedical and clinical domains. In this study, we explicitly evaluate generative models in domain shift settings, with an emphasis on minimizing inference costs.

7.2 Context augmentation

Integrating models with supplementary contexts has shown benefits in knowledge-intensive tasks (Lewis et al., 2020; Guu et al., 2020). Generative models can utilize knowledge prompts from external knowledge sources (Peng et al., 2023; Liu et al., 2021). In our work, we retrieve contexts from the unlabeled clinical note corpus without relying on external resources.

7.3 Compositionality Gap

The compositionality gap has been identified as a challenge in generative models when multiple subtasks are combined (Press et al., 2022). Prior research on large language models has demonstrated that breaking down complex tasks into smaller subproblems can be beneficial (Wei et al., 2022; Press et al., 2022). Small models have been employed for multiple decoding passes (Khot et al., 2021), but there is limited research on reasoning with smaller models that merge these steps, which is essential for real-time applications in the clinical field.

8 Conclusion

In conclusion, we present generative event extraction methods for radiology findings that improve generalization under domain shifts and reduce the inference costs. By decomposing complex tasks into simpler subtask blocks and incorporating target-domain context during the inference process, our approach enables smaller models to achieve performance similar to or better than those obtained with more decoding passes, and comparable to larger models on anatomy normalization. Our methods make efficient inference for extensive clinical notes more feasible. This work offers insights into reasoning with smaller models and using context to compensate the reduced model size.

Limitations

The use of machine learning models in clinical decision-making requires an understanding of the reasoning behind model predictions. Our study focuses on improving the performance of smaller models using context and subtask blocks. While the subtask state labels provide some interpretability, we have not explored its impact on trust among medical professionals. In addition, the relative benefit of the different multi-pass strategies and different types of context appear to depend on the degree of domain mismatch, which should be further explored in future work.

Ethics Statement

Radiology reports contain sensitive patient information and it is crucial to handle this data responsibly, adhering to strict privacy and confidentiality guidelines. The dataset used in this paper was fully-de-identified. We received approval from our institution's IRB prior to conduct the presented research and used HIPAA compliant servers. Additionally, a careful examination is needed to assess potential bias in models used for extracting information from radiology reports prior to implementing real life secondary use applications.

Acknowledgements

This work was supported by NIH/NCI (1R01CA248422-01A1 and 1R21CA258242-01).

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Conference on Neural Information Processing Systems*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Markus Eberts and Adrian Ulges. 2019. Span-based joint entity and relation extraction with transformer pre-training. In *European Conference on Artificial Intelligence*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Yushi Hu, Chia-Hsuan Lee, Tianbao Xie, Tao Yu, Noah Smith, and Mari Ostendorf. 2022. In-context learning for few-shot dialogue state tracking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2627—-2643. Association for Computational Linguistics.

Tushar Khot, Daniel Khashabi, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2021. Text modular networks: Learning to decompose tasks in the language of existing models. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 1264–1279.

Wilson Lau, Laura Aaltonen, Martin L. Gunn, and Meliha Yetisgen-Yildiz. 2020. Automatic assignment of radiology examination protocols using pre-trained language models with knowledge distillation. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2021:668–676.

Eric P. Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J. Smith, Zachary M. Ziegler, Daniel Nadler, Peter Szolovits, Alistair E. W. Johnson, and Emily Alsentzer. 2023. Do we still need clinical language models? *ArXiv*, abs/2302.08091.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2021. Generated knowledge prompting for commonsense reasoning. In *Annual Meeting of the Association for Computational Linguistics*.

Qiuhao Lu, Dejing Dou, and Thien Nguyen. 2022. ClinicalT5: A generative language model for clinical text. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5436–5443, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*.

Kevin Lybarger, Aashka Damani, Martin L. Gunn, Özlem Uzuner, and Meliha Yetisgen-Yildiz. 2021. Extracting radiological findings with normalized anatomical information using a span-based BERT relation extraction model. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2022:339–348.

Kevin Lybarger, Nicholas J Dobbins, Ritche Long, Angad Singh, Patrick Wedgeworth, Özlem Uzuner, and Meliha Yetisgen. 2023. Leveraging natural language processing to augment structured social determinants of health data in the electronic health record. *Journal of the American Medical Informatics Association*. Ocad073.

Kevin Lybarger, Namu Park, Sitong Zhou, Aashka Damani, Alison Brennan, Jagjeet Gill, Nianiella Dorvall, Vy Huynh, Spencer Lewis, Martin L. Gunn, Özlem Uzuner, and Meliha Yetisgen-Yildiz. 2022. A corpus of radiology reports from multiple imaging modalities with fine-grained event-based annotations. In *American Medical Informatics Association Annual Symposium*.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Lidén, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *ArXiv*, abs/2302.12813.

Long Phan, James T. Anibal, Hieu Trung Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. SciFive: a text-to-text transformer model for biomedical literature. *ArXiv*, abs/2106.03598.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *ArXiv*, abs/2210.03350.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Andrew Trotman, Antti Puurula, and Blake Burgess. 2014. Improvements to BM25 and language models examined. *Proceedings of the 19th Australasian Document Computing Symposium*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I. Wang, Victor Zhong, Bailin Wang, Chengzu Li, Connor Boyle, Ansong Ni, Ziyu Yao, Dragomir R. Radev, Caiming Xiong, Lingpeng Kong, Rui Zhang, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2022. UnifiedSKG: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. In *Conference on Empirical Methods in Natural Language Processing*.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online. Association for Computational Linguistics.

A Hierarchical anatomy normalization categories

We normalize detected anatomy spans for applications focusing on anatomy categories rather than specific anatomy terms. We classify at different granularities, a parent-level coarse classification with 16 parent labels and a child-level fine-grained

Parent-level Class	Child-level Classes
Neurological	Undetermined, Spine Cervical, Spine Thoracic, Spine Lumbar, Spine Sacral, Spine Cord, Spine Unspecified, Brain, Nerve, Pituitary, Cerebrospinal Fluid Pathway, Cerebrovascular System, Extraaxial
Cardiovascular	Undetermined, Venous, Arterial, Pulmonary Artery, Heart, Pericardial Sac, Coronary Artery
Thoracic	Undetermined, Mediastinal
Respiratory	Undetermined, Lung, Pleural Membrane, Tracheobronchial
Digestive	Undetermined, Esophagus, Stomach, Intestine, Small Intestine, Large Intestine
Hepato-Biliary	Undetermined, Gallblader, Bile Duct, Pancreas, Liver
Urinary	Undetermined, Kidney, Urinary Bladder, Ureter
Lymphatic	Undetermined
F Reproductive Obstetric	Undetermined, Breast, Ovary, Uterus, Adnexal, Extra-embryonic, Placenta, Fetus, Umbilical Cord, Female Genital Structure
M Reproductive	Undetermined, Prostate, Testis, Epididymis
Musculo-Skeletal	Undetermined, Skeletal and or Smooth Muscle, Bone and or Joint
Body Regions	Undetermined, Entire Body, Pelvis, Lower Limb, Upper Limb
Head Neck	Undetermined, Thyroid, Neck, Ear, Eye, Mouth, Nasal Sinus, Pharynx, Laryngeal
Skin	Undetermined, Skin and or Mucous Membrane, Subcutaneous
Abdomen	Undetermined, Retroperitoneal, Abdominal Wall, Peritoneal Sac, Spleen, Adrenal Gland, Mesentery
Miscellaneous	Undetermined, Adipose Tissue, Connective Tissue, Biomedical Device

Table 8: Hierarchical anatomy normalization categories at parent and child levels.

classification with 72 categories. Each parent-level class includes an "Undetermined" child-level class to account for cases that don't fit into its other specified child classes. The full normalization categories are in Tabel 8.

As shown in Figure 2, MRI exhibits a more imbalanced distribution, with a majority of the anatomies related to the "Neurological" parent-level class. In CT exams, "Respiratory" account for 16% and "Neurological" represent 19% among all finding-related anatomies. For MRI, "Musculo-Skeletal" constitutes 18% while "Neurological" exams make up a substantial 62%. Lastly, in PET, "Head Neck" accounts for 12% and "Musculo-Skeletal" comprises 14%.

B Generative method input and output formats

We document the templates for the input and output, with examples in Tabel 9. For the template with contexts, "prepended contexts" include prior sentences, section headers, exam type metadata, other contexts are "appended contexts".

<table>
<tr><th>TEMPLATE</th></tr>
<tr><td>Template: [Input sentence] [Question] structured knowledge: [Task ontology]
Template with contexts: [Prepended contexts] [Input sentence] [Question] structured knowledge: [Appended contexts] [Task ontology]</td></tr>
<tr><td>Trigger task ontology (for T5-vanilla pipeline: 1st step) Indication | Lesion | Medical_Problem
Anatomy task ontology (for T5-vanilla pipeline: 2nd step, 3rd step) Neurological: Undetermined, Spine Cervical, Spine Thoracic .... (see Table A)
Trigger anatomy task ontology (for all related to one-step building block) trigger types: Indication | Lesion | Medical_Problem anatomy categories: Neurological: Undetermined, Spine, ...</td></tr>
<tr><th>EXAMPLE</th></tr>
<tr><td>Input sentence: 18 x 17 mm hypermetabolic soft tissue density insinuating between the left lobe of the liver and anterior abdominal wall (the R/112) with maximum SUV 14.4 .</td></tr>
<tr><td>Model: T5-vanilla pipeline: first step (trigger span & type)
Question: Question: What are medical findings in this sentence?
Output: trigger: density [Lesion]</td></tr>
<tr><td>Model: second step (anatomy span & type)
Question: Consider the medical finding "density" in the span "hypermetabolic soft tissue density insinuating between the", Question: What anatomy it occurs in? Where is it located?
Output: anatomies: soft tissue [Hepato-Biliary | Liver], left lobe of the liver [Hepato-Biliary | Liver], anterior abdominal wall [Abdomen | Abdominal Wall]</td></tr>
<tr><td>Model: third step (anatomy normalization)
Question: Consider the anatomy "soft tissue" in the span "17 mm hypermetabolic soft tissue density insinuating between", which anatomy category it belongs to among listed options?
Output: anatomies: soft tissue [Hepato-Biliary | Liver]</td></tr>
<tr><td>Model: T5-vanilla one-step (trigger span & type, anatomy span & normalization)
Question: Question: What are medical findings in this sentence? What anatomy they occur in? which anatomy category they belong to among listed options?
Output: trigger: density [Lesion] anatomies: soft tissue [Hepato-Biliary | Liver], left lobe of the liver [Hepato-Biliary | Liver], anterior abdominal wall [Abdomen | Abdominal Wall]</td></tr>
<tr><td>Model: T5 one-step subtask blocks (trigger span & type, anatomy span & normalization)
Question: [same as T5-vanilla one-step]
Output: state: trigger detection answer: density state: trigger classification answer: density [Lesion] state: span detection answer: soft tissue, left lobe of the liver, anterior abdominal wall state: classification answer: soft tissue [Hepato-Biliary | Liver] state: classification answer: left lobe of the liver [Hepato-Biliary | Liver] state: classification answer: anterior abdominal wall [Abdomen | Abdominal Wall]</td></tr>
<tr><td>Model: multitask for trigger classification (trigger type)
Question: Consider the medical finding "density", Question: What is the type of this medical finding?
Output: state: trigger classification answer: density [Lesion]</td></tr>
<tr><td>Model: multitask for anatomy span (anatomy span)
Question: Consider the medical finding "density" in the span "hypermetabolic soft tissue density insinuating between the", Question: Please identify terms that describe the finding's anatomy locations.
Output: state: span detection answer: soft tissue, left lobe of the liver, anterior abdominal wall</td></tr>
</table>

Table 9: Templates and examples for T5 inputs and outputs. The "multitask" rows correspond to auxiliary tasks for the T5 one-step subtask block method. We omit rows for "multitask for anatomy" and "multitask for anatomy normalization", since they use the same question format as the 2nd and 3rd steps of the pipeline approach, but with answers in the subtask block format.

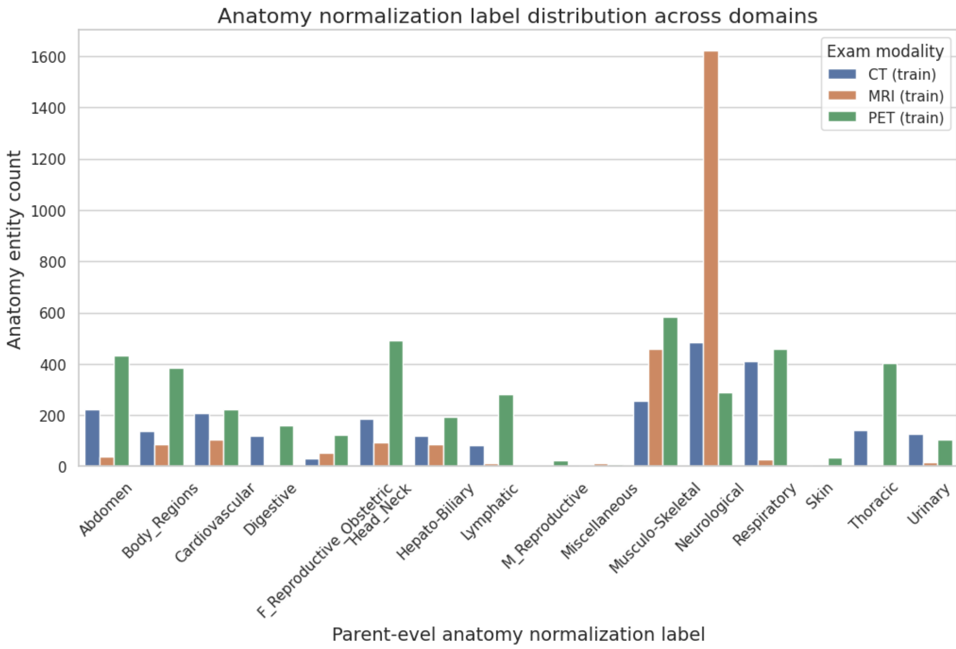


Figure 2: Domain differences in terms of the frequencies of parent-level anatomy normalization labels from the training data.

C Post-processing for the generative event extraction

When matching spans in the input sentence for predicted terms, for single-token terms, we match the corresponding token. For multiple-token phrases, we match phrases using the longest common normalized string to the input sentence. Where multiple matches are found, we choose the first match from the left of the sentence, while for anatomy spans, we choose the closest match to their query triggers.

D Domain-level context retrieval

We conduct a domain-level context search using 50,000 sentences from the target domain (PET) corpus with more than three tokens, plus 1841 sentences from the test set. The retrieved content must not be the input sentence itself. For each input clinical sentence, we identify the most lexically similar sentence from the search pool by selecting the one with the highest BM25 score. We remove punctuation and lowercase each input query when matching it with the search corpus sentences using the BM25 method.

To filter for anatomy-informative sentences, we employ the same BM25 model to match the entire search corpus with a single anatomy string, which was cheaply curated from the anatomy normalization categories and frequently auto-extracted section headers, as shown in Table 10. After filtering, the search corpus is reduced to 36%, shrinking from 51,481 sentences to 18,959 sentences.

E Implementation details

The mSpERT models are trained at a batch size of 15 for 15 epochs.[5] T5 models utilize a maximum input length of 768 tokens and a maximum output length of 512 tokens. When incorporating all types of contexts, we double the input maximum length to 1536 tokens. We train 15 epochs, with a batch size of 8. For the T5 large model, to accommodate a single NVIDIA A100 device, we employ gradient accumulation by using a batch size of 2 and accumulating four times.

F Case study for context benefits

We observe that contexts can aid in disambiguation (e.g. right middle lob) and understanding difficult medical terminology (e.g. biapical). For both examples presented in Table 11, contexts include the term "pulmonary", indicating the anatomies are related to lungs.

[5] We use full event schema for mSpERT models, including all attribute types in the annotations, including anatomy, characteristic, size, size-trend, and count. While T5 models only extract the most important attribute, the anatomy attribute.

Neurological: Spine Cervical, Spine Thoracic, Spine Lumbar, Spine Sacral, Spine Cord, Spine, Brain, Nerve, Pituitary, Cerebrospinal, Cerebrovascular, Extraaxial
Cardiovascular: Venous, Arterial, Pulmonary Artery, Heart, Pericardial Sac, Coronary Artery
Thoracic: Mediastinal
Respiratory: Lung, Pleural Membrane, Tracheobronchial
Digestive: Esophagus, Stomach, Intestine, Intestine, Intestine
Hepato-Biliary: Gallbladder, Bile, Pancreas, Liver
Urinary: Kidney, Urinary Bladder, Ureter
Reproductive: Breast, Ovary, Uterus, Adnexal, Extra-embryonic, Placenta, Fetus, Umbilical Cord, Genital Structure, Prostate, Testis, Epididymis
Musculo-Skeletal: Skeletal, Smooth Muscle, Bone, Pelvis, Limb
Head Neck: Thyroid, Neck, Ear, Eye, Mouth, Nasal Sinus, Pharynx, Laryngeal
Skin: Skin, Mucous Membrane, Subcutaneous
Abdomen: Retroperitoneal, Abdominal, Peritoneal Sac, Spleen, Adrenal, Mesentery, Adipose, Chest, Mediastinum, Osseous, Bones, Extremities, Lungs, Musculoskeletal, Ventricular, Bowel, Pleura, Spleen, Vasculature, Thorax, Gallbladder, Kidneys, Adrenals, Adrenal, Cardio

Table 10: Common anatomy terms for filtering the search scope of domain-level context retrieval. This list is curated from the anatomy task ontology (Table 8) and frequent section headers. Stop words are removed.

Table 11: Error examples with helpful contexts

Error with example	Contexts	Before and after
[ambiguity] **Right middle lobe** nodule (4, 81) measures 3 mm, previously 4 mm	[document-level section header] Scattered bilateral **pulmonary** nodules, as described below	before: Hepato-Biliary | Liver after: Respiratory | Lung
[hard vocabulary] There is **biapical** fibrosis	[domain-level BM25] There is biapical **pulmonary** fibrosis compatible with radiation therapy	before: Musculo-Skeletal | Bone and or Joint after: Respiratory Lung

Intersectionality and Testimonial Injustice in Medical Records

Kenya Andrews, Bhuvani Shah, Lu Cheng
kandre32@uic.edu | bshah46@uic.edu | lucheng@uic.edu
University of Illinois at Chicago
Computer Science

Abstract

Detecting testimonial injustice is an essential element of addressing inequities and promoting inclusive healthcare practices, many of which are life-critical. However, using a single demographic factor to detect testimonial injustice does not fully encompass the nuanced identities that contribute to a patient's experience. Further, some injustices may only be evident when examining the nuances that arise through the lens of intersectionality. Ignoring such injustices can result in poor quality of care or life-endangering events. Thus, considering intersectionality could result in more accurate classifications and just decisions. To illustrate this, we use real-world medical data to determine whether medical records exhibit words that could lead to testimonial injustice, employ fairness metrics (e.g. demographic parity, differential intersectional fairness, and subgroup fairness) to assess the severity to which subgroups are experiencing testimonial injustice, and analyze how the intersectionality of demographic features (e.g. gender and race) make a difference in uncovering testimonial injustice. From our analysis, we found that with intersectionality we can better see disparities in how subgroups are treated and there are differences in how someone is treated based on the intersection of their demographic attributes. This has not been previously studied in clinical records, nor has it been proven through empirical study.

1 Introduction

In medical settings, decisions can have life-critical consequences (Zenios et al., 1999; Kumar Mangla et al., 2023; White and Lo, 2020; Cheng et al., 2021; Cheng and Liu, 2023), making it *essential* to ensure that machine learning tools use there are fair. This fairness is often measured with common fairness metrics such as demographic parity (Dwork et al., 2012) and equal opportunity (Hardt et al., 2016). However, these tools do not consider the intersectionality of the subjects under consideration (Ghosh et al., 2021; Gohar and Cheng, 2023). That is, by focusing solely on factors such as race, gender, or socioeconomic status, we ignore the nuances related to individuals with unique experiences shaped by having multiple features sensitive to marginalization. We theorize that *how various aspects of an individual intersect and contribute to their experiences, via intersectionality, could make instances of injustice more overt - and in some cases may be the sole approach for identifying such instances*. Intersectionality recognizes that power relations based on factors such as race, class, and gender are not mutually exclusive and can interact with each other, affecting all aspects of the social world (Marques, 2018). Therefore, it is important to consider intersectionality when evaluating the fairness of machine learning tools in medical settings.

In clinical settings, it is particularly important that care providers (e.g. physicians) properly acknowledge what their patients are hoping to convey to them in a way that does not diminish what the patient is saying. Moreover, it is imperative for care providers to accurately relay their understanding of their patients' experiences, as others will be dependent upon their previous understandings and evaluations, often recorded in notes, to assist with overseeing and providing care for that patient (Jin, 2021). We have seen that when this does not occur, there are higher instances of death amongst certain marginalized groups (Bowman, 2013). With the rise in using machine learning tools to help make decisions on medical plans and treatments, who often only interact with the notes provided to them and not the actual patient, it is vital they are able to properly see patients. This visibility should be clear despite previous attempts at burying their words behind instances of injustices which hides them as a speaker. Here, we focus on a particular form of injustice - testimonial injustice. Testimonial injustice occurs when someone is assigned less

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 358–372
July 14, 2023 ©2023 Association for Computational Linguistics

credibility due to prejudices about them (Fricker, 2019).

The aim of our study is to examine how testimonial injustice in medical records is affected by the intersectionality of gender and race. These two observable attributes have historically led to marginalization in various societal settings, such as education (Rankin and Thomas, 2020), housing (Roscigno et al., 2009), and healthcare (Krieger, 1990; Chapman et al., 2013). In fact, some forms of marginalization may only be evident in those with multiple marginalized identities - for instance, a Black police woman may not experience the same level of power and privilege as a White male police officer (Martin, 1994). Neglecting to consider the various contributing identities of an individual may further marginalize them. Therefore, it is important to consider intersectionality when identifying and addressing injustices in order to result in more accurate classifications and decisions.

There has been a small amount of work done to understand testimonial injustice in medical records and to our knowledge no prior work on how intersectionality might affect the emergence of testimonial injustice, even in life-critical medical settings. This motivates our contributions to this work: (1) The importance of intersectionality has been spoken about but has not been shown before (particularly in the medical setting). Thus, we perform an empirical study to show there is a difference in how subgroups are treated in medical settings, but this can only be revealed in intersectional views. (2) Practitioners continue to use singular-feature fairness metrics in medical settings. Thus, we provide proof that we should not be using these metrics to detect instances of injustice. This proof has not been provided before, not even in medical settings. Thus, we (3) perform an empirical study to show traditional fairness metrics (i.e. demographic parity) are inefficient when judging people's experiences in healthcare because they produce different results when the entirety of a person is considered. (4) Lastly, not all metrics fit each situation - even in similar settings. Therefore, we analyze if different intersectional fairness metrics might reveal differences in how we recognize intersectionality.

Previous studies have shown that both Black patients and female patients are more likely to experience testimonial injustice in the medical field, as evidenced by the use of biased language in their records (Beach et al., 2021). However, these studies have not examined the specific impact of intersectionality, or how being simultaneously Black and female might affect testimonial injustice. Our work seeks to address this gap by examining the impact of the intersection of ethnicity (Black, Asian, Latino, and White) and gender (Male and Female - though we acknowledge in modern society, there is recognition of genders beyond the traditional binary options, the dataset used here only includes these two genders) on testimonial injustice in medical records.

2 Related Works

Despite the increased use of machine learning tools and a growing focus on intersectionality in the medical community (Holman et al., 2021; Bauer and Lizotte, 2021), there have been limited efforts to understand how intersectionality can impact outcomes in medical settings. Since various healthcare professionals rely on medical records to make treatment decisions and give proper care, it is crucial that such records are written appropriately (Bali et al., 2011). The authors of (Adam et al., 2022) found that even when race is removed from patients' records, models could detect the race of the patient - even when humans could not. Furthermore, they discovered that models trained on these records (i.e. which race has been removed from) still maintain biases in treatment recommendations. Though they only remove race in their work, this further affirms that there are differences in how patients are spoken about in their records based on demographic features, emphasizing the need to study what can occur if we look at multiple demographic features as we do here. In their work, P Goddu et al. explored how stigmatizing language in a patient's medical record can shape the attitudes of physicians-in-training towards the patient and their clinical decision-making. They found that stigmatizing language is associated with more negative attitudes and less aggressive pain management. Building on this work, we examine words that may indicate testimonial injustice, which occurs when someone's statements are diminished due to stereotypes or prejudices about them (Fricker, 2019). It is therefore important to identify instances of stigmatizing language in medical records and take steps to prevent them from occurring as emphasized by Park et al..

In (Beach et al., 2021), the authors use a lexicon look-up to identify testimonial injustice in medical records, analyzing the use of quotation marks, evidential words, and judgmental words in the records of male and female patients who are Black or White. We expand their work, including words that are negative and commonly used stigmatizing words in medical settings. We exclude the search for quotation marks, acknowledging that direct quotations may give rise to uncertainty by suggesting that the statement in question constitutes not a fact, but rather an assertion (Beach et al., 2021). However, we believe that our expanded lexicon will help to identify instances of testimonial injustice. Further in contrast to Beach et al., we consider the records of Black, White, Asian, and Latino patients, exploring how testimonial injustice may differ across the intersection of their identities with gender. The authors found that Black and female patients are most likely to experience testimonial injustice, highlighting the need to examine how different intersectional identities impact experiences of testimonial injustice in medical settings.

Previous research has examined the presence of epistemological bias in medical records based on sensitive attributes to detect instances of experiences injustice i.e. disparate treatment. Himmelstein et al. studied diabetic patients and found that non-Hispanic Black patients were more likely to have stigmatizing language included in their notes than non-Hispanic White patients. Similarly, Sun et al. investigated medical records and racial bias, discovering that Black patients had a 2.54 times higher chance of negative descriptors than White patients. These studies suggest that certain demographics may experience differential treatment in medical settings, which may help explain healthcare disparities. However, these works only examined single demographic features, while we seek to investigate their intersection. We anticipate that studying the intersection of groups will more clearly reveal instances of injustice or discrepancies in treatment. The ongoing use of tools that do not consider intersectionality highlights the importance of this research (Buolamwini and Gebru, 2018).

Guo and Caliskan developed a technique to automatically identify intersectional biases from static word embeddings. They found that their model's highest accuracy was for predicting emergent intersectional bias among African American and Mexican American women. This could be attributed to these groups experiencing more overt biases that are easier to detect. This discovery motivates us to further investigate if biases are more prevalent in high-risk settings such as medical settings, especially for individuals from marginalized groups. However, it can be challenging for humans to identify when a bias is occurring since it can be subtle, as highlighted by Hube and Fetahu. Furthermore, doctors may struggle to recognize their own use of words that cause testimonial injustice since they may be unconsciously influenced by their own biases and take them as facts (FitzGerald and Hurst, 2017; Beeghly and Madva, 2020).

3 Data

3.1 MIMIC-III

Obtaining medical data has been a standing challenge, largely due to HIPAA requirements and privacy constraints. We use the MIMIC-III (Johnson et al., 2016) dataset, which contains features of interest to our experiments: ethnicity/race, gender, patient id, diagnosis, physicians' notes , and so on. This data was collected between 2001-2012 at the Beth Israel Deaconess Medical Center in Boston, MA. The MIMIC-III dataset contains information for 46,146 patients. The distribution of racial groups in the data was highly disproportionate, as shown in Table 1. The two genders represented in this dataset, Female and Male, however, are more balanced. We removed ethnicities that were listed as "unknown/not specified", "multi-race ethnicity", "other", "unable to obtain", and "patient declined to answer" since we cannot clearly denote the race of these patients. We also removed patients whose diagnosis was "newborn" since these patients had notes solely stating they were newly born. We did however include the newborns who had other diagnoses. Only 9 of those patients were Caribbean and 38 were Middle Eastern, thus we removed them from the records as well. We were not able to find any duplicate records in the dataset, with a simple python search.

After data pre-processing, there are 32,864 patients in total for experimentation. We truncated the MIMIC-III feature 'ethnicity' into 'race' such that all ethnicities are represented as the race often associated with them as labeled in the dataset (e.g. original ethnicity in the dataset: 'ASIAN - VIETNAMESE' was truncated to 'Asian'). For ethnicities that were not associated with a particular race, we searched for how they are commonly associated and relabeled them to the race (e.g. original ethnicity in the dataset: 'SOUTH AMERICAN' was relabeled to 'Latino'). Finally, given that many patients had multiple records, we clustered the patients based on their patient_id and combined their records based on patient_id, gender, race, and diagnosis (e.g. 56327, male, Latino, HYPOTENSION). We then run analysis on the physicians' notes to find terms that are testimonially injust.

Race	Gender	Count
White	Female	15,399
Black	Female	2,522
Asian	Female	512
Latina	Female	662
White	Male	20,317
Black	Male	2,041
Asian	Male	690
Latino	Male	1,041

Table 1: Counts of patients by race and gender.

We analyze the distribution of data for MIMIC-III in A.3. Our analysis looked at the occurrence of our four types of words associated with testimonial injustice, namely evidential (Figure 6 and 7), judgmental words (Figure 8 and 9), stigmatizing words (Figure 10 and 11), and negative words. We plot the density distribution of each gender, race, and their intersection as normalized sums of these types of words, where the numerator is the frequency of occurrence of the relevant words for that patient and the denominator is the number of records for that patient. We did not include the plots for negative words due to their limited occurrence in the medical notes of this dataset, however we do use them in our analysis of the results for detecting testimonial injustice. Our observations suggests that the confluence of race and gender better helps us in distinguishing instances of testimonial injustice than either race or gender in isolation. In particular, when race and gender are considered independently, males seem to be treated better than females or White patients are treated generally better than Black patients. However, there is nuance in the difference in the treatment of White males and White females as well as Black males and Black females.

3.2 Testimonial Injustice Terms

In order to assess testimonial injustice in the physicians' notes, we focus on 4 main categories of unjust words: evidential, judgemental, negative, and stigmatizing words that can contribute to someone experiencing testimonial injustice. We use the same evidential and judgmental words from (Beach et al., 2021). Evidential terms do not endorse a statement but allow it to be agnostic (e.g. “complains", “says", “tells me" and so on). When a physician uses these words, they express dismissing what the patient is actually experiencing. Judgment terms cast doubt on the sayer by the hearer (i.e. the physician) by trying to make their statements sound good or bad (e.g. “apparently", “claims", “insists", and so on). Exacerbated racial and ethnic healthcare disparities have been linked to negative words used to describe Black patients as well (Sun et al., 2022). Negative words are included in this study as they typically show active rejection or disagreement, e.g. “challenging", “combative", “defensive", “exaggerate", and so on. Clearly, the use of these words expresses assumptions about the patient and could result in a lower quality of care.

We also include stigmatizing terms as they are commonly used in medical contexts (Himmelstein et al., 2022). Stigmatizing terms are rooted in stereotypes or stigmas about a person (Link and Phelan, 2001) (e.g. “user", “faking", “cheat", and so on). Using stigmatizing terms may alter treatment plans, transmit biases between clinicians, and alienate patients. This lexicon has been proven to consist of words used to diminish specific conditions like diabetes, substance use disorder, and chronic pain (Himmelstein et al., 2022). All of these conditions are known to disproportionately affect racial minority groups. Using all of these terms in our lexicon lookup 4.2 will help us to detect testimonial injustice in these medical records.

4 Methods

Although all marginalized groups invariably experience some degree of injustice, our aim is to bridge the gap in research by highlighting the disparate treatment of subgroups in medical notes. To achieve this goal, we estimate and compare common metrics across different groups (i.e. Asian men, Asian women, Black men, Black women, Latino men, Latina women, White women, and White men) specifically using demographic parity, differential intersectional fairness, and subgroup fairness.

4.1 Normalization

To account for patients who had multiple visits or were admitted to the ICU for multiple days, the physicians' notes were combined for each patient's duration in the ICU. To analyze the potential variance in testimonial injustice among different groups, we summed the frequency of testimonial injustice words in the notes for each patient and then normalized this frequency by dividing it by the number of original records we had for that particular patient. This allowed us to ensure that each patient had an equal standing, regardless of length of hospital stay or number of visits from doctors. By using normalized sums, we were able to compare groups and determine if there were any differences in levels of testimonial injustice. The normalized sums of occurrences of testimonial injustice across each intersection of groups are visualized in Figure 5 in A.1.

4.2 Lexicon Lookup

After normalizing the sums of testimonial injustice for each patient, we performed a lexicon lookup for exact phrase matching. With this, we counted the frequency of occurrence for each testimonial injustice word in the patients' combined and normalized visits. We combined the terms introduced in Section 3.2 commonly associated with being evidentially biased, judgmental, negative, and stigmatizing into a lexicon.

4.3 Defining Fairness

In this work, we define the desired fairness as the following: *a patient's record has **no** terms which are considered testimonial unjust.* However, this is a strict boundary that is unlikely to be met since a term could appear in a patient's record but might not actually be casting doubt on them as a sayer (i.e. testimonial injustice). Thus, we find the greatest number of occurrences of each type of term that indicates testimonial injustice, $m = max_p(t/r)$ (where p are the patients). We determine that if a patient has more than $m * .10$ in that particular type of term, they as experiencing testimonial injustice. For this work, we arbitrarily use 10% of the maximum value for each term. In the future, we will do some experimentation to improve this definition of fairness. To determine if there is disparate treatment amongst groups to this fairness definition, we use fairness metrics - demographic parity, differential intersectional fairness, and subgroup fairness.

4.3.1 Demographic Parity

Demographic parity requires that the difference in two groups being assessed have equal chances of receiving a positive outcome (Dwork et al., 2012). We use this metric as our baseline metric to understand how testimonial injustice might reveal itself if we ignore intersectionality, as has been done with most works in the fairness literature [(Hardt et al., 2016), (Kusner et al., 2017), (Agarwal et al., 2018),and so on]. That is, we are seeking to investigate whether there is a significant difference in the way a patient is spoken about in medical records when the intersection of their race and gender are considered. Demographic parity is a popular fairness metric, but it does not work to reveal fairness or justice; rather it solely reveals equity. We can look at the example of when both groups have high amounts of injustice (i.e. true fairness occurs when neither group experiences injustice, nearly 0) hence, fairness is not detected only equality or when a marginalized group should be afforded more opportunity for the sake of corrective justice due to historical bias hence justice is not enforced. In these cases, demographic parity is still satisfied, but fairness nor justice persists. Demographic parity is defined as:

$$\frac{P(Y = 1|A = a)}{P(Y = 1|A = a')} > 0.8, \tag{1}$$

where Y is the outcome and A is the sensitive attribute. Demographic parity looks to ensure the difference between the two groups receiving a positive outcome is greater than 80%.

4.3.2 Differential Fairness

For intersectionality, we first look at ϵ-Differential fairness (Foulds et al., 2020), which requires that the difference between groups, regardless of their combination of sensitive attributes, not be treated differently within a range. This metric of fairness allows us to include multiple attributes of a person whereas demographic parity only allows us to look at one sensitive attribute per group. Differential fairness is defined as:

$$e^{-\epsilon} < \frac{P(M(x) = y|s_i, \theta)}{P(M(x) = y|s_j, \theta)} < e^{\epsilon}, \tag{2}$$

where ϵ should be small. In our experiments, it is set to 0.01, M is a mechanism (linear regression in our case) that takes an instance, x, from the data to achieve some outcome, y, s values are the cross product of sensitive attributes, and θ is the distribution of x.

4.3.3 Subgroup Fairness

Another common intersectional fairness notion is Statistical Parity Subgroup Fairness or subgroup fairness. We use subgroup fairness to compare our results with the differential fairness metric. Subgroup fairness (Kearns et al., 2018) requires there be no difference in positive outcomes between groups, but we are allowed to ignore an α amount of people. Subgroup fairness is described for each group, a, by:

$$\alpha(a, \mathcal{P}) * \beta(a, M, \mathcal{P}) \leq \gamma, \tag{3}$$

where,

$$\begin{aligned} \alpha(a, \mathcal{P}) &= P_{\mathcal{P}}[a(x) = 1] \\ \beta(a, M, \mathcal{P}) &= |P_{\mathcal{D},\mathcal{P}}[M(x) = 1] - \\ &\quad P_{M,\mathcal{P}}[M(x) = 1 | a(x) = 1]|. \end{aligned}$$

Here M is a classifier, $\mathcal{P}$ is the distribution of patients, $\gamma\epsilon[0, 1]$ indicates the amount of deviation from equity we tolerate. We relax this constraint for our experiments, allowing γ to be 95% of the maximum value of $\alpha(a, \mathcal{P}) * \beta(a, M, \mathcal{P})$ for each term that leads to testimonial injustice. $a(x) = 1$ indicates that individuals with sensitive feature, x, are in group a.

5 Results

When examining the results for demographic parity, we solely focus on instances of race or gender, as this approach only allows for an assessment of one factor at a time. However, for differential fairness and subgroup fairness, we conduct an intersectional analysis with race and gender. For these, we look to see which groups have privilege over another, meaning one group experiences less testimonial injustice in their physicians' notes as opposed to the group they are being compared to.

5.1 Demographic Parity

Gender. In terms of Demographic Parity gender analysis, there was little to no disparate treatment detected across all term types between male and female patients, indicating that there was minimal evidence of injustice in the data based on gender, as observed in Figure 1. The greatest difference was found within evidential words, where female patients experienced the most injustice. Then follows the stigmatizing words and judgment words with the greatest bias against females. The least difference comes from the negative words with males experiencing the least fairness. Negative words occurred the least and stigmatizing words occurred the most across the patient records. With this, gender should not be found to be a significant predictor of the treatment or care received by patients. Therefore, the findings of the analysis should show that a person's gender membership does not have any substantial impact on how they are treated, indicating that the principle of fairness is being upheld.

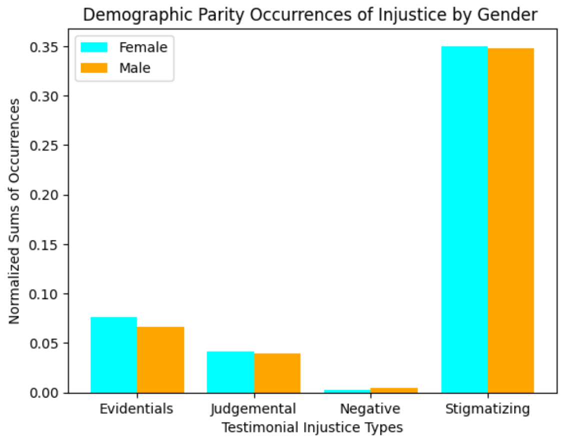


Figure 1: Demographic Parity Occurrences of Injustice by Gender.

Race. In terms of Demographic Parity race analysis, there was little to no disparate treatment detected across all term types between the different races of patients, indicating that there was minimal evidence of injustice in the data based on race, as observed in Figure 2. We observe that Latino patients are the most likely to experience evidential words, while Asian patients were the least likely. Further, for evidential words, White patients have privilege over Black patients, Black patients have privilege over Latino patients, and Asian patients have privilege over White and Latino patients. For judgemental words, Black patients are the most likely, and Asian patients were the least likely to experience judgemental words. Here, we observe that White patients have privilege over Black patients. Latino patients were the most likely and Asian patients were the least likely to experience negative words in their medical records. We note here that negative terms were the least likely to appear in the records of any patient. Black patients were the most likely and Asian patients were the least likely to experience stigmatizing words in their medical records. Another observation is that White patients have privilege over Black patients, Asian patients have privilege over every race of

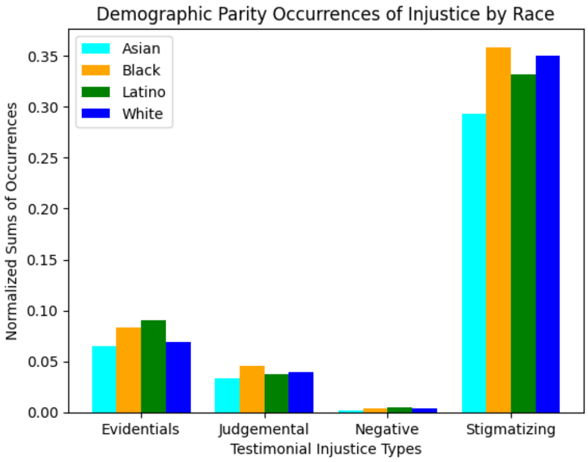


Figure 2: Demographic Parity Occurrences of Injustice by Race.

patients, and Latino patients have privilege over White patients. Stigmatizing words occurred the most in everyone's medical records. With this, race should also not be found to be a significant predictor of the treatment or care received by patients. Therefore, the findings of the analysis should show that a person's racial membership does not have any substantial impact on how they are treated, indicating that the principle of fairness is being upheld.

Since our analysis using demographic parity showed that neither race nor gender affect how a patient experiences testimonial injustice, when we observe their intersection, we should see that the treatment and care received by patients are not affected by the intersectionality of race and gender. This would indicate that the principle of fairness is being upheld regardless of a patient's race or gender. However, we see a different story when we consider intersectionality.

5.2 Differential Fairness

Differential fairness focuses on the intersectionality of race and gender in relation to testimonial injustice. The results of the demographic parity experiments showed, there are no disparities in how groups are treated with respect to testimonial injustice upon race or gender. However, the results of the experiment pertaining to differential fairness show that there are disparities between different intersections of gender and race with respect to the types of terms that lead to testimonial injustice. Specifically, out of 112 comparisons for each intersection of gender and race, 110 violations of differential fairness occurred. This demonstrates that there are underlying injustices occurring in how different groups are treated based on gender and race and that we cannot simply rely on measures that do not consider intersectionality to reveal this.

There were very few instances in which fairness was not violated, such as Asian males to Asian females for evidential and judgmental words, and Asian males to Latina females for negative words. The results showed that Asian females and males were the most privileged, and White males and females were the least privileged when fairness was violated. This may be due to the fact that there are many more records for White patients than all other races of patients. As observed in Figure 3, across all types of terms that lead to testimonial injustice, Black females were the next least privileged after White patients. Black males were found to have more privilege in experiencing testimonial injustice than Black females. The experiment was also conducted with 500 randomly sampled records of each subgroup of patient, and the results there showed that when unfairness is present, Black females are the most marginalized, and Asian males are the least. For these sampled records, across all types of terms that lead to testimonial injustice, Latina females were the most marginalized for evidential words, Black females for judgment words and negative words, and Latino males for stigmatizing words. However, even with the full dataset, Asian males were consistently found to be the most privileged of all the groups represented.

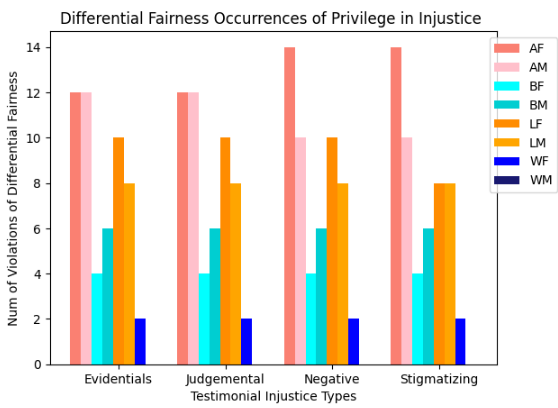


Figure 3: Differential Fairness Occurrences of Injustice by Gender and Race.

5.3 Subgroup Fairness

In this experiment, similar to differential fairness, we focus on the intersectionality of race and gender in relation to testimonial injustice. The results of the demographic parity experiments showed, there were no disparities in how groups were treated with respect to testimonial injustice upon race nor gender. However, the results of the differential fairness experiments showed there are differences in how one is treated based on their race and gender. We conduct an experiment that also looks at intersectionality of groups to compare if there is a difference in how these two metrics reveal disparate treatment amongst the subgroups.

Based on our analysis of demographic parity in detecting testimonial injustice in medical records, we found that the privileged groups by race are Asian and White patients, as well as males. Therefore, for the purpose of intersectional fairness analysis, we consider Asian men and White men as non-sensitive groups. When we conducted a differential fairness analysis, we found that violations occurred 110 times out of 112 comparisons (each intersection of gender and race for each type of term leading to testimonial injustice). We expected similar results (Figure 4) for subgroup fairness analysis. Our subgroup fairness metric detected 69 violations our of the 112 comparisons of subgroups. Though less occurrences of violations are present, this still reveals we must consider intersectionality within the medical setting and in the fairness metrics we use there. If even better highlights that a metric which considers intersectionality is not enough, but we must be careful at which fairness metrics we use based on the tasks at hand.

For evidential terms, we found that Latina females were the most discriminated against, while Asian males were the most privileged. For judgment terms, Black males were the most discriminated against, while Asian males were the most privileged. For negative words, Asian males were the most privileged, while Latino males were the least privileged. For stigmatizing words, Black females were the most discriminated against, while Asian males were again the most privileged. It is important to note that our experiment includes the entire dataset, which is over-representative of White patients. Thus, we can expect even larger disparities in how different groups are treated with a more representative dataset. This does not mean that White patients do not experience discrimination, but rather emphasizes the importance of having a more representative dataset to better understand the degrees to which different groups may experience testimonial injustice in their records.

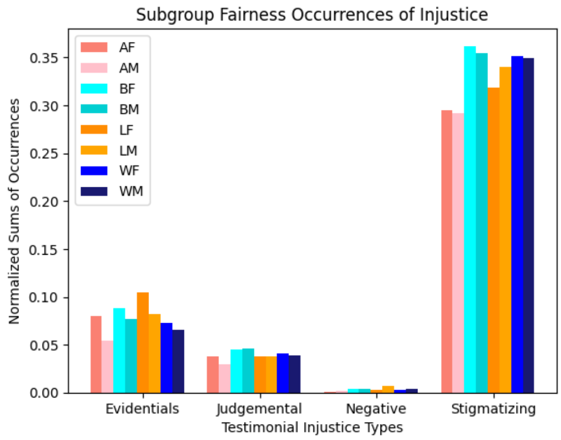


Figure 4: Subgroup Fairness Occurrences of Injustice by Gender and Race.

6 Discussion

When conducting experiments using demographic parity, we compared race or gender. In each case, there were no violations of demographic parity for any patient is treated based on their race or gender alone. If a practitioner takes these results for face value, they might determine there is no form of discrimination happening based on these commonly observed visible attributes. For example, when speaking to a Black male patient who was stigmatized against from the demographic parity view, they would have no evidence in that setting to back their expression of their experience. However, when we look deeper, through the lens of intersectional fairness (i.e., differential fairness and subgroup fairness) at the intersection of race and gender, we can see that a male patient can still experience discrimination (i.e. Black males) and so could a White patient (i.e. White females).

When we look at measures that consider intersectionality, we see disparity in how people are treated based on their race and gender for every type of word we analyzed that could lead to testimonial injustice. We attribute this to: (1) being able to consider multiple aspects about a person that might only reveal themselves at the intersection of race and gender, (2) in differential fairness being able to constrain the range in which we look for violations, as opposed to only looking at it from one side as demographic parity does. To properly see injustices occurring, we must look at all angles from which they could possibly be coming from. This is because someone might only be testimonially injust toward a person who is female, others might only act unjustly because of your membership with a historically marginalized race, and so on. We contend that the better metrics to use for detecting injustices, e.g. testimonial injustice, in medical records are ones which consider intersectionality. Still, we see differences in how these measures show which groups are experiencing privilege, thus we must be careful in understanding the goals of the fairness metrics we use.

7 Conclusions

The objective of this empirical study was to investigate the potential benefits of intersectionality in detecting testimonial injustice, using medical records as a real-world application. Demographic parity, differential intersectional fairness, and subgroup fairness were used to examine whether there are differences in the extent of testimonial injustice experienced by individuals based on the intersection of their demographic attributes and if intersectionality helps reveal this. Our results showed (1) when we allow ourselves to use metrics that consider intersectionality, as opposed to sole factors of who a person is, we can better see disparities in how they are treated in terms of detecting testimonial injustice in medical records, (2) there are differences in how someone is treated based on the intersection of their demographic attributes (3) different intersectional fairness metrics do reveal these injustices differently. While demographic parity did not show a clear disparate impact based on gender or race, differential intersectional fairness and subgroup fairness – two intersectional fairness measures – revealed that there was disparate treatment based on both gender and race. These findings suggest that intersectionality should be considered when detecting testimonial injustice, especially in medical settings.

8 Limitations and Future Work

Data. A challenge we faced was that MIMIC-III was unevenly distributed across the races (e.g. ethnicities) for the patients represented. We had significantly more White and Black patients than any other race of people and even still many more White than Black patients. Therefore we continue to express the need for more representative, inclusive, and balanced datasets. Further, the dataset did include ethnic breakdowns, but due to the lack of patients present in those ethnic groups we could not include Caribbean or Middle Eastern patients as well as many other subgroups in our analysis. We would like to use a more comprehensive dataset in the future, potentially from a facility that consistently services marginalized and privileged communities. If we had more time, we would like to partner with a medical facility that regularly serves marginalized and non-marginalized groups, steadily, to develop a dataset which captures more features that could reveal some bias and ensure they are more descriptive (i.e. has_insurance) to get higher quality data.

Better Feature Selection and Using More Demographic Features. To ensure the quality of the aforementioned data, we will perform a causal analysis to identify the specific features that cause testimonial injustice. We anticipate that variables such as age and education level of patients need be included, as these factors have been shown to affect how patients are treated, particularly in the medical field (Dunsch et al., 2018; DeVoe et al., 2009).

Fairness Metrics. Existing and popular, fairness metrics cannot be generalized to fit in settings where intersectionality must be considered. Another challenge we faced was having a lack of good baselines to use when analyzing intersectional differences. Intersectionality is highly unexplored, in the future we would like to develop our own metric which can be more beneficial in detecting intersectional disparate treatment between individuals.

Additional Analysis. We plan to conduct additional analysis to understand if specific physicians treat similar patients similarly based on the intersection of their demographic features. Further, we plan to perform statistical significance testing on differences in how patients were treated based on the intersection of their demographic features and the occurrences of specific physicians' use of testimonial unjust terms to other patients.

Acknowledgements

This paper is based upon work supported in part by the NSF LSAMP Bridge to the NSF Program on Fairness in AI in Collaboration with Amazon under Award No. IIS-1939743, titled FAI: Addressing the 3D Challenges for Data-Driven Fairness: Deficiency, Dynamics, and Disagreement (Kenya Andrews). This work is also supported in part by the Cisco Research Gift Grant (Lu Cheng). Any opinion, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation, Amazon, or Cisco Research.

References

Hammaad Adam, Ming Ying Yang, Kenrick Cato, Ioana Baldini, Charles Senteio, Leo Anthony Celi, Jiaming Zeng, Moninder Singh, and Marzyeh Ghassemi. 2022. Write it like you see it: Detectable differences in clinical notes by race lead to differential model recommendations. New York, NY, USA. Association for Computing Machinery.

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR.

Amit Bali, Deepika Bali, Nageshwar Iyer, and Meenakshi Iyer. 2011. Management of medical records: facts and figures for surgeons. *Journal of maxillofacial and oral surgery*, 10:199–202.

Greta R. Bauer and Daniel J. Lizotte. 2021. Artificial intelligence, intersectionality, and the future of public health. *American Journal of Public Health*, 111(1):98–100. PMID: 33326280.

Mary Catherine Beach, Somnath Saha, Jenny Park, Janiece Taylor, Paul Drew, Eve Plank, Lisa A Cooper, and Brant Chee. 2021. Testimonial injustice: linguistic bias in the medical records of black patients and women. *Journal of general internal medicine*, 36(6):1708–1714.

Erin Beeghly and Alex Madva. 2020. *An introduction to implicit bias: Knowledge, justice, and the social mind*. Routledge.

Sue Bowman. 2013. Impact of electronic health record systems on information integrity: quality and safety implications. *Perspectives in health information management*, 10(Fall).

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Elizabeth N. Chapman, Anna Kaatz, and Molly Carnes. 2013. Physicians and implicit bias: How doctors may unwittingly perpetuate health care disparities. *Journal of General Internal Medicine*, 28(11):1504–1510.

Lu Cheng and Huan Liu. 2023. *Socially Responsible AI: Theories and Practices*. World Scientific.

Lu Cheng, Kush R Varshney, and Huan Liu. 2021. Socially responsible ai algorithms: Issues, purposes, and challenges. *Journal of Artificial Intelligence Research*, 71:1137–1181.

Jennifer E DeVoe, Lorraine S Wallace, and George E Fryer Jr. 2009. Patient age influences perceptions about health care communication. *Family medicine*, 41(2):126.

Felipe Dunsch, David K Evans, Mario Macis, and Qiao Wang. 2018. Bias in patient satisfaction surveys: a threat to measuring healthcare quality. *BMJ global health*, 3(2):e000694.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

Chloë FitzGerald and Samia Hurst. 2017. Implicit bias in healthcare professionals: a systematic review. *BMC medical ethics*, 18(1):1–18.

James R. Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2020. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1918–1921.

Miranda Fricker. 2019. Testimonial injustice. *Contemporary Epistemology: An Anthology*, pages 149–163.

Avijit Ghosh, Lea Genuit, and Mary Reagan. 2021. Characterizing intersectional group fairness with worst-case comparisons. In *Proceedings of 2nd Workshop on Diversity in Artificial Intelligence (AIDBEI)*, volume 142 of *Proceedings of Machine Learning Research*, pages 22–34. PMLR.

Usman Gohar and Lu Cheng. 2023. A survey on intersectional fairness in machine learning: Notions, mitigation, and challenges. In *IJCAI*.

Wei Guo and Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. New York, NY, USA. Association for Computing Machinery.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Gracie Himmelstein, David Bates, and Li Zhou. 2022. Examination of Stigmatizing Language in the Electronic Health Record. *JAMA Network Open*, 5(1):e2144967–e2144967.

Daniel Holman, Sarah Salway, Andrew Bell, Brian Beach, Adewale Adebajo, Nuzhat Ali, and Jabeer Butt. 2021. Can intersectionality help with understanding and tackling health inequalities? perspectives of professional stakeholders. *Health Research Policy and Systems*, 19(1):97.

Christoph Hube and Besnik Fetahu. 2018. Detecting biased statements in wikipedia. Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Jill Jin. 2021. What to Consider When Reading Your Medical Notes. *JAMA*, 326(17):1756–1756.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2564–2572. PMLR.

Nancy Krieger. 1990. Racial and gender discrimination: Risk factors for high blood pressure? *Social Science Medicine*, 30(12):1273–1281.

Sachin Kumar Mangla, Ankur Chauhan, Tanmoy Kundu, and Abbas Mardani. 2023. Emergency order allocation of e-medical supplies due to the disruptive events of the healthcare crisis. *Journal of Business Research*, 155:113398.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems*, 30.

Bruce G Link and Jo C Phelan. 2001. Conceptualizing stigma. *Annual review of Sociology*, 27(1):363–385.

Ana Cristina Marques. 2018. Patricia hill collins and sirma bilge, intersectionality.

Susan E Martin. 1994. "outsider within" the station house: The impact of race and gender on black women police. *Social Problems*, 41(3):383–400.

Anna P Goddu, Katie J O'Conor, Sophie Lanzkron, Mustapha O Saheed, Somnath Saha, Monica E Peek, Carlton Haywood, and Mary Catherine Beach. 2018. Do words matter? stigmatizing language and the transmission of bias in the medical record. *Journal of general internal medicine*, 33:685–691.

Jenny Park, Somnath Saha, Brant Chee, Janiece Taylor, and Mary Catherine Beach. 2021. Physician Use of Stigmatizing Language in Patient Medical Records. *JAMA Network Open*, 4(7):e2117052–e2117052.

Yolanda A. Rankin and Jakita O. Thomas. 2020. The intersectional experiences of black women in computing. New York, NY, USA. Association for Computing Machinery.

Vincent J Roscigno, Diana L Karafin, and Griff Tester. 2009. The complexities and processes of racial housing discrimination. *Social Problems*, 56(1):49–69.

Michael Sun, Tomasz Oliwa, Monica E. Peek, and Elizabeth L. Tung. 2022. Negative patient descriptors: Documenting racial bias in the electronic health record. *Health Affairs*, 41(2):203–211. PMID: 35044842.

Douglas B. White and Bernard Lo. 2020. A Framework for Rationing Ventilators and Critical Care Beds During the COVID-19 Pandemic. *JAMA*, 323(18):1773–1774.

Stefanos A Zenios, Lawrence M Wein, and Glenn M Chertow. 1999. Evidence-based organ allocationaccess the "journal club" discussion of this paper at http://www.elsevier.com/locate/ajmselect/. *The American Journal of Medicine*, 107(1):52–61.

A Appendix

A.1 Normalized Sums of Unjust Terms

	ETHNICITY	GENDER	evidentials	judge	neg	stig	tot_records	all_sum
0	WHITE	F	1127.461072	628.037035	36.943441	5411.386544	2998983	7203.828092
1	BLACK	F	221.803952	113.947568	10.194630	913.227860	1343317	1259.174010
2	ASIAN	F	40.882661	19.303682	0.413122	151.041180	80437	211.640645
3	LATINO	F	69.251953	24.994622	1.560188	211.187412	170878	306.994175
4	MIDDLE EASTERN	F	0.694151	0.165846	0.023256	3.583392	291	4.466645
5	CARIBBEAN	F	0.600000	0.000000	0.000000	0.733333	23	1.333333
6	WHITE	M	1327.707553	789.633200	87.403889	7093.839355	3963336	9298.583998
7	BLACK	M	156.845940	94.634272	8.078835	723.663970	1919285	983.223017
8	ASIAN	M	37.644775	20.407761	1.521424	201.057112	109762	260.631073
9	LATINO	M	84.873032	39.084754	6.806472	354.373769	276852	485.138027
10	MIDDLE EASTERN	M	1.134574	0.714829	0.165254	12.310840	7460	14.325497
11	CARIBBEAN	M	0.808468	0.375000	0.000000	3.211022	435	4.394489

Figure 5: Normalized sums of occurrences of unjust terms for patients based on race and gender. Higher numbers indicate higher counts of terms.

A.2 Abbreviations

Abbreviation	Full Form
F	Female
M	Male
W	White
B	Black
A	Asian
L	Latino
WF	White Female
BM	Black Male
...	...

Table 2: Abbreviations of Demographic Features and their Combinations.

A.3 Intersectional Analysis of Terms

In conducting analysis on the MIMIC-III dataset, we plot the distributions of the occurrences of each term which can lead to testimonial injustice. The position of the peak in the distribution graph provides insight into which subgroups are experiencing a stronger degree of injustice. The more right-skewed the peak of the distribution is, the higher amount of injustice experienced by that particular subgroup. Naturally, the height of the peak speaks to the confidence of the severity to which that subgroup is experiencing injustice based on their word count.

In comparing Figures 6 and 7 notice in terms of race, Asian patients experience evidential terms the second least, after White patients. Still, Asian Females have the second most highest occurrences of evidential terms, which is a clear contradiction, showing the importance of observing intersectional experiences. In Figure 7, we observe the normalized distribution of evidential terms used for patients across different intersections of races and genders. White men, Asian men, and White females show lower amounts of evidential terms in their records, while Latina females, Asian females, and Black females have higher occurrences of evidential terms in their medical records.

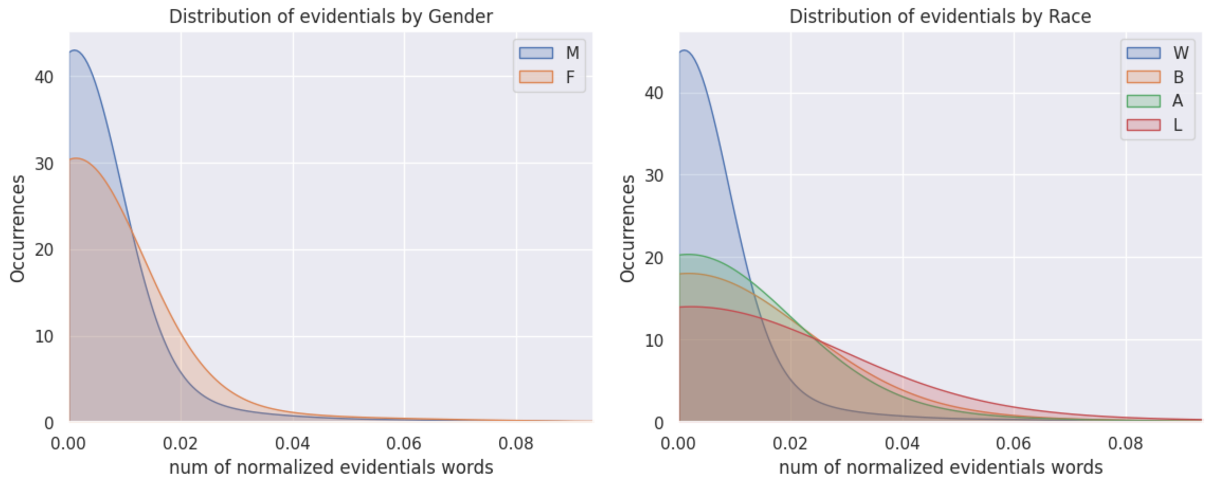


Figure 6: Distribution of Evidential terms in medical notes, refer to legend in Table 2 to see the full text of the abbreviated terms. Left: Shows distribution of gender-only. Right: Shows the distribution of the intersection of race-only.

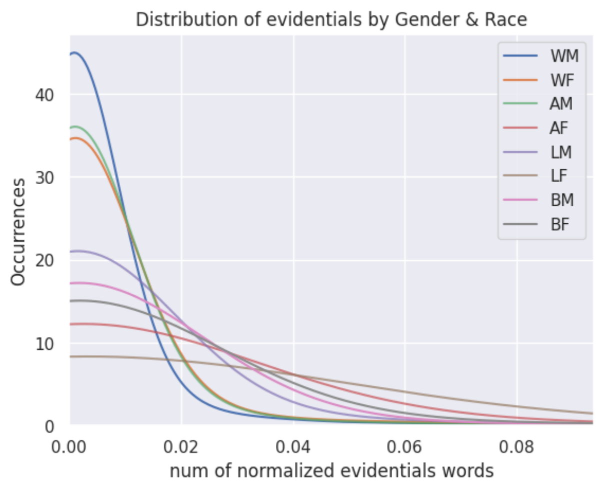


Figure 7: Distribution of Evidential terms considering intersectionality in medical notes, refer to legend in Table 2 to see the full text of the abbreviated terms.

From the normalized distributions of the occurrence of judgement terms in the medical records, in Figure 8 we can observe that female patients as opposed to male patients and Black patients as apposed to the other races, studied here, have the most occurrences of judgement terms. Figure 9 emphasizes just how much worse Black women are impacted than any other subgroup. Black men and White women are the next two most vulnerable groups to experiencing judgemental terms in their physicians' notes. Latino men, White men, and Latina females have the least occurrences of judgement terms in their records.

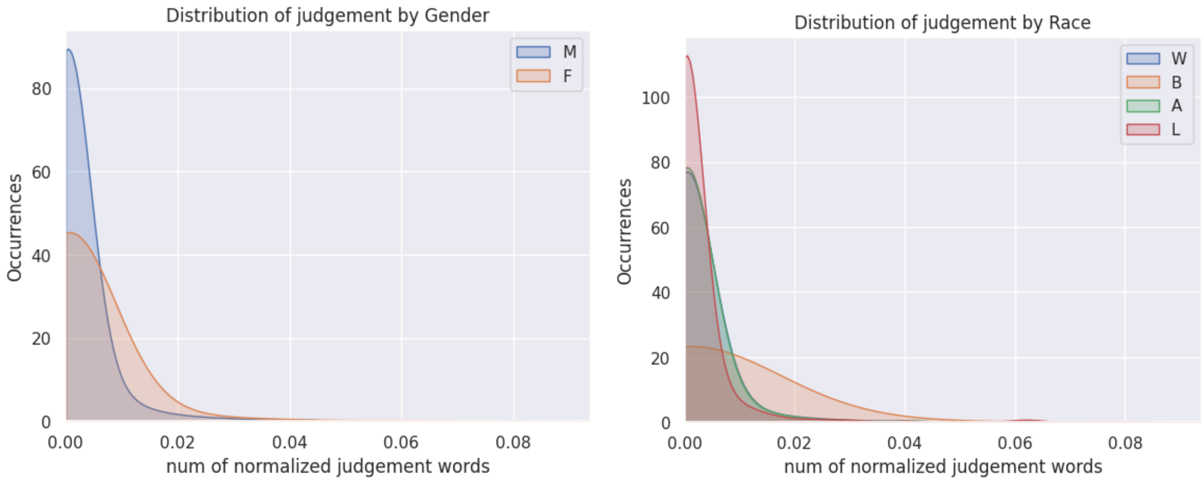


Figure 8: Distribution of Judgement terms in medical notes, refer to legend in Table 2 to see the full text of the abbreviated terms. Left: Shows distribution of gender-only. Right: Shows the distribution of the intersection of race-only.

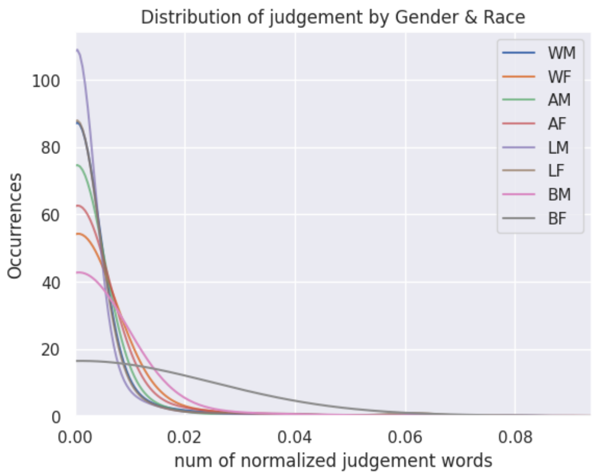


Figure 9: Distribution of Judgement terms in medical notes, refer to legend in Table 2 to see the full text of the abbreviated terms. Refer to Figure 8 to see gender-only and race-only graphs.

From Figure 11 we observe the distributions of normalized stigmatizing terms used for patients over the intersection of their race and gender. Asian men, followed by Asian females and White males have experienced the least stigmatizing language in the physicians' notes, while Black females and Latino men have been faced with it the most. Figure 6 suggests that Latino and Black patients receive similar treatment, however, Figure 7 highlights that stigmatizing language is more prevalent in the medical records of Black females and Latino males compared to any other subgroups.

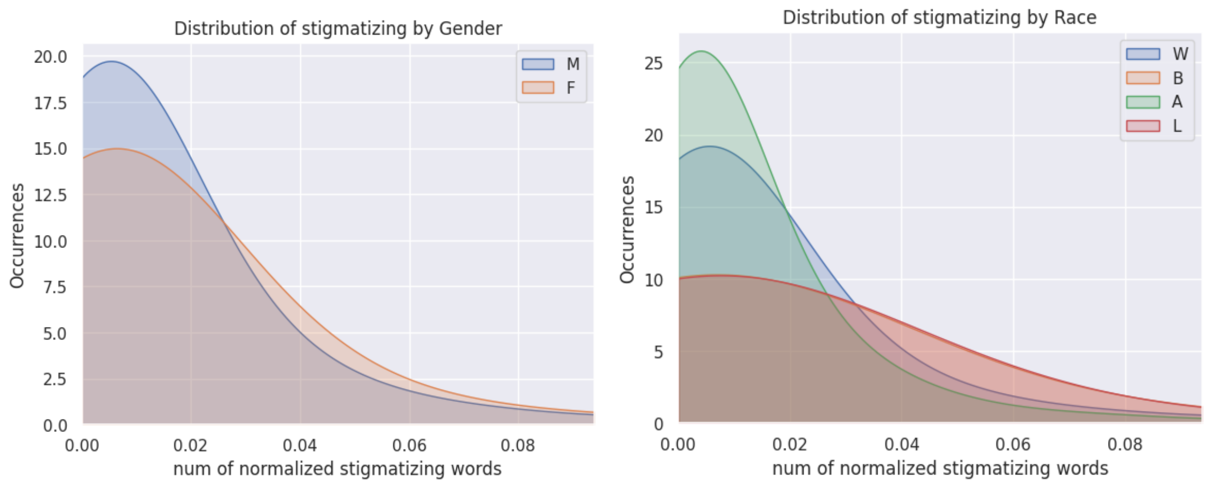


Figure 10: Distribution of Stigmatizing terms in medical notes, refer to legend in Table 2 to see the full text of the abbreviated terms. Left: Shows distribution of gender-only. Right: Shows the distribution of the intersection of race-only.

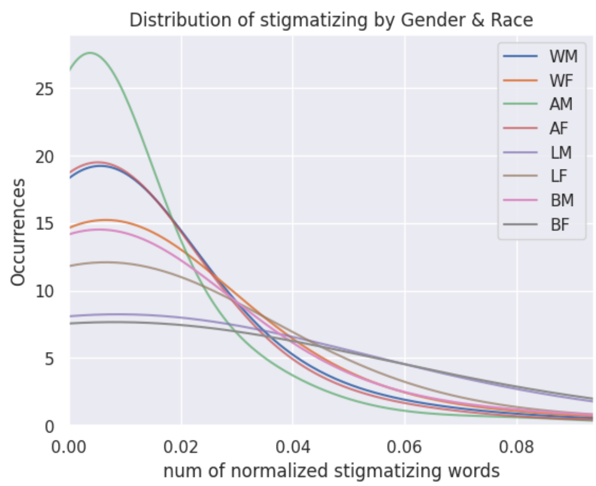


Figure 11: Distribution of Stigmatizing terms in medical notes, refer to legend in Table 2 to see the full text of the abbreviated terms. Refer to Figure 10 to see gender-only and race-only graphs.

To conclude, these graphs specifically their variations show the importance of exploring intersectionality while providing medical care. For example, Black females face challenges that are unique to their intersectional identity as both black and female. This intersectionality can result in compounded experiences of discrimination and marginalization. Furthermore, the fact that Asian and White males consistently occupy the most privileged subgroup highlights systemic inequalities and the need for continued efforts to address these disparities.

Interactive Span Recommendation for Biomedical Text

Louis Blankemeier
Stanford University
lblankem@stanford.edu

Theodore Zhao
Microsoft Health AI
theodorezhao@microsoft.com

Robert Tinn
Microsoft Health AI
robert.tinn@microsoft.com

Sid Kiblawi
Microsoft Health AI
sidkiblawi@microsoft.com

Yu Gu
Microsoft Health AI
aiden.gu@microsoft.com

Akshay S. Chaudhari
Stanford University
akshaysc@stanford.edu

Hoifung Poon
Microsoft Research
robert.tinn@microsoft.com

Sheng Zhang
Microsoft Research
zhang.sheng@microsoft.com

Mu Wei
Microsoft Health AI
muhsin.wei@microsoft.com

Joseph S. Preston
Microsoft Health AI
sam.preston@microsoft.com

Abstract

Motivated by the scarcity of high-quality labeled biomedical text, as well as the success of data programming, we introduce *KRISS-Search*. By leveraging the Unified Medical Language Systems (UMLS) ontology, KRISS-Search addresses an interactive few-shot span recommendation task that we propose. We first introduce *unsupervised KRISS-Search* and show that our method outperforms existing methods in identifying spans that are semantically similar to a given span of interest, with $>$ 50% AUPRC improvement relative to PubMedBERT. We then introduce *supervised KRISS-Search*, which leverages human interaction to improve the notion of similarity used by unsupervised KRISS-Search. Through simulated human feedback, we demonstrate an enhanced F1 score of 0.68 in classifying spans as semantically similar or different in the low-label setting, outperforming PubMedBERT by 2 F1 points. Finally, supervised KRISS-Search demonstrates competitive or superior performance compared to PubMedBERT in few-shot biomedical named entity recognition (NER) across five benchmark datasets, with an average improvement of 5.6 F1 points. We envision KRISS-Search increasing the efficiency of programmatic data labeling and also providing broader utility as an interactive biomedical search engine.

1 Introduction

One of the major challenges in developing machine learning models for biomedical text analysis is the scarcity of high-quality labeled data. Manual annotation of biomedical text is a time-consuming process that demands specialized expertise, leading researchers to investigate alternative methods such as weak supervision (Zhang et al., 2022a; Yakimovich et al., 2021; Poon et al., 2021; Lang and Poon, 2020) and active learning(Naseem et al., 2021; Ren et al., 2020) to address this bottleneck. Programmatic data labeling (Ratner et al., 2016, 2017b,a), a form of weak supervision in which domain experts develop heuristics (labeling functions) to provide noisy labels for large datasets, has been shown to be effective in leveraging domain expertise. However, developing diverse and high-quality labeling functions can be challenging, as it requires knowledge of the programmatic rule specification. Some techniques have been proposed to suggest labeling functions to users (Boecking et al., 2021; Zhao et al., 2021; Li et al., 2021), but they still rely on users' understanding of rule specifications to evaluate or modify the labeling functions.

To address this challenge, we introduce an interactive span recommendation task. Our key idea is to train a single model and adapt it to human feedback, enabling it to understand and treat similarity at various levels of granularity. This approach eliminates the need to train multiple models for different notions of similarity. Conventional entity linking is one such notion of similarity, where a user may want to identify all mentions of a specific concept, such as "hypertension disease" from the Unified Medical Language System (UMLS). However, a user may want the flexibility to iden-

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 373–384
July 14, 2023

tify not only mentions of "hypertension disease" but also those of "hypertension treatments" and "hypertension comorbidities" simultaneously. This task extends beyond entity linking and can be more broadly described as interactive span recommendation.

To tackle the interactive span recommendation task, we propose *KRISS-Search*, a method that enables domain experts to develop span recommendation models for searching unlabeled corpora. A crucial aspect of model performance in KRISS-Search is the choice of embedding space. The UMLS ontology offers a comprehensive set of biomedical concepts organized hierarchically. We adapt the UMLS-based self-supervised training technique of KRISSBERT to generate the embedding space used by our method.

We evaluate two versions of KRISS-Search. *Unsupervised KRISS-Search* takes a single user-selected query span from a biomedical corpus as input and returns semantically similar spans. However, in some cases, this single measure of similarity may not adequately overcome the inherent ambiguity in identifying spans based on one example. To address this limitation, we introduce *supervised KRISS-Search*, which employs active learning to incorporate human feedback and refine the concept of similarity used in the unsupervised version. In the context of programmatic data labeling, we envision unsupervised KRISS-Search recommending terms for users to incorporate into labeling functions and supervised KRISS-Search directly generating noisy labels, providing a more flexible alternative to labeling functions.

Our main contributions can be summarized as follows:

1. We demonstrate that unsupervised KRISS-Search outperforms PubMedBERT (Gu et al., 2020) by 51% area under the precision-recall curve (AUPRC) in returning spans with *exact* concept unique identifier (CUI) matches to the CUI associated with the query span. KRISS-search further outperforms PubMedBERT by 54% in returning spans with *similar* associated CUIs.

2. By extending unsupervised KRISS-Search to supervised KRISS-Search through human-feedback and active learning, we surface spans associated with specific concepts (CUIs) with an F1 of 0.68, outperforming PubMedBERT by 2 F1 points.

3. We demonstrate that supervised KRISS-Search performs comparably or outperforms PubMedBERT across five benchmark tasks in the few-shot biomedical NER setting. On average, supervised KRISS-Search outperforms PubMedBERT by 5.6 F1 points, demonstrating the flexibility of our method to handle various levels of granularity.

2 Methods

In this paper, we compare various training strategies for the BERT-base (Devlin et al., 2018) (100 million parameters) architecture in order to address our proposed task. While the training strategies discussed in this paper are specific to the BERT-base architecture, they can also be applied to larger models. The methods we evaluate can be characterized as "contextual," "in-domain," "contrastive," and "interactive." "Contextual" methods use the surrounding context to make recommendations, while "in-domain" methods are trained on data specifically related to the biomedical domain. "Contrastive" methods utilize semantic similarity and dissimilarity during the training process. "Interactive" methods involve human participation to guide model training. The four training strategies we compare are BERT, PubMedBERT, unsupervised KRISS-Search, and supervised KRISS-Search. Each strategy implements an additional descriptor in the order they were listed, with supervised KRISS-Search implementing all four.

To highlight the distinctions between BERT, PubMedBERT, unsupervised KRISS-Search, and supervised KRISS-Search, consider the following example. In the sentence "The patient received a pt assay," the query span "pt" refers to the concept "prothrombin time assay". BERT, which is not specifically tailored to the biomedical domain or designed to employ contrastive or interactive techniques, may surface a false positive "platinum," which shares the same abbreviation "pt" but is not relevant to the biomedical domain. Similarly, PubMedBERT, which is trained on biomedical data but does not utilize contrastive learning, may generate a false positive "physical therapy," which is in the biomedical domain but semantically dissimilar to the query span. In contrast, both unsupervised and supervised KRISS-Search utilize contrastive learning, which makes them more likely to recommend semantically similar spans, such as "prothrombin time assay", as this similarity is explicitly incorporated into the training process. Now consider another example: "decrease in right lung mass compared to prior imaging". Here, the user is interested in the query span "decrease in right lung mass", which represents a relationship between "decrease" and "right lung mass". In this scenario, the concept of similarity is complex and may require the interactive feature of supervised KRISS-Search to surface similar spans.

2.1 Efficiently Embedding the Corpus

We posit that KRISSBERT (Zhang et al., 2022b) serves as an excellent foundation for our method, as it is trained using a contrastive learning approach based on the UMLS ontology that enables it to effectively predict the correspondence of multiple entities to the same underlying concept, a task known as entity-linking. However, in its original form, KRISSBERT is not computationally tractable for our use case. The KRISSBERT model (Zhang et al., 2022b) uses the [CLS] token to represent the contextual embedding of a span and places entity tokens between the span and its context to communicate the span of interest to the model. As such, generating embeddings for X spans requires X forward passes. This can prove computationally intractable when the number of spans to embed is large. To address this issue, the KRISS-Search method removes the entity tokens from the mention representations and instead aggregates the final layer embeddings of the tokens in a span to generate the span's embedding. Fig. 2a shows how KRISSBERT uses entity tokens (corresponding embeddings shown in orange) to denote the entity and [CLS] embeddings to compute the contrastive loss. Fig. 2b shows how KRISS-Search removes the entity tokens and aggregates the final layer embeddings of the entity tokens to compute the loss. The dummy text snippets in Fig. 2 provide an example of a positive pair where "patient discharge" and "released" correspond to the same concept and are thus pulled together in the embedding space during contrastive training. The entity encoder is left unchanged and is trained jointly with the mention encoder, as we hypothesize that the hierarchical UMLS ontology embedded in the entity encoder is useful for the task. For training, we used a single Tesla V100 16GB GPU.

These modifications increase computational efficiency by reducing the number of forward passes required for generating embeddings. If we pass 512

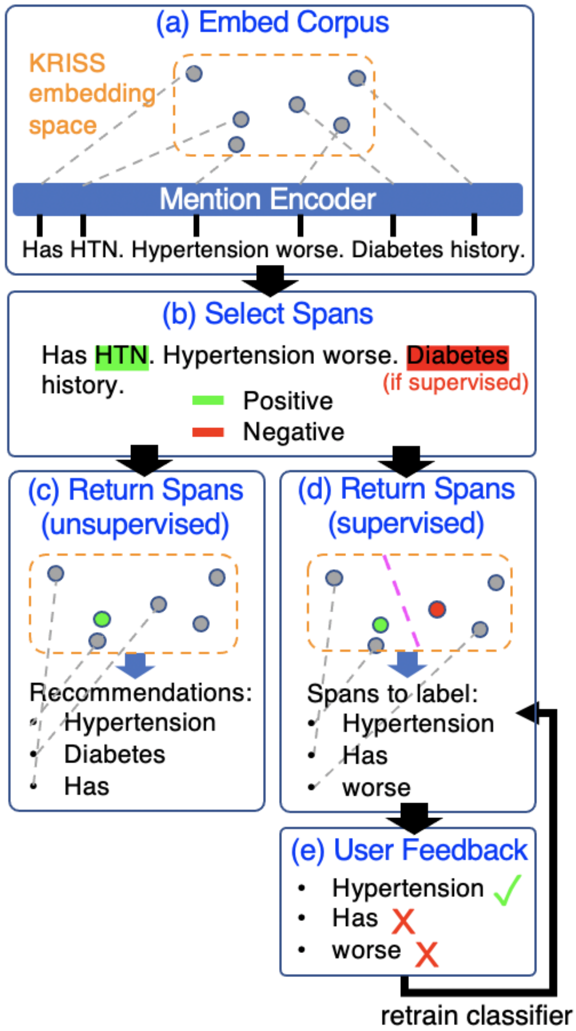
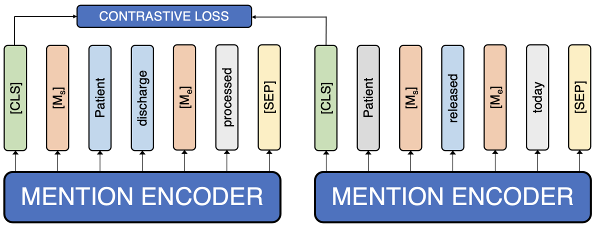
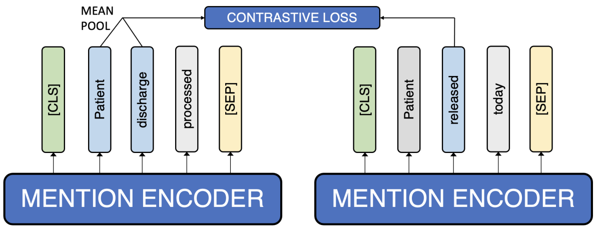


Figure 1: The KRISS-Search method consists of the following steps: (a) embed the corpus using the KRISSBERT embedding space, which places mentions of the same concept closer together and different concepts further apart; (b) the user selects spans to seed supervised and unsupervised KRISS-Search. For unsupervised KRISS-Search, the user selects a single positive query span. For supervised KRISS-Search, the user selects any number of positive and negative spans; (c) in unsupervised KRISS-Search, nearest neighbors to the positive query span are returned; (d) in supervised KRISS-Search, active learning is used to train a light-weight classifier to refine recommendations, with examples closest to the decision boundary being returned for subsequent active learning; (e) the user provides feedback on the returned spans, which can be used to retrain the light-weight classifier and return to step (d).



(a) KRISSBERT mention encoder training with entity tokens. The [CLS] token is used for computing the contrastive loss. This is the approach used in the original KRISSBERT paper.



(b) KRISSBERT mention encoder training without entity tokens. Span token embeddings (blue) are agggregated to generate the span embeddings and compute the contrastive loss. This is the strategy adopted for KRISS-Search.

Figure 2: A comparison of the mention encoder training with and without the entity tokens.

tokens (the maximum sequence length of BERT-base) into our model during a single forward pass, our method reduces inference time by N $\times$ 512 where N is the maximum span length that we embed. Additionally, our approach allows us to leverage the contrastive loss while still maintaining per-token embeddings. We use the same hyperparameters to retrain KRISSBERT and observe marginally degraded performance on validation data for the original KRISSBERT entity linking task. We note that this is expected as the KRISSBERT hyperparameters optimize validation performance of the original model. As the goal of this paper is not entity linking, we leave re-selecting hyperparameters to future work.

To further increase the efficiency of our method, we also filter the embeddings, discarding spans where the tokenization (Honnibal and Montani, 2017) of the span triggers a stop token, punctuation token, or whitespace token based on the assumption that such spans are not generally of interest.

2.2 Unsupervised KRISS-Search

The unsupervised KRISS-Search task involves returning a ranked list of spans from the corpus that are semantically similar to a query span, as determined by the L2 distance of their embedding to the query span embedding.

Evaluation: For evaluation of unsupervised KRISS-Search, we use the n2c2 dataset (2019 n2c2/UMass Lowell shared task 3) (Luo et al., 2020). This dataset contains 100 discharge summaries labeled with CUI annotations. We choose this dataset as it represents a domain shift from the PubMed abstracts used to train KRISSBERT. Additionally, n2c2 is annotated with diverse entities, including medical problems, treatments, and tests from established ontologies (Liu et al., 2005; Spackman et al., 1997).

To evaluate the quality of the retrieved spans, we assess the model's ability to retrieve (1) spans with associated CUIs that match the CUI associated with the query span (*same* evaluation type in Tables 1, 2, and 3) and (2) spans with associated CUIs that are closely related to the CUI associated with the query span (*related* evaluation type in Tables 1, 2, and 3). Related CUIs are generated by sampling a parent CUI of the query-associated CUI and returning its children using the UMLS hierarchy (Bodenreider, 2004). The *same* evaluation type experiments indicate how well each approach is at returning specific concepts of interest, while the *related* evaluation type experiments measure how well each approach can return more loosely related concepts.

We adopt a relaxed evaluation measure where spans that overlap with a concept mention are associated with the concept. We apply relaxed evaluation as we hypothesize that for our task, generating precise span boundaries is less important than providing the user with a greater number of recommendations. We represent spans with the mean of the span token embeddings. We choose the test query spans, used in Tables 1, 2, 3, as well as Figures 5 and 10, as follows. For 255 CUIs with more than 25 mentions in the corpus and corresponding span embeddings, we randomly sample one span for each of the 255 CUIs. We select CUIs that appear more than 25 times hypothesizing the difficulty of comparing approaches using low-prevalence CUIs.

To assess the model performance, we calculate the average precision, recall, and F1 metrics for a varying number of retrieved spans (Fig. 1 and Fig. 10). Specifically, we evaluate the performance at $1 \times N$, $2 \times N$, and $3 \times N$, where N represents the total number of mentions of a specific CUI in the dataset. It is important to note that N varies across different CUIs. The precision, recall, and F1 metrics are computed based on the number of correctly retrieved mentions of a specific CUI relative to the total number of CUI mentions present in the dataset. The denominator of precision corresponds to the number of nearest neighbors retrieved, while the denominator of recall corresponds to the total number of mentions in the corpus for each CUI. These average values are not optimally informative as performance across different CUIs varies widely for all methods. As such, we also report per-query measures (Table 1). We compute average per-query percent recall improvement of KRISS-Search compared to PubMedBERT ($\overline{\%\Delta}$ in Table 1) and the frequency with which unsupervised KRISS-Search outperforms PubMedBERT with respect to recall ("Win Rate" in Table 1). We also compute p-values testing the null hypothesis that the means of the recalls from unsupervised KRISS-Search and PubMedBERT are the same using a two-sample t-test ("P-Val" in Table 1).

Additionally, we compute AUPRC values across the 255 test query spans for both the *same* and *related* experiments ($\overline{\text{AUPRC}}$ in Table 2). As with the recall measures, we compute average per-query percent AUPRC improvement of KRISS-Search compared to PubMedBERT ($\overline{\%\Delta}$ in Table 3), the frequency with which unsupervised KRISS-Search outperforms PubMedBERT with respect to AUPRC ("Win Rate" in Table 3), and p-values testing the null hypothesis that the means of the AUPRCs from unsupervised KRISS-Search and PubMedBERT are the same using a two-sample t-test ("P-Val" in Table 3).

2.3 Supervised KRISS-Search

To incorporate human feedback, we train a light-weight classifier with KRISSBERT embeddings as input. We cache the KRISSBERT embeddings to reduce the latency that would result from fine-tuning KRISSBERT and embedding the corpus at each active learning iteration. Our active learning strategy is as follows. First, the user selects a small number of positive and negative seed examples. We then train the light-weight classifier on these seed examples. Leveraging this trained model, we generate a small number of additional examples to be labeled and added to the training dataset. We then retrain the classifier from scratch, repeating this procedure until the label quality appears satisfactory.

Concept Retrieval: To measure the performance of supervised KRISS-Search in retrieving specific concept mentions, we use same 2019 n2c2 entity linking dataset that was used to evaluate unsupervised KRISS-Search. We simulate human feedback with the ground truth labels. We adopt a least confidence (LC in Table 5 and Table 4) active learning strategy where we return examples closest to the decision boundary for labeling. Furthermore, we use a logistic regression linear probe as the classifier, 5 active learning iterations, 15 seed examples, and 15 labeled examples per active learning iteration. Furthermore, we hypothesize that the contrastive loss makes distance to positively labeled examples a useful feature. Thus, we append the square of the L2 distance from the mean of the positively labeled embeddings to the KRISSBERT embeddings as an additional input feature, which we refer to as sum of squares (SS in Table 5 and Table 4). For these experiments, we use 28 concepts with greater than 100 mentions and corresponding embeddings, as we require additional spans for active learning. For evaluation, we compute performance on retrieving all ground truth mentions in the corpus.

Few-Shot Biomedical Named Entity Recognition: We evaluate our method on the BLURB NER datasets (Gu et al., 2020) to ground our method in benchmarked tasks and demonstrate the flexibility of our method to handle various notions of similarity. Here, we adopt strict evaluation as is conventional in NER and to be consistent with previous work evaluated on these tasks. We hypothesize that mean pooling aggregation does not sufficiently represent span boundaries, as it discards spatial information about span embeddings. Thus, we concatenate the first token embedding with the last token embedding and append the length of the span. To provide a fair comparison between the traditional NER approaches and KRISS-Search, we equalize the number of labeled words used for training. We empirically choose the total number of labeled words to be equal to the number of words in 75 randomly sampled sentences that are used for BERT and PubMedBERT training. For all methods, we use the same single layer perceptron as the light-weight classifier. During BERT and PubMedBERT training, we save training checkpoints, and for testing, we choose checkpoints with the best performance on the validation sets. We forgo this approach with KRISS-Search, as we assume that the user has not labeled validation sets. We report results (Table 4) using the random sampling baseline (RSB), least confidence active learning (LC), and a spatial refinement strategy (SpR).

2.4 KRISS-Search with Spatial Refinement

Supervised KRISS-Search is different from standard active learning tasks in that the examples (spans) are not independent, rather they have spatial relationships. Specifically, since one span can overlap with other spans in the sample set, we apply the following spatial refinement (SpR in Table 5, Table 4, and Fig. 7) strategies for KRISS-Search:

- When the span presented to the user overlaps with a true positive span, the user can modify the boundaries and label the correct span (Fig. 3).
- In NER tasks aiming for exact span recovery, only one span from an overlapping group of spans can be correct, in which case we predict only the span with highest probability and mark all the other spans as negative.

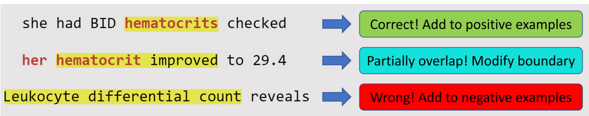


Figure 3: An example human feedback interface in Supervised KRISS-Search with spatial refinement (SpR). Yellow highlighting depicts spans presented to the user. Red bold letters are ground truth positive spans. For any recommended span, the user provides feedback by choosing from the following options: 1. mark the span as exactly correct (green button); 2. refine the boundaries of the span if it overlaps with a true span (cyan button); 3. mark the span as wrong (red button).

3 Results

3.1 Unsupervised KRISS-Search

Fig. 4 demonstrates a performant example on a test query for the "prothrombin time assay" CUI. Here, we show recall for unsupervised KRISS-Search (blue), PubMedBERT (red), and BERT (green) vs. the number of nearest neighbors for the *same* evaluation type. For this example, unsupervised KRISS-Search has an edge in terms of recall and thus precision, requiring fewer nearest neighbors to retrieve a similar number of positive spans. Fig. 8 in A demonstrates a similar outcome for this example using the *related* evaluation type.

Fig. 5 shows the mean recall, precision, and F1 across the 255 test query spans for the *same* evaluation type. Across the 255 corresponding concepts, an average (standard deviation) of 47% (16%) of mentions are unique. We observe that on average, unsupervised KRISS-Search has an edge over both PubMedBERT and BERT in terms of recall, precision, and F1. The error bars indicate $\pm$ 1 standard deviation. These error bars are large as the performance across CUIs varies.

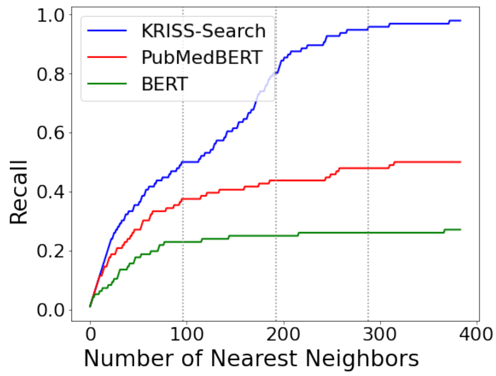


Figure 4: Recall using the *same* evaluation type (CUIs associated with returned spans must match the query associated CUI exactly). Query span is "PT", corresponding to the concept "prothrombin time assay". The vertical dotted lines indicate $1 \times N$, $2 \times N$, and $3 \times N$.

As in Fig. 5 with the *same* evaluation type, Fig. 10 in A aggregates the results across 255 test query spans for the *related* evaluation type. Overall, it appears that the benefit of unsupervised KRISS-Search over PubMedBERT and BERT is still substantial when we make the evaluation less rigid and allow for more diverse spans.

In, Table 1 we compare the aggregate performance of unsupervised KRISS-Search and PubMedBERT. $\overline{\%\Delta}$ indicates that the average per-query percent improvement of unsupervised KRISS-Search over PubMedBERT is substantial. Furthermore, the win rates indicate that unsupervised KRISS-Search does better than PubMedBERT across most of the test queries. The P-values indicate that for number of nearest neighbors equals $1 \times N$, $2 \times N$, $3 \times N$, and both evaluation types, the benefit of unsupervised KRISS-Search over PubMedBERT is statistically significant.

Fig. 6 shows the precision-recall curves for the same performant prothrombin time assay example previously evaluated using the *same* evaluation type. We note that for this example, the benefit of unsupervised KRISS-Search (AUPRC = 0.60) over both PubMedBERT (AUPRC = 0.31) and BERT (AUPRC = 0.11) is substantial. Fig. 9 in A shows similarly beneficial results for the *related* evaluation type.

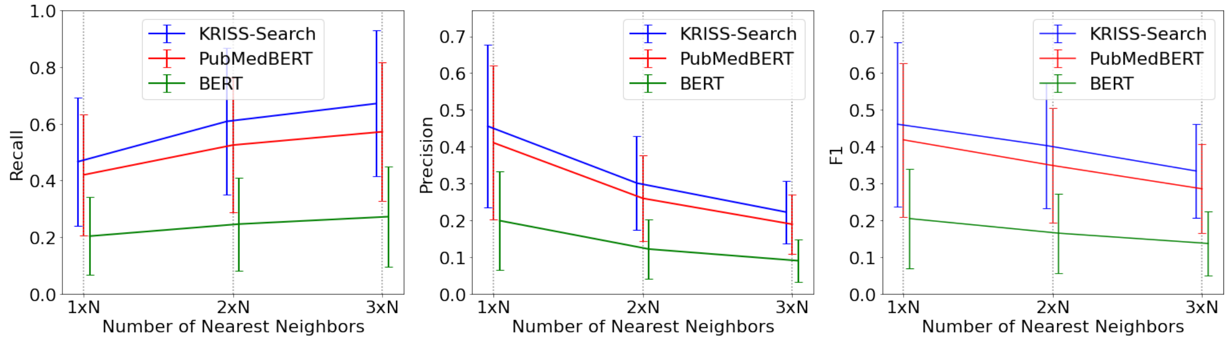


Figure 5: Mean recall (left), precision (center), and F1 (right) of unsupervised KRISS-Search (blue), PubMedBERT (red), and BERT (green) in retrieving concepts with the same CUI for number of nearest neighbors equals $1 \times N$, $2 \times N$, and $3 \times N$ across 255 test query spans. The error bars indicate ± 1 standard deviation. The three plots are staggered slightly to make the errors bars more visible.

Table 1: Comparison of unsupervised KRISS-Search and PubMedBERT with respect to recall across 255 test query spans. #NN refers to the number of nearest neighbors.

Eval Type	#NN	$\overline{\%\Delta}$	Win Rate	P-Val
Same	1xN	+ 24%	0.61	4.2e-3
	2xN	+ 29%	0.69	2.4e-5
	3xN	+ 31%	0.73	7.0e-7
Related	1xN	+ 26%	0.69	3.0e-4
	2xN	+ 35%	0.73	1.7e-6
	3xN	+ 35%	0.75	4.0e-7

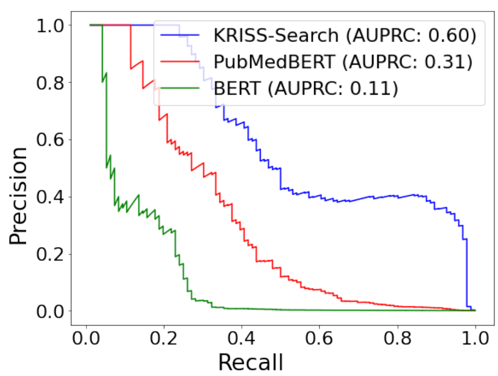


Figure 6: Precision-recall curves using the *same* evaluation type on an example query span with the text “PT”, corresponding to the concept “prothrombin time assay”.

Table 2: Average AUPRC scores from unsupervised KRISS-Search, PubMedBERT, and BERT across 255 test query spans. Results are presented as mean $\pm$ 1 standard deviation

Eval Type	Model	$\overline{\text{AUPRC}}$
Same	BERT	0.14 ± 0.12
	PubMedBERT	0.37 ± 0.23
	KRISS-Search	0.43 ± 0.25
Related	BERT	0.10 ± 0.09
	PubMedBERT	0.26 ± 0.19
	KRISS-Search	0.33 ± 0.23

Table 3: AUPRC comparison of unsupervised KRISS-Search and PubMedBERT.

Eval Type	$\overline{\%\Delta}$	Win Rate	P-Val
Same	+ 51%	0.71	4.5E-03
Related	+ 54%	0.76	1.6E-04

From Table 2 we observe that unsupervised KRISS-Search statistically significantly outperforms PubMedBERT ("P-Values" in Table 3). Although the average AUPRC decreases when moving from the *same* to the *related* evaluation type (as seen in Table 2), the average percentage change (as represented by $\overline{\%\Delta}$) increases (as seen in Table 3). This suggests that KRISS-Search's performance does not decline as steeply when transitioning from the *same* to the *related* evaluation type. We further assess whether this advantage persists when considering only unique mentions as positive spans. Utilizing the *same* evaluation type, we observe average AUPRCs of 0.24 ± 0.14, 0.20 ± 0.13, and 0.14 ± 0.11 for KRISS-Search, PubMedBERT, and BERT, respectively.

3.2 Supervised KRISS-Search

In concept retrieval on the n2c2 dataset, we outperform PubMedBERT and achieve an average F1 score of 0.68 ± 0.14 (using least confidence active learning, the sum of squares feature, and spatial refinement).

Table 5 shows an ablation study which demonstrates the utility of least confidence active learning (LC vs. the random sampling baseline), the sum of squares feature (SS), and spatial refinement (SpR). Furthermore, for the most performant configurations, the KRISS-Search embeddings outperform the PubMedBERT embeddings.

Fig. 7 shows the concept retrieval performance curves for an example "White Blood Count" span. This figure illustrates that as the supervised KRISS-Search iterations progress, incorporating human feedback consistently enhances the model's F1 performance. Furthermore, utilization of least confidence sampling (LC), sum of squares feature (SS) and spatial refinement (SpR) techniques results in less recall degradation while achieving the highest F1 score performance.

Table 4 shows that our method significantly outperforms BERT and also performs comparably to or outperforms PubMedBERT by an average of 5.6 F1 points. This is significant given that our method was not designed for NER. Our performance here indicates that supervised KRISS-Search can generalize to coarse-grained biomedical concepts and strict evaluation.

4 Conclusion

We demonstrate that unsupervised KRISS-Search outperforms existing embedding methods for biomedical interactive span recommendation. Supervised KRISS-Search utilizes humans-in-the-loop to achieve high levels of performance on both granular and course grain span recommendation. Future work will investigate whether KRISS-Search does indeed address the initial motivation - aiding programmatic data labeling as part of an interactive biomedical NLP system. Nonetheless, we envision KRISS-Search being broadly useful as a general purpose interactive biomedical search engine.

5 Limitations

One drawback of our method is that given a maximum span length, we always miss longer spans. For example, the BC2GM and JNLPBA NER datasets contain lengthy spans so we do not do as well on those tasks. Another drawback of our method is that it requires embedding the full corpus. One of our methods for making this tractable introduces another limitation - span filtering based on token types may discard spans that are useful to the user. Additionally, although we demonstrate that our method can be robust to training time (A.1), we have not explored principled methods for selecting the model checkpoint in supervised KRISS-Search, as the user does not label a validation set. Methods for making the process more rigorous should be explored, especially for out of distribution tasks.

6 Ethics Statement

The authors have evaluated the potential consequences of their research, including both positive and negative effects. Furthermore, the authors have ensured compliance with the guidelines outlined in the ACM Code of Ethics and Professional Conduct, and confirm that this work is in accordance with those principles.

References

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database-Issue):267–270.

Benedikt Boecking, Willie Neiswanger, Eric Xing, and Artur Dubrawski. 2021. Interactive weak supervision: Learning useful heuristics for data labeling. In *International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing. *CoRR*, abs/2007.15779.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Hunter Lang and Hoifung Poon. 2020. Self-supervised self-supervision by combining deep learning and probabilistic logic. *CoRR*, abs/2012.12474.

Jiacheng Li, Haibo Ding, Jingbo Shang, Julian McAuley, and Zhe Feng. 2021. Weakly supervised named entity tagging with learnable logical rules.

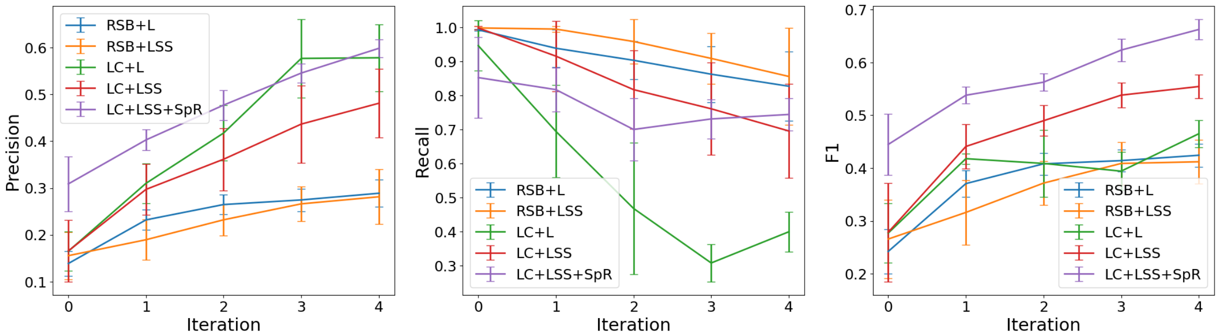


Figure 7: "White Blood Count" concept retrieval example across active learning iterations: precision (left), recall (middle), and F1 score (right). L stands for linear and denotes using the span embeddings without any additional features. LSS represents an additional sum of squares feature, which is the squared distance to the mean of positively labeled embeddings. LC denotes least confidence active learning, while RSB is the random sampling baseline. SpR represents spatial refinement human feedback. Values are mean ± 1 standard deviation across 3 active learning runs.

Table 4: Few-Shot biomedical NER results. For the KRISS-Search methods, the reported values represent the mean of 3 runs using different random seed example for active learning. PMB refers to PubMedBERT and KS refers to KRISS-Search. RSB, LC, and SpR refer to random sampling baseline, least confidence active learning, and spatial refinement respectively.

Dataset	BERT	PMB	KS (RSB)	KS (LC)	KS (SpR)
BC5-chem	0.69	0.73	0.67	0.82	0.84
BC5-disease	0.49	0.60	0.52	0.71	0.74
NCBI-disease	0.55	0.63	0.45	0.65	0.70
BC2GM	0.48	0.54	0.31	0.49	0.56
JNLPBA	0.55	0.59	0.37	0.47	0.53

Table 5: Ablation study with PubMedBERT and KRISS-Search on the concept retrieval task. The table compares least confidence active learning (LC) versus the random sampling baseline. It also measures how the sum of squares feature (SS), which denotes the squared distance to mean of the positively labeled embeddings, impacts performance. Furthermore, it measures the impact of spatial refinement human feedback (SpR). Values represent mean ± 1 standard deviation.

LC	SS	SpR	PubMedBERT	Kriss-Search
✓	✓	✓	0.66 ± 0.13	0.68 ± 0.14
✓	✓		0.61 ± 0.13	0.63 ± 0.13
✓		✓	0.57 ± 0.14	0.58 ± 0.14
✓			0.56 ± 0.14	0.56 ± 0.15
	✓	✓	0.34 ± 0.12	0.41 ± 0.16
	✓		0.33 ± 0.11	0.35 ± 0.11
		✓	0.33 ± 0.12	0.31 ± 0.12
			0.39 ± 0.12	0.37 ± 0.12

S. Liu, Wei Ma, R. Moore, V. Ganesan, and S. Nelson. 2005. Rxnorm: prescription for electronic drug information exchange. *IT Professional*, 7(5):17–23.

Yen-Fu Luo, Sam Henry, Yanshan Wang, Feichen Shen, Ozlem Uzuner, and Anna Rumshisky. 2020. The 2019 n2c2/UMass Lowell shared task on clinical concept normalization. *Journal of the American Medical Informatics Association*, 27(10):1529–e1.

Usman Naseem, Matloob Khushi, Shah Khalid Khan, Kamran Shaukat, and Mohammad Ali Moni. 2021. A comparative analysis of active learning for biomedical text mining. *Applied System Innovation*, 4(1).

Hoifung Poon, Hai Wang, and Hunter Lang. 2021. Combining probabilistic logic and deep learning for self-supervised learning. *CoRR*, abs/2107.12591.

Alexander Ratner, Stephen H. Bach, Henry R. Ehrenberg, Jason Alan Fries, Sen Wu, and Christopher Ré. 2017a. Snorkel: Rapid training data creation with weak supervision. *CoRR*, abs/1711.10160.

Alexander J. Ratner, Stephen H. Bach, Henry R. Ehrenberg, and Chris Ré. 2017b. Snorkel: Fast training set generation for information extraction. In *Proceedings of the 2017 ACM International Conference on Management of Data*, SIGMOD '17, page 1683–1686, New York, NY, USA. Association for Computing Machinery.

Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data programming: Creating large training sets, quickly. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Xiaojiang Chen, and Xin Wang. 2020. A survey of deep active learning. *CoRR*, abs/2009.00236.

Kent A. Spackman, Ph. D, Keith E. Campbell, Ph. D, Roger A. Côté, and D. Sc. (hon. 1997. Snomed rt: A reference terminology for health care. In *J. of the American Medical Informatics Association*, pages 640–644.

Artur Yakimovich, Anaël Beaugnon, Yi Huang, and Elif Ozkirimli. 2021. Labels in a haystack: Approaches beyond supervised learning in biomedical applications. *Patterns*, 2(12):100383.

Jieyu Zhang, Cheng-Yu Hsieh, Yue Yu, Chao Zhang, and Alexander Ratner. 2022a. A survey on programmatic weak supervision.

Sheng Zhang, Hao Cheng, Shikhar Vashishth, Cliff Wong, Jinfeng Xiao, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022b. Knowledge-rich self-supervision for biomedical entity linking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 868–880, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xinyan Zhao, Haibo Ding, and Zhe Feng. 2021. Glara: Graph-based labeling rule augmentation for weakly supervised named entity recognition. *CoRR*, abs/2104.06230.

A Appendix

A.1 KRISS-Search Hyperparameters

For the conventional NER methods, we choose the following hyperparameters based on performance on the validation set. For KRISS-Search, the task of choosing hyperparameters for the single layer perceptron is less straightforward as we would like our method to generalize to settings where we do not have a labeled validation set for hyperparameter tuning. We hypothesize that we can include an L2 regularization term and train for many epochs without overfitting, eliminating the need for selecting a precise number of training iterations. We thus increase the default regularization coefficient from the scikit-learn MLP classifier default value of 1e-4 to 1e-3. Furthermore, we choose the Adam optimizer, hypothesizing that it is less sensitive than other optimization methods to initial learning rate. We selected an initial learning rate of 1e-4, a train batch size of 64, and 200 training iterations based on our hypothesis that these hyperparameters would result in training that is not sensitive to the number of training iterations. To validate this hypothesis, we also conducted additional experiments with only 100 training iterations, and found that the performance differences between the two sets of experiments were negligible. This suggests that our chosen hyperparameters are indeed robust and do not greatly affect the outcome of the training.

A.2 Recall of "PT" example using related evaluation type

Fig. 8 shows results for the same prothrombin time assay CUI example as was used in Fig. 4 but with the *related* evaluation type. We note here that the number of nearest neighbors corresponding to $1 \times N$, $2 \times N$, and $3 \times N$ is greater as expected.

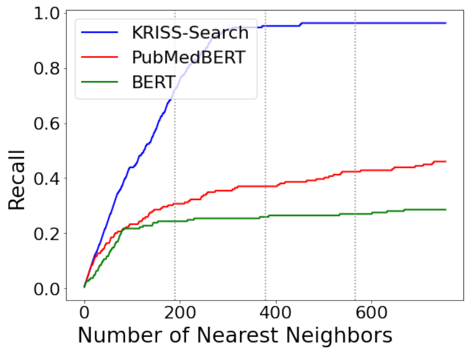


Figure 8: Recall using the *related* evaluation type on an example query span with the text "PT", corresponding to the concept "prothrombin time assay".

A.3 Aggregate recall, precision, and F1 using related evaluation type

Fig. 10 aggregates the results across 255 test query spans for the *related* evaluation type. The benefit of unsupervised KRISS-Search over PubMedBERT and BERT is substantial when we make the evaluation less rigid and allow for more diverse spans as compared to the *same* evaluation type.

A.4 AUPRC of "PT" example using related evaluation type

Fig. 9 shows the precision-recall curves for the prothrombin time assay example using the *related* evaluation type. As with the *same* evaluation type, the benefit of unsupervised KRISS-Search (AUPRC = 0.77) over both PubMedBERT (AUPRC = 0.26) and BERT (AUPRC = 0.13) is significant.

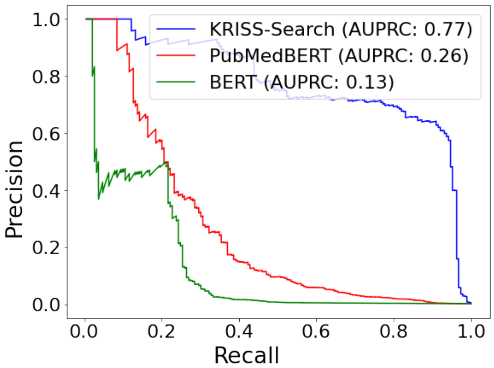


Figure 9: Precision-recall curves using the *related* evaluation type on an example query span with the text “PT”, corresponding to the concept “prothrombin time assay”.

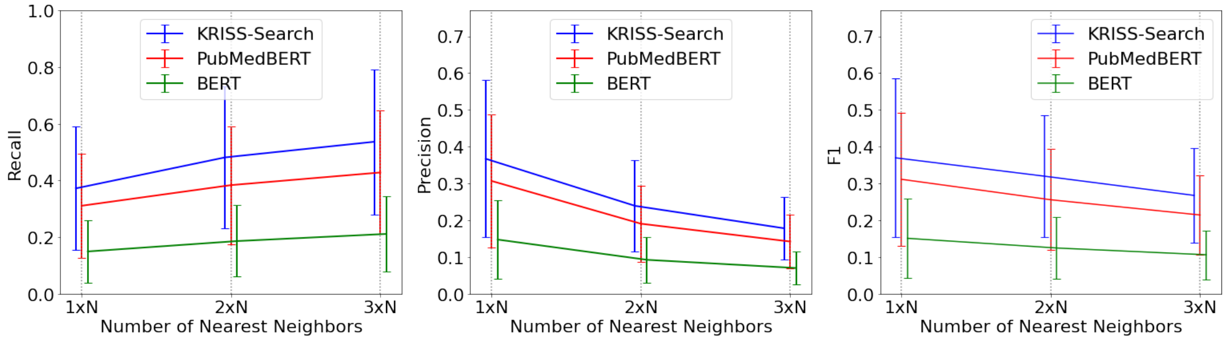


Figure 10: Aggregate recall (left), precision (center), and F1 (right) of unsupervised KRISS-Search (blue), PubMedBert (red), and BERT (green) in retrieving concepts with related CUIs for number of nearest neighbors equals $1 \times N$, $2 \times N$, and $3 \times N$ across 255 test query spans.

Prompt-based Extraction of Social Determinants of Health Using Few-shot Learning

Giridhar Kaushik Ramachandran[1]*, Yujuan Fu[2]*, Bin Han[3]*
Kevin Lybarger[1], Nicholas J Dobbins[2], Özlem Uzuner[1], Meliha Yetisgen[2]
[1] Department of Information Sciences and Technology, George Mason University, Fairfax, VA, USA
[2] Department of Biomedical Informatics & Medical Education, University of Washington, Seattle, WA, USA
[3] Information School, University of Washington, Seattle, WA, USA

Abstract

Social determinants of health (SDOH) documented in the electronic health record through unstructured text are increasingly being studied to understand how SDOH impacts patient health outcomes. In this work, we utilize the Social History Annotation Corpus (SHAC), a multi-institutional corpus of de-identified social history sections annotated for SDOH, including substance use, employment, and living status information. We explore the automatic extraction of SDOH information with SHAC in both standoff and inline annotation formats using GPT-4 in a one-shot prompting setting. We compare GPT-4 extraction performance with a high-performing supervised approach and perform thorough error analyses. Our prompt-based GPT-4 method achieved an overall 0.652 F1 on the SHAC test set, similar to the 7th best-performing system among all teams in the n2c2 challenge with SHAC.

1 Introduction and related work

Social determinants of health (SDOH) are the conditions in which people work and live that impact quality of life and health (Centers for Disease Control and Prevention, 2022). Understanding SDOH can assist in clinical decision-making (Daniel et al., 2018; Friedman and Banegas, 2018). SDOH is documented in the electronic health record (EHR) through unstructured clinical narratives and structured data; however, the clinical narrative includes a more detailed description of many SDOH events. To utilize the text-encoded SDOH information in secondary use applications, including clinical decision-support systems, the SDOH information must be automatically extracted (Daniel et al., 2018; Singh et al., 2017).

SDOH extraction has been explored using rule-based systems and data-driven models that use supervised learning (Hatef et al., 2019; Patra et al., 2021; Yu et al., 2022; Han et al., 2022) on a variety of corpora (Uzuner et al., 2008; Stemerman et al., 2021; Yetisgen and Vanderwende, 2017). Recent SDOH extraction work utilizes large language models (LLMs) like BERT and T5, where models are fine-tuned to the SDOH extraction task (Lybarger et al., 2022; Romanowski et al., 2023). Recent advancements in LLMs, including larger models like Generative Pretrained Transformer (GPT)-based models (Brown et al., 2020; OpenAI, 2023), allow for new training paradigms, including few-shot or zero-shot learning. Recent developments in LLMs like the GPT-4 (OpenAI, 2023) and med-PaLM models have shown their capability to understand the clinical text and achieve/exceed human-level performance in US medical licensing exams (Singhal et al., 2022). This high performance may be attributed to (1) high model parameter counts, (2) large pre-training datasets, and (3) instruction tuning and optimization with Reinforcement Learning Human Feedback (RLHF) (Ouyang et al., 2022). Recent clinical information extraction (IE) work (Liu et al., 2023; Hu et al., 2023) comparing BERT-based fine-tuning approaches to zero-shot learning indicates GPT models can extract entities and relations with reasonable performance; however, there are many open questions related to the use of recent LLMs, like GPT-4, in clinical IE tasks.

In this work, we explore the extraction of SDOH using GPT-4 in a one-shot prompting setting with event-based SHAC (Lybarger et al., 2021). We compare prompt-based extraction approaches with a high-performing supervised BERT-based model(Lybarger et al., 2022) that has been fine-tuned to SDOH extraction from SHAC. We investigate two different one-shot prompting strategies for GPT-4, including prompts aimed at generating BRAT standoff format and inline annotations. We report an overall performance of 0.861 F1 from the fine-tuned model, evaluated on the withheld test set. The highest-performing one-shot GPT-4 approach

* Equal contribution

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 385–393
July 14, 2023 ©2023 Association for Computational Linguistics

achieved an overall 0.652 F1 for SDOH event extraction. Our initial study shows that GPT-4 can extract SDOH information from text with limited training examples.

2 Data, Task, & Evaluation

The 2022 National NLP Clinical Challenges SDOH extraction task (n2c2/UW SDOH Challenge) used SHAC for model development and evaluation (Lybarger et al., 2023). SHAC contains 4405 deidentified social history sections of notes from MIMIC-III (Johnson et al., 2016) and the University of Washington (UW). SHAC includes training, development, and test partitions for both sources (MIMIC-III and UW). SHAC was annotated using BRAT (Stenetorp et al., 2012), a web-based annotation tool, to capture five SDOH event types: substance use (*Alcohol, Drug, Tobacco*), employment status (*Employment*), and living status (*LivingStatus*). Figure 1 (A) presents an annotated sample in BRAT from the SHAC UW training set.

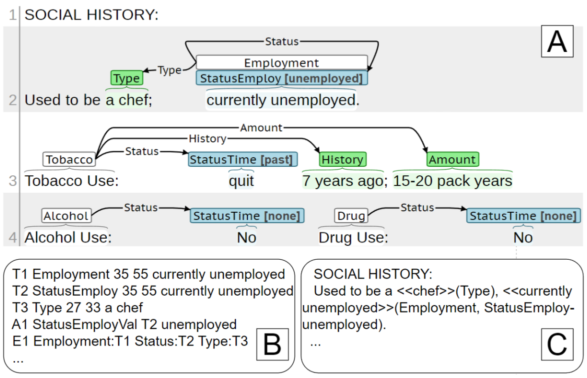


Figure 1: A. Sample note with SDOH events, visualized in the BRAT website. B. Standoff annotations in the BRAT format (.ann). C. Inline annotations.

The n2c2/UW SDOH Challenge evaluation criteria interpret the extraction as a slot-filling task (Lybarger et al., 2023). Each event comprises a single trigger span and at least one required argument. *Trigger any overlap* equivalence requires the predicted trigger to overlap with the true trigger of the same event type. Arguments can be classified into two categories: *span-only* (a multi-word span and argument type) and *labeled* (a multi-word span, argument type, and *subtype* label). Arguments can be equivalent only when attached to equivalent triggers. In addition to trigger *any overlap* equivalence, *span-only* argument equivalence is evaluated by *exact match*, and *labeled* arguments equivalence requires the correct argument and subtype labels (*span agnostic*) (Lybarger et al., 2023).

We evaluated performance using the n2c2/UW SDOH Challenge criteria, as well as on more lenient evaluation criteria that still assess the clinical meaning of extraction. In the lenient criteria, trigger equivalence is relaxed to a minimum-distance metric (*minimum distance*), where gold triggers are paired (aligned) with the closest predicted trigger of the same event type, and the closest predicted trigger is counted as a true positive. In the lenient criteria, the *span-only* arguments use the *any overlap* criteria and the *labeled* arguments are evaluated as previously described.

3 Methods

We benchmark the SDOH extraction task using two methods: (1) multi-label variation of the Span-based Entity and Relation Transformer (SpERT)(Eberts and Ulges, 2020) architecture, mSpERT (Lybarger et al., 2022) benchmarked for SHAC as a high-performing fine-tuned baseline, and (2) prompt-based one-shot learning with GPT-4. Inspired by performance gains of few-shot learning, relative to zero-shot learning in prior work (Brown et al., 2020; Liévin et al., 2022), we use one-shot prompting with GPT-4 for the SDOH extraction task in this short study. We experiment with two distinct output formats - (1) BRAT-style standoff annotations (GPT-standoff) and (2) Inline annotations (GPT-inline).

We conduct the GPT-4 one-shot experiments through OpenAI's GPT-4 Chat Completion Application Programming Interface (API)[1], because of GPT-4's proprietary nature and significant hardware requirements. The API allows users to provide instructions via three role variables. Our prompts are structured in the following order:

1. ***system***: defines the desired role, personality traits, and task instructions for GPT-4. We use the *system* variable to assign GPT-4 the role of an annotator along with the paraphrased annotation guideline.
2. ***user***: provides an example note for one-shot learning.
3. ***assistant***: provides the gold annotations for the example note in *user*. This is an example prediction for one-shot learning.

Following the above definitions, we end with a *user* message containing a note to be annotated and

[1] https://platform.openai.com/docs/api-reference/chat

indicate the *assistant* should respond. We randomly sample a note from a subset of the SHAC-UW training set containing all five SDOH event types. Appendix A.2 contains formats and snippets of our prompts.

3.1 GPT-standoff

To assess GPT-4's ability to comprehend the task and generate structured outputs, we prompt the model to generate predictions in the BRAT standoff format (Stenetorp et al., 2012) used by SHAC. The BRAT standoff format includes pairs of text (*.txt) and annotation (*.ann) files. The event annotation is characterized by three BRAT annotation frames in the annotation file: (1) *Text bounds (T)* include a text span (e.g. "currently unemployed"), span label (e.g. 'Employment'), and character indices (e.g. "35 55") for marking both triggers and arguments; (2) *Attributes (A)* adds a subtype label to T, and (3) *Events (E)* characterize an SDOH event through linking a trigger and at least one argument. A visual representation is provided in Figure 1 (B).

We provide the paraphrased annotation guideline in the *system* and the note via *user* variables and elicit GPT-4 to output the desired annotation file through the *assistant* variable. Our preliminary experimentation indicated that though GPT-4 was able to correctly extract relevant text spans, it had some shortcomings: (1) some generated lines did not conform to the BRAT standoff format, and (2) the generated character indices did not correspond with the identified text spans. We post-processed the generated outputs to ensure compliance with the BRAT standoff format and updated the character indices to correspond with the first occurrence of the generated text span (<3% spans occur more than once). Please refer to Appendix A.3 for original and post-processed examples.

3.2 GPT-inline

Prior fine-tuned IE work utilized inline markers to infuse entity information in the body of narratives (Romanowski et al., 2023; Phan et al., 2021). In our work, we instruct GPT-4 to generate a version of the note with inline markers that identify the SDOH triggers and arguments. These markers encode all spans inside double-angle brackets ("« »"), with trigger, argument, and subtype labels appended (Figure 1, (C)). Similar to the GPT-standoff model, the GPT-inline model elicits GPT-4 the desired annotation format through the *assistant* variable. This method does not prompt the model to make trigger-argument connections, and we use a heuristic search to associate each argument with the nearest event trigger, constrained by the allowable trigger-argument connections defined by the annotation guideline (See details in the Appendix. A.4). The GPT-inline output is post-processed into BRAT standoff format for evaluation.

4 Results and Discussion

Table 1 contains the overall performance of the prompt-based GPT-4 models on the withheld SHAC-UW test set, which was the evaluation data for Subtask C of n2c2/UW SDOH Challenge (Lybarger et al., 2023). The GPT-standoff and GPT-inline models achieve an overall F1 of 0.625 and 0.652, respectively. This performance is much lower than the mSpERT model and the highest-performing n2c2 systems, which utilized the entire training set in supervised model fine-tuning. GPT-inline achieved performance similar to the 7th best n2c2 system from IBM, which utilized BERT (Lybarger et al., 2023). This one-shot performance indicates that the natural language understanding capabilities of GPT-4 allow the prompt-based methods to leverage the annotation schema and achieve moderate performance. The results suggest that fine-tuning at least a portion of the training set may be needed to achieve high performance. We also observed that a generative architecture could lead to a new set of errors, and some of them may be eliminated in post-processing.

Method	P	R	F1
Fine-tuned			
Microsoft (T5)	0.891	0.887	0.889
CHOP (BERT)	0.874	0.888	0.881
mSpERT	0.868	0.854	0.861
...	...	...	...
IBM (BERT)	0.538	0.788	0.640
GPT-4 one-shot + post-processing			
GPT-standoff	0.621	0.628	0.625
GPT-inline	0.650	0.654	0.652

Table 1: Comparison of overall micro-averaged performance of SDOH triggers and arguments between select top-performing models in the n2c2 challenge, mSpERT (fine-tuned baseline), and the GPT-4 models.

Table 2 contains trigger and argument micro-averaged F1 scores for each event type using the n2c2 and lenient evaluation criteria. Comparing overall performances (last row in the table), mSpERT outperforms both our GPT-4 models. But the performance gap between the fine-tuned model and the one-shot GPT models is smaller from the

Field	Argument	# True Labels	n2c2 Evaluation (F1)			Lenient Evaluation (F1)		
			mSpERT	GPT-standoff	GPT-inline	mSpERT	GPT-standoff	GPT-inline
Trigger								
Alcohol	-	403	0.964	0.861	0.938*	0.967	0.972*	0.952
Drug	-	473	0.929	0.824	0.861*	0.942	0.935*	0.898
Tobacco	-	434	0.963	0.825	0.917*	0.970	0.965*	0.939
Employment	-	153	0.908	0.803*	0.709	0.915	0.921*	0.766
LivingStatus	-	354	0.886	0.590	0.749*	0.903	0.844*	0.811
Labeled Argument								
Alcohol	StatusTime	403	0.913	0.763	0.734	0.913	0.856*	0.750
Drug	StatusTime	473	0.857	0.706*	0.646	0.868	0.783*	0.673
Tobacco	StatusTime	434	0.917	0.694	0.738*	0.926	0.813*	0.764
Employment	StatusEmploy	153	0.868	0.657	0.627	0.875	0.759*	0.679
LivingStatus	StatusTime	354	0.833	0.572	0.709*	0.850	0.787*	0.760
	TypeLiving	354	0.871	0.560	0.725*	0.891	0.770*	0.774
Span-only Argument								
Alcohol	All types	178	0.699	0.388*	0.172	0.783	0.694*	0.354
Drug	All types	418	0.625	0.219*	0.104	0.688	0.426	0.381
Tobacco	All types	375	0.775	0.420*	0.322	0.830	0.714*	0.537
Employment	Duration, History, Type	96	0.675	0.169	0.109	0.735	0.677*	0.500
LivingStatus	Duration, History	11	0.421	0.063	0.074	0.526	0.159	0.105
	Overall	5066	0.861	0.625	0.652*	0.882	0.791*	0.728

Table 2: Micro-averaged F1 comparison between mSpERT and GPT-4 one-shot models. SpERT outperforms the GPT-4 one-shot models in all trigger and argument extraction (with the exception of *Alcohol* and *Employment* triggers). For better readability, we only mark performance significance among GPT-4 one-shot models. * indicates performance significance among GPT-4 one-shot methods, with 10,000 bootstrap samples and a p-value threshold of 0.05.

n2c2 evaluation to the lenient evaluation, which can largely be attributed to higher trigger extraction performance, as argument equivalence requires trigger equivalence in both evaluations. The GPT-standoff model can identify the presence of SDOH events, but the identified triggers may not overlap with the gold trigger. The lenient evaluation only requires the same trigger type present in the social history text. The relatively lower performance for the GPT-inline model in *Employment* trigger extraction can be attributed to the model frequently identifying an *StatusEmploy* labeled argument without predicting an *Employment* trigger. The GPT-inline extractions may capture meaningful employment information but do not adhere to the annotation guidelines. For both GPT models, the extraction of *LivingStatus* triggers is relatively more challenging. Although the notes contain many plausible candidate spans for *LivingStatus* triggers, these spans were not annotated in SHAC since they did not contain information to resolve the associated *TypeLiving* labeled argument. The GPT models capture these false positive *LivingStatus* triggers often without *TypeLiving* labeled arguments. For argument extraction, GPT-standoff significantly outperforms GPT-inline in four arguments under the n2c2 evaluation and eight arguments in lenient evaluation. The GPT-inline does not link annotated arguments to triggers, and a distance metric (character count) is used to link them, which contributes to the GPT-in-line's relatively lower performance. The labeled arguments are required for each event, and we observe that labeled argument performance is 0.1 F1 lower than the corresponding trigger performance. When multiple substance events are present in a note with differing *StatusTime* labels (e.g. current and past), we observe the GPT-standoff model tends to output the same *StatusTime* label for all substance events. The GPT-inline model correctly captures both substance trigger and *StatusTime* spans with the correct label but fails to correctly link the triggers with the right *StatusTime* spans because multiple *StatusTime* spans can have the same distance to a trigger span.

5 Conclusions

We investigate the efficacy of two prompt-based approaches for extracting SDOH from social history sections using GPT-4. Although the supervised model achieves higher performance, our findings indicate that GPT-4's one-shot learning capabilities serve as a promising starting point for extracting SDOH events without the need for annotated data. Possible gains in future work may be achieved with a combination of few-shot and active learning.

6 Limitations

We only explored one-shot prompting strategies with GPT-4. More examples (few-shot) may improve performance. We prompted GPT-4 with only a single randomly selected sample that included all of the annotated event types. Our post-processing included simple rules to process the generated output and may be improved. The quality of the sample and the selection method may influence performance. We explored two prompting styles. Future work could explore more prompting methods such as question & answering and chain-of-thought (Wei et al., 2023) and fine-tuning non-proprietary LLMs.

7 Ethics statement

Our experimentation utilized OpenAI API to extract SDOH information from SHAC with GPT-4. SHAC is a fully de-identified corpus of social history sections. The use of such external API/models could introduce ethical problems related to privacy, identifiability, and other unintended consequences if the data sets are not fully de-identified. Additionally, a careful examination is needed to assess potential bias in LLMs for extracting SDOH prior to implementing real-life secondary use applications. We received approval from the Institutional Review Board (IRB) prior to conducting the presented research. As our GPT-4 one-shot experiments are conducted on the SHAC-UW test set, broader use of the model may need necessary precautions.

8 Acknowledgements

This work was supported in part by the National Institutes of Health and the National Library of Medicine (NLM) (Grant numbers R01 CA248422-01A1, R15 LM013209). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Centers for Disease Control and Prevention. 2022. Social determinants of health at CDC.

Hilary Daniel, Sue S Bornstein, and Gregory C Kane. 2018. Addressing social determinants to improve patient care and promote health equity: An american college of physicians position paper. *Annals of internal medicine*, 168(8):577–578.

Markus Eberts and Adrian Ulges. 2020. Span-based joint entity and relation extraction with transformer pre-training. In *24th European Conference on Artificial Intelligence*.

Nicole L Friedman and Matthew P Banegas. 2018. Toward addressing social determinants of health: a health care system strategy. *The Permanente journal*, 22.

Sifei Han, Robert Zhang, Lingyun Shi, Russell Richie, Haixia Liu, Andrew Tseng, Wei Quan, Neal Ryan, David Brent, and Fuchiang (Rich) Tsui. 2022. Classifying social determinants of health from unstructured electronic health records using deep learning-based natural language processing. *Journal of Biomedical Informatics*, 127:103984.

Elham Hatef, Masoud Rouhizadeh, Iddrisu Tia, Elyse Lasser, Felicia Hill-Briggs, Jill Marsteller, and Hadi Kharrazi. 2019. Assessing the availability of data on social and behavioral determinants in structured and unstructured electronic health records: A retrospective analysis of a multilevel health care system. *Journal of Medical Internet Research*, 7.

Yan Hu, Iqra Ameer, Xu Zuo, Xueqing Peng, Yujia Zhou, Zehan Li, Yiming Li, Jianfu Li, Xiaoqian Jiang, and Hua Xu. 2023. Zero-shot Clinical Entity Recognition using ChatGPT. *arXiv preprint arXiv:2303.16416*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, et al. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035.

Valentin Liévin, Christoffer Egeberg Hother, and Ole Winther. 2022. Can large language models reason about medical questions? *arXiv preprint arXiv:2207.08143*.

Zhengliang Liu, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Wei Liu, Dinggang Shen, Quanzheng Li, et al. 2023. Deid-gpt: Zero-shot medical text de-identification by gpt-4. *arXiv preprint arXiv:2303.11032*.

Kevin Lybarger, Nicholas J Dobbins, Ritche Long, Angad Singh, Patrick Wedgeworth, Ozlem Uzuner, and Meliha Yetisgen. 2022. Leveraging Natural Language Processing to Augment Structured Social Determinants of Health Data in the Electronic Health Record. *arXiv preprint arXiv:2212.07538*.

Kevin Lybarger, Mari Ostendorf, and Meliha Yetisgen. 2021. Annotating social determinants of health using active learning, and characterizing determinants using neural event extraction. *Journal of biomedical informatics*, 113:103631.

Kevin Lybarger, Meliha Yetisgen, and Özlem Uzuner. 2023. The 2022 n2c2/uw shared task on extracting social determinants of health. *Journal of the American Medical Informatics Association*.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Braja Patra, Mohit Sharma, Veer Vekaria, Olga Patterson, Benjamin Glicksberg, Lauren Lepow, Euijung Ryu, Joanna Biernacka, Al'ona Furmanchuk, Thomas George, William Hogan, Yonghui Wu, Xi Yang, Bian Jiang, Myrna Weissman, Jaya Priya, J. Mann, Mark Olfson, Thomas Jr, and Jyotishman Pathak. 2021. Extracting social determinants of health from electronic health records using natural language processing: A systematic review. *Journal of the American Medical Informatics Association*, pages 1–12.

Long N. Phan, James T. Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. Scifive: a text-to-text transformer model for biomedical literature.

Brian Romanowski, Asma Ben Abacha, and Yadan Fan. 2023. Extracting social determinants of health from clinical note text with classification and sequence-to-sequence approaches. *Journal of the American Medical Informatics Association*. Ocad071.

Gopal K Singh, Gem P Daus, Michelle Allender, et al. 2017. Social determinants of health in the united states: addressing major health inequality trends for the nation, 1935-2016. *International journal of MCH and AIDS*, 6(2):139.

Karan Singhal, Shekoofeh Azizi, Tao Tu, et al. 2022. Large language models encode clinical knowledge.

Rachel Stemerman, Jaime Arguello, Jane Brice, Ashok Krishnamurthy, Mary Houston, and Rebecca Kitzmiller. 2021. Identification of social determinants of health using multi-label classification of electronic health record clinical notes. *American Medical Informatics Association, issuing body*, 4.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Ozlem Uzuner, Ira Goldstein, Yuan Luo, and Isaac Kohane. 2008. Identifying patient smoking status from medical discharge records. *American Medical Informatics Association*, 15:14–24.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Meliha Yetisgen and Lucy Vanderwende. 2017. Automatic identification of substance abuse from social history in clinical text. In , pages 171–181.

Zehao Yu, Xi Yang, Chong Dang, Songzi Wu, Prakash Adekkanattu, Jyotishman Pathak, Thomas George, William Hogan, Yi Guo, Jiang Bian, and Yonghui Wu. 2022. A study of social and behavioral determinants of health in lung cancer patients using transformers-based natural language processing models. *AMIA Annual Symposium proceedings. AMIA Symposium*, 2021:1225–1233.

A Appendix

A.1 The SHAC annotation schema

Event type	Argument type	Argument subtypes	Span examples
Alcohol, Drug, & Tobacco	StatusTime*	{none, current, past}	"denies," "no history"
	Amount	–	"1 pack", "3 drinks"
	Duration	–	"since last week"
	Frequency	–	"1-2x/week", "daily"
	History	–	"when he was young"
	Type	–	"smokeless," "methamphetamine"
Employment	StatusEmploy*	{employed, unemployed, retired, on disability, student, homemaker}	"works," "unemployed"
	Duration	–	"since last week"
	History	–	"10 years ago"
	Type	–	"remote office work"
LivingStatus	StatusTime*	{current, past, future}	"lives," "lived"
	TypeLiving*	{alone, with family, with others, homeless}	"with husband," "alone"
	Duration	–	"since he was 12"
	History	–	"until 2 years ago"

Table 3: Annotation guideline summary. *indicates the argument is required

A.2 Prompt Methods

The exact prompting messages are listed below. The annotation guideline is used for the SHAC dataset creation. We removed all the annotation examples in the guideline, as some of the examples are from MIMIC-III. We also remove invalid references to tables.

A.2.1 Message 1 - *System*

GPT-4 role definition for standoff annotation
You are an expert medical annotator and understand the BRAT standoff format very well. You are given a document that contains the following list of entities and events:"

GPT-4 role definition for inline annotation
"You are an expert medical annotator who adds annotations as inline markers in documents. You are given a document to annotate the following list of entities, events, and attributes:

Annotation guideline
The annotation involves the identification of SDOH events, where each SDOH event is represented by a trigger and set of entities. The trigger consists of a multi-word span (word or phrase) and a label indicating the type of SDOH (e.g. employment or tobacco use). All annotated phenomena are defined in terms of the span (words associated with phenomena) and the span type (e.g. amount, status, etc.), and some annotated phenomena, like status, will also include a span label (e.g. current or past). If an annotated span includes a noun, the selected span should include the entire applicable noun phrase. If an annotated span includes a verb, the selected span should include applicable adverbs. The inclusion of entire noun phrases and relevant adverbs is applicable to all annotated phenomena. You are given a document that contains the following list of events and entities:

[EVENTS]

- Alcohol: Alcohol event indicates the usage of alcohol. The trigger span should be a noun phrase describing a general substance type, like "alcohol" if present. Alcohol may be abbreviated ETOH in chart notes. If general substance types are not present, the span should be a relevant verb (e.g. "drink") or a noun phrase describing a specific substance type (e.g. "beer"). Examples of Alcohol spans include: ["alcohol", "beer", "drink"]. ...

[ENTITIES]
<Span only entities>

- Amount: Linked to substance use events (Alcohol, Drug, Tobacco). Amount indicates the quantity of the substance used by the patient (e.g. "three drinks" or "2 packs") and do not include qualitative descriptions (e.g. "significant" or "heavy") . Amount spans should not include frequency information, and frequency and amount spans should not overlap. Examples of Amount entities include: ["one glass", "up to three drinks", "1 ppd", "half a pack", "4 packs"].

  ...

A.2.2 Message 2 - *User*

Annotation instructions for standoff annotation
Based on this annotation guideline, please annotate the following document in the BRAT standoff format.

Annotation instructions for inline annotation
Based on this annotation guideline, please annotate the following document with inline markers.

An example document
HABITS:
Tobacco Use: denies Alcohol Use: denies Drug Use: denies

SOCIAL HISTORY:
Residence: [LOCATION] with husband and kids
Job: no longer works

A.2.3 Message 3 - *Assistant*

Annotated example in standoff annotation
T1 LivingStatus 88 97 Residence
T2 TypeLiving 110 131 with husband and kids
T3 StatusTime 88 97 Residence
...
T11 StatusEmploy 137 152 no longer works
E1 LivingStatus:T1 Type:T2 Status:T3
E2 Tobacco:T4 Status:T6
...
A1 TypeLivingVal T2 with_family
A2 StatusTimeVal T3 current
...

Annotated example for inline annotation
HABITS:
«Tobacco Use»(Tobacco): «denies»(StatusTime-none) «Alcohol Use»(Alcoho): «denies»(StatusTime-none) «Drug Use»(Drug): «denies»(StatusTime-none)

SOCIAL HISTORY:
«Residence»(LivingStatus, StatusTime-current): [LOCATION] «with husband and kids»(TypeLiving-with_family)
«Job»(Employment): «no longer works»(StatusEmploy-unemployed)

A.2.4 Message 4 - *User*

The annotation instructions for standoff and inline annotations are the same as in Message 2.

A.3 Examples of Post-processing

The example document before and after post-processing can be found in Table. 4

	Before Post-processing	After Post-processing
GPT-standoff	T1 LivingStatus 26 35 Residence T2 TypeLiving 43 64 with husband and kids T3 StatusTime 75 97 Residence ... T11 StatusEmploy 110 123 no longer works A1 TypeLivingVal T2 with_family A2 StatusTimeVal T3 current ... E1 LivingStatus:T1 Type:T14 Status:T3 ...	T1 LivingStatus 88 97 Residence T2 TypeLiving 110 131 with husband and kids T3 StatusTime 88 97 Residence ... T11 StatusEmploy 137 152 no longer works A1 TypeLivingVal T2 with_family A2 StatusTimeVal T3 current ... E1 LivingStatus:T1 Status:T3 ...
GPT-inline	... SOCIAL HISTORY: «Residence»(LivingStatus, StatusTime-current): [LOCATION] «with husband and kids»(TypeLiving-with_family) «Job»(Employment) Status: «no longer works»(StatusEmploy-unemployed).	... SOCIAL HISTORY: «Residence»(LivingStatus, StatusTime-current): [LOCATION] «with husband and kids »(TypeLiving-with_family) «Job»(Employment): «no longer works »(StatusEmploy-unemployed)

Table 4: GPT-4 responses before and after post-processing.

A.4 Heuristics for Trigger-argument Connections in the GPT-inline Outputs

We use the sample example in Table. 4 to demonstrate our heuristics for finding the trigger-arguments from GPT-inline outputs and output to BRAT standoff format. For example, after post-processing, the T and A arguments can be directly extracted:

T1 LivingStatus 88 97 Residence
T2 TypeLiving 110 131 with husband and kids
A1 TypeLivingVal T2 with_family
T3 StatusTime 88 97 Residence
A2 StatusTimeVal T3 current
T4 Employment 132 135 Job
T5 StatusEmploy 137 152 no longer works
A6 StatusEmployVal T5 unemployed
...

The above examples contain two trigger spans: T1 and T4. For the rest argument spans, we want to link each of them to its closest trigger span, constrained by the annotation guideline. For example, the distance between the argument T2 and the trigger T1 is 132 (T4 start) - 131 (T2 end) = 1 (character index), and the distance between T2 and the trigger T4 is 131 (T2 start) - 97 (T1 end) = 34 (character index). T1 is closer to the trigger T4. However, because the *TypeLiving* argument can only be attached to the *LivingStatus* trigger, T1 is attached to its closet *LivingStatus* trigger T4. Note that it is possible that a trigger does not contain any arguments or an argument is not attached to any trigger in GPT-inline outputs. Arguments in the above example can be summarized into BRAT events as:

E1 LivingStatus:T1 Type:T2 Status:T3
E2 Employment:T4 Status:T5
...

Teddysum at MEDIQA-Chat 2023: an analysis of fine-tuning strategy for long dialog summarization

Yongbin Jeong[1] Ju-Hyuck Han[2] Kyung Min Chae[2] Yousang Cho[2]
Hyunbin Seo[1] KyungTae Lim[3] Key-Sun Choi[4] YoungGyun Hahm[1]
[1]*Teddysum* [2]*Dept. of Medical Engineering, Konyang University*
[3]*Dep. Applied AI, Seoul National University of Science and Technology*
[4]*Dept. of Medical Artificial Intelligence, Konyang University*
{ybjeong, errol.seo, hahmyg}@teddysum.ai,
{21856503, 23856503, 22806506, kschoi}@konyang.ac.kr, ktlim@seoultech.ac.kr

Abstract

In this paper, we introduce the design and various attempts for Task B of MEDIQA-Chat 2023. The goal of Task B in MEDIQA-Chat 2023 is to generate full clinical note from doctor-patient consultation dialogues. This task has several challenging issues, such as lack of training data, handling long dialogue inputs, and generating semi-structured clinical note which have section heads. To address these issues, we conducted various experiments and analyzed their results. We utilized the DialogLED model pre-trained on long dialogue data to handle long inputs, and we pre-trained on other dialogue datasets to address the lack of training data. We also attempted methods such as using prompts and contrastive learning for handling sections. This paper provides insights into clinical note generation through analyzing experimental methods and results, and it suggests future research directions.

1 Introduction

Multi-turn dialogue summarization in the medical field is an important research area. Medical professionals need to make crucial decisions while consulting with various patients, so creating clinical notes from doctor-patient consultations, which record consultation details and diagnoses, is an essential task for both doctors and patients. However, having doctors write entire clinical notes is time-consuming and reduces consultation efficiency. herefore, it is important to develop technologies that can automatically generate clinical notes from conversation content, allowing doctors to simply review and modify the results, which can shorten consultation times. MEDIQA-Chat 2023's shared task(Ben Abacha et al., 2023) is a benchmark task for summarizing, classifying, and generating clinical dialogue data, and Task B(Yim et al., 2023) is a problem of generating a full clinical note from doctor-patient conversations.

This paper describes how we designed and addressed Task B in MEDIQA-Chat 2023 shared task. Task B is a problem that takes clinical consultation dialogues between doctors and patients as input and generates a summarized full clinical note, as shown in Table 1. This Task has three main challenging features that differentiate it from previous tasks:

1. **Long sequences**: This task takes long conversations as input, with an average of 1,246 words per conversation based on the training data, and generates long outputs with an average of 390 words.

2. **Structured output**: The output is semi-structured data, divided into sections. Some sections are composed of typical paragraph forms, while others are briefly represented using symbols like bullet points.

3. **Low-resourced**: The number of training data pairs is only 64, making it a relatively small dataset.

To effectively address these three challenging issues, we propose the following three methods:

1. **Long sequences**: Utilizing the DialogLED(Zhong et al., 2021) model, which is suitable for processing long conversation inputs.

2. **Structured output**: Implementing a robust model for specific section information by adding a prompt feature.

3. **Low-resourced**: Additional pre-training with outside knowledge using AMI(Carletta et al., 2005) and ICSI(Janin et al., 2003) datasets.

In this study, we first attempted to train the DialogLED model, which is pre-trained on long conversations, with the entire input and output for the

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 394–402
July 14, 2023 ©2023 Association for Computational Linguistics

Input	Output
[doctor] hi , martha . how are you ? [patient] i'm doing okay . how are you ? [doctor] i'm doing okay . so , i know the nurse told you about dax . i'd like to tell dax a little bit about you , okay ? [patient] okay . [doctor] martha is a 50-year-old female with a past medical history significant for congestive heart failure , depression and hypertension who presents for her annual exam . so , martha , it's been a year since i've seen you . how are you doing ? [patient] i'm doing well . i've been traveling a lot recently since things have , have gotten a bit lighter . and i got my , my vaccine , so i feel safer about traveling . i've been doing a lot of hiking . uh , went to washington last weekend to hike in northern cascades, like around the mount baker area . [doctor] nice . that's great . i'm glad to hear that you're staying active , you know . i , i just love this weather . i'm so happy the summer is over . i'm definitely more of a fall person . [patient] yes , fall foliage is the best	CHIEF COMPLAINT Annual exam. HISTORY OF PRESENT ILLNESS Martha Collins is a 50-year-old female with a past medical history significant for congestive heart failure, depression, and hypertension who presents for her annual exam. It has been a year since I last saw the patient. . . . REVIEW OF SYSTEMS • Ears, Nose, Mouth and Throat: Endorses nasal congestion from allergies. • Cardiovascular: Denies chest pain or dyspnea on exertion. . . . PHYSICAL EXAMINATION • Cardiovascular: Grade 3/6 systolic ejection murmur. 1+ pitting edema of the bilateral lower extremities.

Table 1: Task B Input and Output example

long sequence problem. And, we pre-trained the model with the AMI and ICSI datasets for solving the low-resource problem before training with the MEDIQA-chat Task B(Yim et al., 2023) dataset. As a result, the ROUGE-1 score for the validation data was relatively high as 0.575 compared to the previous performance. However, when analyzing the results, there were some error cases where the necessary sections did not appear well, or the section name was incorrect, causing the subsequent context to flow in the wrong direction. Taking inspiration from how people create clinical notes while considering the overall section structure, we devised a way to give the model hints about the section that needs to be created. Like determining the title of an article in advance, we added prompts to the input to help the model understand which section to create and generated summary sentences for only that section. We then combined the summary sentences for each section to create a complete summary note. However, this approach increased the recall for sections but had the problem of many sections appearing that did not need to appear. The biggest issue was that the content discussed in other sections was repeated. In about 70% of cases, the content was repeated, resulting in a lower overall note score of 0.449 based on ROUGE-1. In addition, we tried various methods such as using only the section name as a prompt, changing the prompt sentence, or wrapping the section name in tokens, but these did not make a significant difference in performance.

To address these issues, we incorporated contrastive learning(Chen et al., 2020). We set the summary of the section corresponding to the prompt as the positive sample for contrastive learning and applied the cross-entropy loss, as usual, to generate summaries close to it. To avoid generating repetitive summaries similar to other sections, we used the summaries of other sections as negative samples and set the loss so that cosine similarity would decrease. As a result of this contrastive learning, the occurrence of repetitive content decreased by nearly 60%, indicating that contrastive learning had some influence.

This paper introduces various attempts for Task B, such as using the DialogLED model, creating partial summaries for each section using prompts, changing prompts, and employing contrastive learning to improve Task B's performance. While we have not yet found a perfect solution, we aim to provide valuable insights for considering approaches to Task B through quantitative and qualitative analysis of data and experimental results. Source code and all the trained models are available on our GitHub repository[1].

2 Related Work

Abstractive summarization of extended conversations is typically approached using generative models such as BART(Lewis et al., 2019). When the input exceeds the model's input length constraint, rudimentary techniques such as truncating the end or middle portions of the input are employed. Methods like Presumm(Liu and Lapata, 2019) address long inputs by extracting key sentences and subsequently performing abstractive summarization on them. Alternatively, models like DialogBERT(Gu et al., 2020) process inputs at the utterance level, encoding and merging them to accommodate the entire conversation. Recently, a trend towards models like Longformer(Beltagy et al., 2020), DialogLM, and DialogLED(Zhong et al., 2021) has emerged, as these models can handle longer inputs owing to their expanded input length capabilities. Prominent datasets for summarizing long conversations include AMI(Carletta et al., 2005) and ICSI(Janin et al., 2003) datasets. These datasets, like Task B, involve summarizing lengthy dialogues. However, unlike Task B, where the domain is clinical, the domain for these datasets is meetings, and the summaries are composed of unstructured paragraphs. And, models such as clinicalBERT(Huang et al., 2019), which are trained in the clinical domain, have emerged, but they have limitations in terms of input length compared to recent models.

3 Dataset and task design

The Task B of MEDIQA-Chat 2023, as presented in Table 1, involves generating a full clinical note based on clinical dialogues between docker and patient. The input comprises multi-turn conversations, with specialized terminology from the clinical domain such as disease names and medication names appearing frequently. The output is organized into distinct sections, including 'CHIEF COMPLAINT' and 'HISTORY OF PRESENT ILLNESS.' Notably,

[1]https://github.com/teddysum/MEDIQA-Chat-2023-Teddysum

	Train	**Valid**
# of data	67	20
# of avg input Char	6443	6124
# of avg output Char	2649	2716
# of avg input Word	1246	1169
# of avg output Word	390	400
# of avg input Token	1480	1401
# of avg output Token	573	589
# of max input Token	3437	2020
# of max output Token	1192	1054
# of input over 512 tokens	67	20
# of output over 512 tokens	40	11

Table 2: Data Statistic

not all sections are required to appear, and there is no predefined structure specifying which sections should be included in the note. Among the sections, 'CHIEF COMPLAINT' is the most common, appearing in 63 out of 64 training samples, whereas 'PAST MEDICAL HISTORY' is the least common, occurring only once. Furthermore, the same section can be represented using different names, such as 'CC:' (four instances) and 'CHIEF COMPLAINT' (59 instances).

Table 2 illustrates the overall statistics of the dataset, highlighting the limited number of training samples (67). Furthermore, the input length is relatively long, averaging 1,246 words per training sample, which translates to an average of 1,480 tokens when tokenized using the BERT(Devlin et al., 2019) tokenizer. The longest document consists of 3,437 tokens, indicating that conventional models such as BART(Lewis et al., 2019), with a maximum input length of approximately 1,024 tokens, may not be suitable for directly processing inputs for this task.

Considering these aspects, the challenges to be addressed in this task, distinct from other summarization tasks, can be summarized as handling long dialogues, working with a small dataset, generating semi-structured notes, and adapting to the intricacies of the clinical domain.

4 Method

To address these challenges, we employed two primary approaches. The first approach involves using the entire dialogue as input and generating the full clinical note as output, thereby adopting an end-to-end learning method. The second approach consists of creating notes for each section individually, and

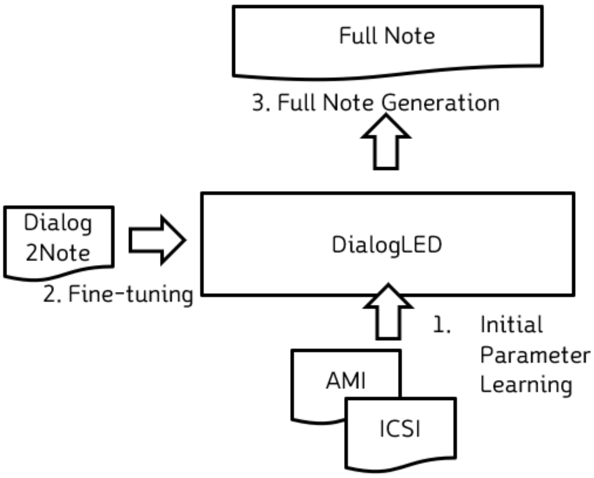


Figure 1: Fine-tuning to full clinical note

subsequently merging the section-specific notes to form a comprehensive full clinical note.

4.1 Full fine-tuning to clinical note

In the first approach, we designed a model based on the DialogLED model, which has been pre-trained on long conversational data, to handle the lengthy dialogue inputs. Although the DialogLED model is trained on dialogue data, it has not been trained specifically for summarization tasks. To overcome the limitations posed by the small dataset, we conducted pre-training using the AMI and ICSI datasets, which consist of summaries of long conversations. Subsequently, we fine-tuned the model using the MEDIQA dataset as illustrated in Figure 1. During this process, we did not explicitly handle the section structure, allowing the model to learn in an end-to-end manner. While this approach demonstrated the advantage of generating accurate content for the full clinical note, it fell short in creating section headers and properly distinguishing content between sections.

4.2 Prompt-based partial fine-tuning

The second approach involves handling the section structure through a prompt-based method. As the first approach struggled to generate section headers effectively, this method aims to produce content for each section, create section headers through post-processing, and combine the resulting partial notes to form a full note. We added a special token, <CMD>, to the prompt as in "$conversation<CMD>Based on this conversation, make a summary of the $SectionName," which is appended after the dialogue. Various forms of prompts were experimented with, as demonstrated in Table 3.

Style	Prompt
Style 1: simple prompt	$conversation<CMD>Based on this conversation, make a summary of the $SectionName
Style 2: special tokens (before/after)	$conversation<CMD>Based on this conversation, make a summary of the <SEC>$SectionName</SEC>
Style 3: section name only	$conversation<CMD> $SectionName

Table 3: Prompt styles

4.3 Prompt-based contrastive learning

Prompt-based partial fine-tuning method often led to the inclusion of content that should have appeared in other sections, particularly when the model's understanding of the prompt was inadequate. To address this issue, we also conducted experiments using a contrastive learning approach for training. In the traditional learning approach, the loss function computes the cross-entropy value for the correct summary. In contrast, the contrastive learning approach utilizes content from other sections as negative samples to prevent the generation of content from different sections. When the model's predicted value is y, the correct summary for the target section is p, and the summary for a randomly chosen section, excluding the target section, is n, the loss can be calculated as follows.

$$CEloss = CrossEntropy(y, p)$$

$$CSloss = CosineSimilarity(y, n)$$

$$loss = a * CEloss + (1 - a) * CSloss$$

5 Experiments

5.1 Implementation Details

In this study, we utilized the pre-trained DialogLED model[2]. The batch size was set to 4, with a maximum input length of 5120 and a maximum output length of 1024. We monitored the training process for up to 80 epochs. The AdamW optimizer was employed, with a learning rate of 2e-5, an epsilon value of 1e-8, 50 warm-up steps, and a

[2]https://huggingface.co/MingZhong/DialogLED-large-5120

learning rate decay of 0. For the training, 4 A100 GPUs(80GB) were used with 12GB dedicated to model uploading and 47GB used for uploading one batch of training and evaluation data. The first model, which fine-tuned the full note in one go, took approximately 3 minutes per epoch, while the second model, which generated summaries for individual sections, took around 13 minutes per epoch.

For the pre-training of the DialogLED parameters, the AMI and ICSI datasets were utilized. A total of 170 samples, comprising 117 from the AMI dataset and 53 from the ICSI dataset, were used as training data, while 20 from the AMI dataset and 25 from the ICSI dataset were used as the validation set. The input data was pre-processed to match the input format of the DialogLED model, with speakers and utterances represented using colons, such as "A: utterance."

During pre-processing, the MEDIQA-chat format, which presents the speaker in brackets followed by their utterance (e.g., "[doctor] utterance"), was transformed to match the conversation format learned by the DialogLED model (e.g., "doctor: utterance") using colons. In the second approach, where summaries were generated for each section, section names were heuristically defined and unified, as shown in Table 4, by converting synonymous section headers like 'CC:' and 'CHIEF COMPLAINT' to 'CHIEF COMPLAINT'. For the 'ASSESSMENT AND PLAN' section, instances where the section appeared as 'ASSESSMENT AND PLAN' or separately as 'ASSESSMENT' and 'PLAN', or only one of them appeared, were combined into a single 'ASSESSMENT AND PLAN' section.

5.2 Experimental Results

Since the ground truth to the test set in Task B is not open to the public, all evaluations were conducted on the validation set.

5.2.1 Quantitative Analysis

Table 5 presents the scores for generating full notes. The FT to Full Note model generates full notes given the entire conversation as input, while the models labeled PT partially generate notes based on prompt-based approach. As shown in Table 3, there are three styles for the prompts. PT contrastive refers to the model trained using a contrastive learning approach for prompt-based models. As the results indicate, the highest score for full notes is achieved by the FT to Full Note model, while the

Original section name	Unified Section name	# of appearance
ASSESSMENT	ASSESSMENT AND PLAN	29
ASSESSMENT AND PLAN	ASSESSMENT AND PLAN	34
PLAN	ASSESSMENT AND PLAN	32
EXAM	PHYSICAL EXAMINATION	4
PHYSICAL EXAM	PHYSICAL EXAMINATION	44
PHYSICAL EXAMINATION	PHYSICAL EXAMINATION	16
HISTORY OF PRESENT ILLNESS	HISTORY OF PRESENT ILLNESS	45
HPI:	HISTORY OF PRESENT ILLNESS	4
REVIEW OF SYSTEMS	REVIEW OF SYSTEMS	50
CC:	CHIEF COMPLAINT	4
CHIEF COMPLAINT	CHIEF COMPLAINT	59
RESULTS	RESULTS	52
CURRENT MEDICATIONS	MEDICATIONS	8
CURRENT MEDICATIONS:	MEDICATIONS	1
MEDICATIONS	MEDICATIONS	19
PAST HISTORY	HISTORY	9
PAST MEDICAL HISTORY:	HISTORY	1
MEDICAL HISTORY	HISTORY	18
SOCIAL HISTORY	HISTORY	28
SURGICAL HISTORY	HISTORY	7
FAMILY HISTORY	HISTORY	10
IMPRESSION	IMPRESSION	4
INSTRUCTIONS	INSTRUCTIONS	32
VITALS	VITALS	23
VITALS REVIEWED	VITALS	3

Table 4: Section Name Definition

Models	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	BLEURT
FT to Full Note	0.575	0.288	0.315	0.692	0.405
PT Style 1	0.431	0.232	0.207	0.662	0.336
PT Style 2	0.439	0.234	0.201	0.661	0.324
PT Style 3	0.447	0.235	0.208	0.66	0.342
PT Contrastive	0.414	0.215	0.192	0.669	0.350

Table 5: Full note generation score

Models	SH P	SH R	SH F1	Context Repetition Rate
FT to Full Note	0.91	0.47	0.62	0.08
PT Style 1	0.65	0.86	0.74	0.67
PT Style 2	0.67	0.87	0.76	0.71
PT Style 3	0.66	0.85	0.74	0.72
PT Contrastive	0.65	0.87	0.74	0.64

Table 6: Error statistic analysis. SH means Section Header.

Models	ROUGE-1	BERTScore	BLEURT
L to L	0.408	0.646	0.309
L to S	0.575	0.692	0.405
S to L	0.518	0.668	0.387
S to S	0.563	0.69	0.397

Table 7: Validation loss and score in fine-tuning to full note. L signifies the point at which the validation loss converges, and S denotes the point of validation score convergence. The prefix L and S pertain to the training phase on the AMI and ICSI datasets, while the suffix L and S refer to the training on Task B data. For instance, L to S indicates that the model was trained on Task B using the point at which the validation loss converges during the training on the AMI and ICSI datasets, and the recorded score corresponds to the highest validation score for that model.

PT models exhibit similar scores overall.

Table 6 provides statistics on the errors that occurred during the generation of full notes. Section Header P, R, and F1 represent the Precision, Recall, and F1 score calculated based on the presence of Section headers in both the Validation set and the model output. Specifically, True Positive (TP) occurs when a Section header appearing in the Validation set also appears in the model output, False Negative (FN) occurs when a Section header present in the Validation set does not appear in the model output, and False Positive (FP) occurs when the model output generates a Section header that does not exist in the Validation set. As shown, the FT to Full Note model exhibits high precision but low recall, leading to a reduced overall F1 score. In contrast, the PT models demonstrate better generation of section headers due to their higher recall.

The Context Repetition Rate refers to the proportion of summary content in a section that appears in other sections. In the FT to Full Note approach, the repetition rate is relatively low at 0.08, indicating that there is minimal repetition of content. In contrast, the PT approach exhibits a repetition rate near 0.7, suggesting that the majority of sections contain similar content to other sections. Although the PT method generates Section Headers more effectively than the FT to Full Note approach, the overall score is lower due to this repetition issue. To address this problem, contrastive learning was implemented; however, the PT Contrastive method still exhibits a repetition rate of 0.64. While this is a decrease in repetition, it still remains a relatively high value. The criterion for determining repetition is a cosine similarity of 0.5 or higher between the summary sentences of two sections.

Taking these factors into account, the FT to Full Note approach demonstrates a good ability to summarize the overall context but falls short in handling individual sections. As a result, it tends to generate frequently occurring sections and fill them with substantial content, leading to lower performance in section header evaluation. However, since the entire note is generated at once, content from one section does not appear in other sections, maintaining distinctiveness. On the other hand, while the prompt based partial generation method yields better evaluation results in section creation, the repetition issue remains inevitable, and even methods like contrastive learning cannot easily resolve it. This is because the model generates summaries

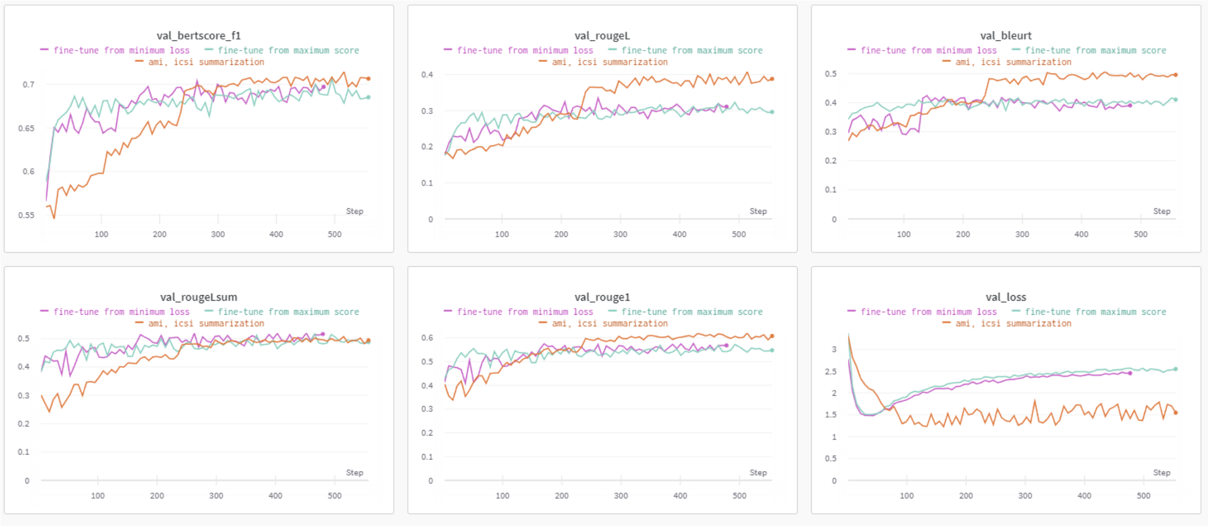


Figure 2: Training graph

separately for each section, so it does not know the content of summaries in other sections, leading to the learning of frequently occurring content repetition.

Additionally, during the learning process for Task B data after initial parameter tuning with the AMI and ICSI datasets, it was discovered that the points at which validation loss and validation score converge are different. As shown in Figure 2, when training to the AMI and ICSI data, the validation loss converged at 100 steps, but the validation score continued to rise, converging at approximately 300 steps. To analyze this, we evaluated two models: one trained on Task B data from the model at 100 steps and the other from the model at 300 steps. The results are shown in Table 7. When using a model trained on AMI and ICSI data, it can be seen that using a model from the point where validation loss converges performs better than using a model with a higher validation score. This suggests that the model may have started overfitting from the point where the validation loss converged. Therefore, a model from the point of validation loss convergence is more suitable for application to other domains like Task B, as it makes more general predictions and generates content. The mismatch between the timing of validation loss and validation score convergence also occurs when training on Task B data, which seems to be due to the small amount of data. In other words, when a large model learns from a small dataset, it appears to perform better when it is somewhat overfitted and dependent on the dataset, rather than at the point with the lowest validation loss, which seems to be more generally and appropriately trained.

Futhermore, as seen in Figure 2, the validation loss for the AMI and ICSI datasets converges and no longer increases, while for Task B, the validation loss increases again after converging. This more clearly shows the issue of overfitting. To address this issue, we can consider increasing the amount of dataset, constructing a model that is more suitable for semi-structured datasets to prevent overfitting, or modifying the loss function to be more appropriate for the dataset.

5.2.2 Qualitative Analysis

We also conducted a qualitative evaluation of the model results. We evaluated the model by comparing all the inferred results on the validation set. We compared the generated summaries with the correct summaries by section, examining whether the necessary sections were created, whether unnecessary sections were created, and what differences existed between the content of the created sections and the correct summary. First, we verified whether the results of the quantitative evaluation were consistent with those of the qualitative evaluation. Additionally, we qualitatively analyzed the phenomenon where the point of convergence for the validation score and the validation loss differed.

When evaluating the performance of the model qualitatively, the results were similar to those of the quantitative evaluation. The FT to Full Note model actually performed better than the PT model. In particular, it demonstrated high performance in frequently occurring and easy sections, such as the 'CHIEF COMPLAINT' section. In the FT model, as in the quantitative analysis, there were many instances where the necessary sections were not created. Upon closer inspection, it was found that the content was often generated but not separated by section headers and instead placed in a single, generic section. This appears to be due to the imbalance in the frequency of section occurrences in the data, data scarcity, and the model's limitations in understanding the semi-structured structure. In the case of the PT model, 'content repetitions', in which the contents of other sections appear redundantly, was remarkably high, as would be evaluated in the quantitative analysis.

Secondly, we analyzed the qualitative differences between the points where the validation loss converged and the points where the validation score converged. As a result, in the case of the FT to Full Note model, when viewed qualitatively, it showed a better ability to create sections and fewer instances of empty content within the sections at the points where the score converged. In the case of the PT model, as the number of epochs increased, the overlap of content with other sections occurred more frequently. This suggests that the model may consider it more advantageous to create content, even if it is incorrect, rather than not create it by mistake, for content that appears most generically. In other words, it can be seen as overfitting to the correct answers in the data. However, in sections such as 'IMPRESSION', 'INSTRUCTIONS', and 'VITALS', the model tended to produce identical outputs at low epochs and different outputs as the number of epochs increased. This suggests that the model is gaining some understanding of the prompts as it learns.

Additionally, we analyzed the learning process for Task B after training on AMI and ICSI data. We qualitatively compared the differences between the models trained from the point where validation loss converged (L to S) and the models trained from the point with high validation scores (S to S). As a result, the S to S models tended to overfit to the AMI and ICSI data, causing many inappropriate words that do not exist in the MEDIQA data to appear. However, for patient information such as names and ages, the S to S models performed better than the L to S models. This suggests that the S to S models overfit to the AMI and ICSI data and learned to reference the original text more extensively. Although the L to S models achieved higher quantitative scores than the S to S models, the accuracy of information such as patient names and ages is more important for clinical notes. Therefore, the S to S models can be considered better suited for this task.

6 Conclusion

In this paper, we have conducted various experiments for Task B and analyzed the results. We utilized the DialogLED model to handle long inputs and employed additional data, such as the AMI and ICSI datasets, to address the limited data issue. As a result, the model generating the entire note at once achieved high scores for the full note but low scores for section header creation. In contrast, the approach of creating section summaries separately and then combining them for a complete note had high scores for section headers but low scores for the entire full note due to repetition issues. Analyzing these results, it appears that generating the entire content at once is a more appropriate approach since it is challenging to avoid repetitive content when creating section summaries separately. In conclusion, to learn the characteristics of Task B, which requires the creation of a semi-structured structure, a hybrid approach that combines generating the entire content at once and creating section headers separately is needed, rather than relying on simple fine-tuning.

Limitations

In the experiments conducted by the team, several limitations were observed. Firstly, the handling of the semi-structured structure was not accomplished perfectly. When generating the entire content at once, many sections did not appear, and when creating sections separately, the content of the notes was repetitive. Additionally, addressing the clinical domain properly was not achieved.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (2022-0-00078, Explainable Logical Reasoning for Medical Knowledge Generation)

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *ArXiv*, abs/2004.05150.

Asma Ben Abacha, Wen-wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen. 2023. Overview of the mediqa-chat 2023 shared tasks on the summarization and generation of doctor-patient conversations. In *ACL-ClinicalNLP 2023*.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Maël Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska Masson, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre D. Wellner. 2005. The ami meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. *ArXiv*, abs/2002.05709.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

Xiaodong Gu, Kang Min Yoo, and Jung-Woo Ha. 2020. Dialogbert: Discourse-aware response generation via learning to recover and rank utterances. In *AAAI Conference on Artificial Intelligence*.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *ArXiv*, abs/1904.05342.

Adam L. Janin, Don Baron, Jane Edwards, Daniel P. W. Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. The icsi meeting corpus. *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, 1:I–I.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Annual Meeting of the Association for Computational Linguistics*.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *ArXiv*, abs/1908.08345.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Acibench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Submitted to Nature Scientific Data*.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Dialoglm: Pre-trained model for long dialogue understanding and summarization. In *AAAI Conference on Artificial Intelligence*.

Rare Codes Count: Mining Inter-code Relations for Long-tail Clinical Text Classification

Jiamin Chen[1,2], **Xuhong Li**[2], **Junting Xi**[1], **Lei Yu**[1], **Haoyi Xiong**[2]
[1]Beihang University, Beijing, China
[2]Baidu Inc., Beijing, China
{lixuhong, xionghaoyi}@baidu.com, {jiaminchen, drinking, yulei}@buaa.edu.cn

Abstract

Multi-label clinical text classification, such as automatic ICD coding, has always been a challenging subject in Natural Language Processing, due to its long, domain-specific documents and long-tail distribution over a large label set. Existing methods adopt different model architectures to encode the clinical notes. Whereas without digging out the useful connections between labels, the model presents a huge gap in predicting performances between rare and frequent codes. In this work, we propose a novel method for further mining the helpful relations between different codes via a relation-enhanced code encoder to improve the rare code performance. Starting from the simple code descriptions, the model reaches comparable, even better performances than models with heavy external knowledge. Our proposed method is evaluated on MIMIC-III, a common dataset in the medical domain. It outperforms the previous state-of-art models on both overall metrics and rare code performances. Moreover, the interpretation results further prove the effectiveness of our methods. Our code is publicly available[1].

1 Introduction

The International Classification of Diseases (ICD) is a worldwide diagnostic tool published and maintained by the World Health Organization (WHO). The ICD coding, a task of assigning ICD codes according to the electronic medical records (EMRs), facilitates a lot of activities in health care, such as morbidity and mortality statistical analysis, medical billing and decision support systems (W. et al., 2020; Sutton et al., 2020). Since the traditional manual EMRs coding is time-consuming and prone to error (O'malley et al., 2005), its automation has always been attracting attention since 1990s (de Lima et al., 1998). Most of the existing methods treat the automatic ICD coding as a supervised multi-label document classification task (Xie and Xing, 2018; Mullenbach et al., 2018). By learning the text representations with an RNN (Vu et al., 2020), CNN (Mullenbach et al., 2018; Liu et al., 2021) or Transformer (Biswas et al., 2021) based encoder, the model extracts the code-relevant features via a trainable query matrix and predicts the codes with multiple binary classifiers.

Rare Code Prediction. Although the introduction of deep learning methods significantly improves the overall metrics for ICD coding, the extremely long-tail distribution over labels still makes the prediction for rare diseases or procedures challenging. Taking the MIMIC-III Dataset as an example, among all the discharge summaries, the most frequent code appears 20,053 times while the codes which occur less than 100 times constitute 12% of the whole dataset. In the supervised methods, learning the distinguished representation for each code through training samples requires rich data resources, leading to better performances on frequent codes than less frequent ones. Collecting sufficient documents for rare codes can be very difficult and expensive, which makes rare code prediction a critical task in automatic ICD coding.

Regarding this subject, several research directions have been explored. For example, some unsupervised methods have been proposed (W. et al., 2020; Song et al., 2020), but there remain clear margins compared to the supervised ones. Most of the previous works with supervised methods focus on the top 50 most frequent codes and extend the model usage on infrequent codes. But their results on rare codes are far from satisfactory. A few studies concerning the few-shot literature (Wang et al., 2021; Yuan et al., 2022) are proposed to improve the rare code performance by enriching the code descriptions via external knowledge sources. However, accessing heavy external sources can be

[1]https://github.com/jiaminchen-1031/Rare-ICD

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 403–413
July 14, 2023

complicated, and it is possible to introduce unexpected bias and noise facing immense knowledge.

In this work, we propose a more efficient method by strengthening the inter-code relations to improve the rare code performances. The existing supervised learning methods with label-wise attention (Vu et al., 2020; Mullenbach et al., 2018; Liu et al., 2021; Biswas et al., 2021) can hardly capture the helpful inter-code relations for rare codes. And we consider it as the reason for their bad performance on rare codes. We show in Figure 1 the correlations between code representations in the traditional label-wise attention method and our method. As suggested in the left figure, for frequent codes, the model can well learn their strong or weak correlations with the other codes via sufficient training samples. On the contrary, for rare codes, the model fails in building the useful connections and presents irrelevance with most of the codes. By enhancing the relations, as shown in the right figure for our method, the model performance on rare codes can be effectively improved.

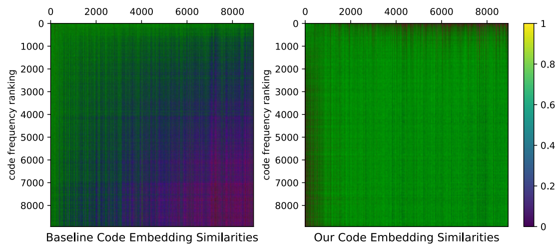


Figure 1: Code correlations by their embedding similarities in the traditional method (Vu et al., 2020) (*left*) and ours (*right*). The axis is arranged according to label frequency, where 0 indicates the most frequent and a greater value means less frequent.

Inter-code Relations. As indicated in Figure 2, we present the inter-code relations in this work by co-occurrence and hierarchy. Generally, code co-occurrence is acquired by counting the co-appearing times of two diseases in the same clinical text from a group of data. Revealing this information explicitly is helpful for the model to incorporate the relations between different codes. However, for rare codes, due to the lack of related samples, their co-occurrence relations with other codes can be incomplete or biased. To alleviate this issue, we propose to introduce a parent-child structure for each code, thus being able to explore the co-occurrence under different levels and merge them to complement the co-occurrence for rare codes. Here, we extend the definition of being hierarchical, follow the ICD-9 Official Guidelines released by the U.S. Federal Government's Department of Health and Human Services, and propose a **code-organ-system** hierarchical structure: level-0 (code itself), level-1 (codes of the similar organs) and level-2 (codes of the same system).

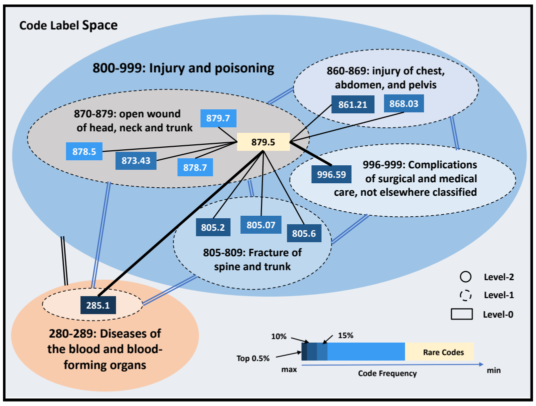


Figure 2: The inter-code relations exploited in this work. Taking the rare code 879.5 as an example, we build its relations with other codes by co-occurrence and enhance them with the code-organ-system hierarchy and the connections between different categories.

Our Contributions. In this work, we propose a novel method to improve rare code prediction by enhancing the connections between frequent and rare codes. The inter-code relations are explored via code descriptions, the code-organ-system hierarchy, and co-occurrence, which can be easily accessed without the necessity of bringing heavy external knowledge. Although quite a few studies have concentrated on inter-code relations (Tsai et al., 2021; Yan et al., 2010; Cao et al., 2020a), to our knowledge, we are the first to bond the inter-code relations specifically with rare code prediction and propose to exploit the inter-code relations under the code-organ-system hierarchy to tighten the weak connections for rare codes. We evaluate our method on the MIMIC-III-full dataset by their metrics on all codes and rare codes, where it outperforms the previous state-of-art models in automatic ICD coding.

2 Related Works

Automatic ICD Coding. Medical text categorization has been an important task in medical NLP for a quite long time. Early works adopt traditional machine learning methods for coding (Larkey and Croft, 1996; Pestian et al., 2007; Perotte et al., 2014). With the rising of neural networks, the automatic ICD coding began to be considered as

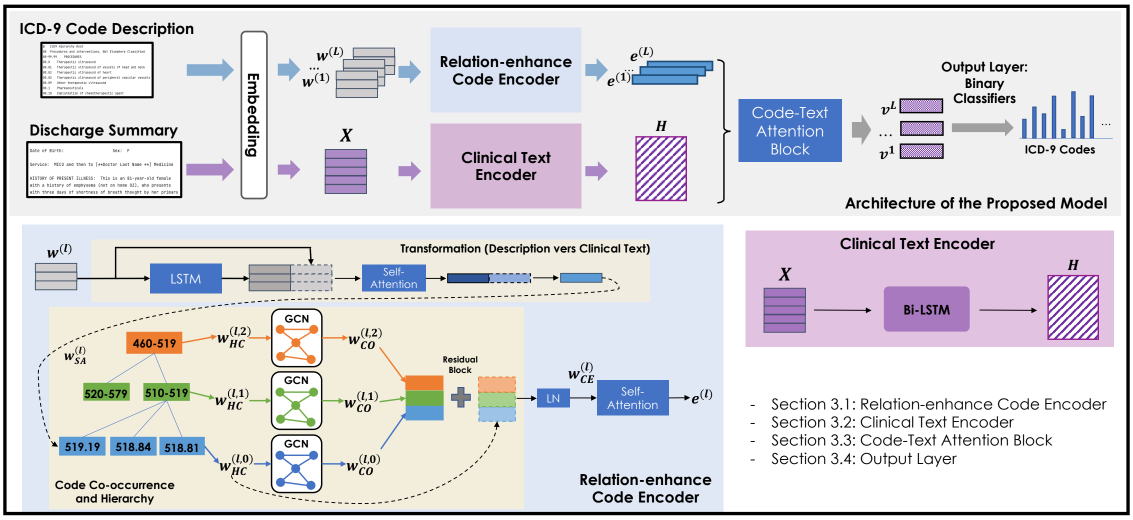


Figure 3: The architecture of our model.

a multi-label classification task. Mullenbach et al. (2018) propose a convolutional neural network text encoder and an attention layer to capture the important features of each code. Vu et al. (2020) further develop the label-wise attention layer with randomly initialized label representations and propose to use LSTM as text encoder. Various CNN (Li and Yu, 2020; Liu et al., 2021), RNN (Vu et al., 2020) and Transformer (Biswas et al., 2021) variants have also been used to encode the clinical documents. Some works (Xie et al., 2019; Cao et al., 2020a) propose to use GCN to integrate the code hierarchy and co-occurrence into the representation learning. Yuan et al. (2022) propose to enrich the code semantic information by introducing its synonyms from the **U**nited **M**edical **L**anguage **S**ystem (UMLS). Concerning the benchmark for evaluation, although some ICD-10 datasets have been collected and used in previous works (Cao et al., 2020b; Koopman et al., 2015), MIMIC-III for ICD-9 codes is still the only available dataset for clinical documents up till now.

Few-shot Learning for Long-tail ICD Codes. Few-shot learning targets at achieving good performances to the classes where a few samples are available (Medina et al., 2020). Due to the long tail of medical document dataset, some methods concerning ICD coding have also been proposed with similar strategies of few-shot learning. Current strategies can be divided into two types. The first works on improving the training process to achieve a better performance, such as proposing a novel optimization mechanism (Li et al., 2017) and modifying the loss function (Lin et al., 2017). For ICD coding, some works introduce different weights to the loss terms to help rare code prediction, such as Focal Loss (Lin et al., 2017) in Effective-CAN (Liu et al., 2021) and label-distribution-aware margin (Cao et al., 2019) in TransICD (Biswas et al., 2021).

The other type aims at learning a similarity function between frequent and few-shot labels (Vinyals et al., 2016). Matching networks (Geng et al., 2020) give predictions by searching the few-shot labeled support set through cosine similarities. In ICD coding, this strategy is usually achieved by introducing external knowledge to obtain the similarities. Vu et al. (2020) use the code formulation rules to apply a hierarchical joint learning mechanism. Some works bring in code relations from reliable sources, such as Wikipedia (Wang et al., 2021), code synonyms from UMLS (Yuan et al., 2022) and knowledge graphs (Xie et al., 2019).

3 Approach

In this section, we introduce the whole architecture of our model, which is illustrated in Figure 3. First, both the medical records and code descriptions are tokenized and embedded via a shared Word2Vec (Mikolov et al., 2013). Then we adopt a dual encoder architecture to encode code descriptions and medical records respectively. The embedding of code descriptions is put into a **Relation-enhanced Code Encoder** (Section 3.1) to strengthen the connections between codes, especially between the rare codes and frequent codes. We exploit the co-occurrence and the hierarchical structure of ICD codes via a series of modules inside the proposed code encoder. In parallel, the embedding of clinical texts is fed into a **Clinical Text Encoder** (Section 3.2) for contextualization. The outputs of these two encoders interact in the **Code-Text Attention Block** (Section 3.3), where the important words are highlighted and combined to generate a new code-specific vector with the representation of each code served as the query. The combination of the weighted words for each code is finally fed into the corresponding binary classifier in the **Output Layer** (Section 3.4) to calculate the probabilities.

3.1 Relation-enhanced Code Encoder

This encoder aims to identify the useful relations between different codes and enhance them via the built representations. We obtain the code descriptions from the World Health Organization (2016). Through the pretrained Word2Vec, the description for the code l is transformed into $\mathbf{w}^{(l)} \in \mathbb{R}^{N_c \times d_e}$. We denote N_c the number of words in each code description, L the total number of code space and d_e the embedding size.

Contextual Transformation. Unlike discharge summaries, the code descriptions are usually noun phrases instead of sentences. During the pretraining of Word2Vec, we construct their embeddings directly from contextual information. As the contexts of a word in the sentence and the noun phrase are different, a gap exists in their embeddings between the words in code descriptions and clinical texts. To solve the gap, we propose the following module to align the words in them. The major differences between the clinical texts and code descriptions are the word order and writing style. Therefore, we feed the word embeddings of code descriptions into an LSTM, concatenate the output and input, then put the concatenated results into a self-attention layer to combine the most important temporal features, and finally generate an overall representation $w_{\mathrm{SA}}^{(l)}$ for code l:

$$\tilde{w}^{(l)} = [\mathbf{w}^{(l)} \oplus \overrightarrow{\mathrm{LSTM}}(\mathbf{w}^{(l)})] \quad , \tag{1}$$

$$\alpha_{\mathrm{SA}}^{(l)} = \mathrm{softmax}(\tilde{w}^{(l)} \cdot W_{\mathrm{Att}} + b_{\mathrm{Att}}) \quad , \tag{2}$$

$$w_{\mathrm{SA}}^{(l)} = \alpha_{\mathrm{SA}}^{(l)} \cdot \tilde{w}^{(l)} \cdot W_{\mathrm{SA}} \quad , \tag{3}$$

where $\alpha_{\mathrm{SA}}^{(l)}$ refers to the self-attention weight, $W_{\mathrm{SA}} \in \mathbb{R}^{(d_e+u_{\mathrm{SA}})\times d_e}$, $W_{\mathrm{Att}} \in \mathbb{R}^{(d_e+u_{\mathrm{SA}})\times 1}$ and $b_{\mathrm{Att}} \in \mathbb{R}^1$ are shared trainable vectors for all codes with u_{SA} the hidden size of LSTM and softmax is applied at the row level.

Code Hierarchy. Due to the lack of samples, the collected co-occurrence relations for the rare codes can be incomplete or biased. To further exploit the inter-code relations for rare codes, we introduce the code-organ-system hierarchical structure, where three levels are defined: level-0 (itself), level-1 (similar organs), and level-2 (same system). Observed from sufficient samples, the codes concerning similar organs or systems have some intrinsic co-occurrence links. These links can be utilized to enrich the connections for rare codes and make them more reliable and less biased via shared embeddings. We define the embedding of each level as the average of all the codes belonging to same categories. Thus, we obtain three embeddings which describe code l from different levels:

$$w_{\mathrm{HC}}^{(l,p)} = \begin{cases} w_{\mathrm{SA}}^{(l)}, & p = 0 \\ \dfrac{1}{|C_p|} \displaystyle\sum_{i \in C_p} w_{\mathrm{SA}}^{(i)}, & p \in \{1, 2\} \end{cases} , \tag{4}$$

where $p \in \{0, 1, 2\}$ is the level number and C_p the category which code l belongs to.

Code Co-occurrence. After obtaining the embeddings at different levels for each code, we adopt three GCNs to exploit the co-occurrence on these three levels. The inputs of these GCNs are the corresponding level embedding $w_{\mathrm{HC}}^{(l,0)}$, $w_{\mathrm{HC}}^{(l,1)}$, $w_{\mathrm{HC}}^{(l,2)}$ and the adjacency matrix $A^{(1)}$, $A^{(2)}$, $A^{(3)}$ based on the co-appearing times for the three levels. The co-occurring times are sampled from a group of data, which is later analysed in Section 4.6. We use a standard convolution computation (Kipf and Welling, 2017):

$$E_{i+1} = \mathrm{ReLU}(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} E_i W_i) \quad , \tag{5}$$

where $\tilde{A} = A + I$, I is the identity matrix, $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$. Thus, we have the output node representations $[w_{\mathrm{CO}}^{(1,p)}; ...; w_{\mathrm{CO}}^{(L,p)}]$ with $w_{\mathrm{CO}}^{(l,p)}$ for code l at level p.

Code Embedding. We merge the three level representations in this module and obtain a specific representation $\mathbf{e}^{(l)}$ for code l in the label space:

$$w_{\mathrm{CE}}^{(l,p)} = \mathrm{LayerNorm}(w_{\mathrm{CO}}^{(l,p)} + w_{\mathrm{HC}}^{(l,p)}) \quad , \tag{6}$$

$$\mathbf{e}^{(l)} = \sum_{p=0}^{2} \alpha_{\mathrm{CE}}^{(l,p)} w_{\mathrm{CE}}^{(l,p)} \quad , \tag{7}$$

where $\alpha_{\mathrm{CE}}^{(l,p)}$ is calculated using the same self-attention method as Eq. 2.

3.2 Clinical Text Encoder

A sequence of words from electronic medical records is transformed into word embedding via the same Word2Vec with embedding size d_e. Assuming the number of words N_w, the input word embedding matrix can be written as

$$\mathbf{X}_w = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{N_w}] \in \mathbb{R}^{N_w \times d_e} \quad . \quad (8)$$

Then we use a Bidirectional LSTM to capture the contextual information of the words. Although the Transformer-based models have taken a great leading place in various NLP applications in recent years, they are not that favorable for this task. We argue this point for the following reasons and verify it with empirical evidence: Unlike other NLP tasks, its vocabulary is domain-specific and thus low-dimensional. Using a Transformer-based encoder may add difficulty and redundancy to the training, costing more time and space. Besides, the sentences in EMR-like documents are not necessarily long and quite concentrated in their meanings. The long dependency issue is not very phenomenal in this case.

Finally, we compute the document representation by concatenating the output $\overrightarrow{\mathbf{h}_i}$ of $\overrightarrow{\mathrm{LSTM}}$ and $\overleftarrow{\mathbf{h}_i}$ of $\overleftarrow{\mathrm{LSTM}}$ for word $\mathbf{x}_i$. All the representations of words in the document formulate the document representation $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_{N_w}] \in \mathbb{R}^{N_w \times 2u}$ with $\mathbf{h}_i = \overrightarrow{\mathbf{h}_i} \oplus \overleftarrow{\mathbf{h}_i}$ and u the hidden size.

3.3 Code-Text Attention Block

After the above modules, we obtain a text representation $\mathbf{H} \in \mathbb{R}^{N_w \times 2u}$, introduced in Section 3.2, and the code representations $\mathbf{e} = [\mathbf{e}^{(1)}, \mathbf{e}^{(2)}, ..., \mathbf{e}^{(L)}] \in \mathbb{R}^{L \times d_c}$ in Section 3.1. The traditional label-wise attention (Vu et al., 2020; Mullenbach et al., 2018) generate the attention weights from text encoding H. Here we propose a more code-constrained version, involving both text and code representations:

$$\mathbf{A} = \mathrm{softmax}(\mathbf{e} \cdot \tanh(\mathbf{H}W_H)) \quad , \quad (9)$$

$$\mathbf{V} = \mathbf{A} \cdot \mathbf{H} \quad , \quad (10)$$

where W_H is a matrix $\in \mathbb{R}^{2u \times d_c}$ which maps the document representation to the code embedding space to avoid dimension mismatch. $\mathbf{A} \in \mathbb{R}^{L \times N_w}$ denotes a code-specific weight, which is measured by how similar the documents are represented with each code. Afterwards, $\mathbf{A}$ is multiplied with the document $\mathbf{H}$ to generate a code-specific representation $\mathbf{V} \in \mathbb{R}^{L \times 2u}$.

3.4 Output Layer

With the code-specific representation $\mathbf{V}$, we calculate the probability for each code by passing each vector $\mathbf{V}_l \in \mathbb{R}^{2u}$ into a fully connected layer following an activation function $\mathrm{sigmoid}$ to produce the binary prediction score for code l. The training objective is to minimise the binary cross entropy between the prediction score $\hat{y}_l$ and target y_l:

$$\sum_{l \in \mathbb{L}} -y_l \log(\hat{y}_l) - (1 - y_l) \log(1 - \hat{y}_l) \quad . \quad (11)$$

4 Experiments

4.1 Dataset

The Medical Information Mart for Intensive Care III, denoted MIMIC-III, is a large open-source dataset (Johnson et al., 2016), containing the medical records of over forty thousand patients in the Beth Israel Deaconess Medical Center between 2001 and 2012. In terms of our problem, we use specifically the discharge summaries and their corresponding ICD-9 codes. Following the previous works (Shi et al., 2017; Mullenbach et al., 2018), we rearrange the data, where in total 52,726 discharge summaries with 8,929 ICD-9 codes served as labels are formulated. We split the dataset with the settings of Mullenbach et al. (2018). The data statistics are shown in Table 1.

MIMIC-III-Full			
	Train	Dev	Test
# Doc.	47,723	1631	3372
Avg words per Doc.	1484	1785	1792
Avg codes per Doc.	16.1	23.2	20.9

Table 1: Statistics of MIMIC-III-Full dataset.

4.2 Implementation Details

We follow the preprocessing schema of Vu et al. (2020) in our experiments. The CBOW is utilized during the pretraining stage, with the embedding size $d_e = 128$ on the processed text. All the documents are truncated with maximum 4,000 words. A data augmentation strategy (**w/ Sentence Permutation**) are applied in the experiment, where multiple sentences in the same document are shuffled in a random order to generate a new sample for training (Kim and Ganapathi, 2021). In our experiment, we use the 3-fold augmentation, i.e. increasing the training set three times. To show that our method

Model	AUC		F1		P@k	
	Macro	Micro	Macro	Micro	k=8	k=15
CAML (Mullenbach et al., 2018)	89.5	98.6	8.8	53.9	70.9	56.1
DR-CAML (Mullenbach et al., 2018)	89.7	98.5	8.6	52.9	69.0	54.8
TransICD (Biswas et al., 2021)	89.7	98.5	8.4	51.1	67.9	53.3
MultiResCNN (Li and Yu, 2020)	91.0	98.6	8.5	55.2	73.4	58.4
HyperCore (Cao et al., 2020a)	93.0	98.9	9.0	55.1	72.2	57.9
LAAT (Vu et al., 2020)	91.9	98.8	9.9	57.5	73.8	59.1
JointLAAT (Vu et al., 2020)	92.1	98.8	10.7	57.5	73.5	59.0
MSMN (Yuan et al., 2022)	95.0	**99.2**	10.3	58.4	75.2	59.9
Baseline (Label-wise attention)	89.4	98.6	9.0	56.2	73.9	58.8
Ours	95.0	**99.2**	10.3	58.0	**75.3**	59.9
Ours w/ Sentence Permutation	**95.2**	**99.2**	**10.8**	58.2	75.1	59.9
Ours w/ Enriched Descriptions	**95.2**	**99.2**	**10.8**	**58.6**	**75.3**	**60.3**

Table 2: Results on MIMIC-III-full, i.e. all codes. We compare our models with all the baselines by their values reported in the original papers for overall metrics.

can be complementary with other existing methods, we conduct experiments with the synonyms enriched descriptions (**w/ Enriched Descriptions**) in MSMN (Yuan et al., 2022) as well.

For the text encoder, we set LSTM hidden size as 256 with 2 layers. The dropout rate is 0.3. We add an extra linear layer after LSTM output with the dimension of 256. Our model is trained with AdamW (Loshchilov and Hutter, 2019) at a learning rate of $1e^{-3}$ on a single NVIDIA Tesla A100 (40GB). The batch size is set as 32. The early stopping mechanism is applied, in which the training will be stopped if there is no improvement of the micro-F1 score on the validation set in ten successive epochs. We run the experiment with 3 random seeds and report the average.

4.3 Evaluation Metrics

To make a comparison with other previous works on ICD-9 prediction, we evaluate the model performance by **F1**, **AUC** (area under the ROC curve) and **P@k**. F1 and AUC are calculated in two manners: macro-averaged, i.e. a simple average of all labels, and micro-averaged, i.e. aggregating the contributions of all classes to compute the average. P@k denotes the precision of top-k predicted results. K is conventionally set as the average number of labels for each document. However, it is not applicable for this task, since the number of labels for each document varies widely.

The macro-averaged metrics cannot thoroughly represent the rare code performance due to the huge gap in prediction performance between rare codes and frequent codes. Our experiment is designed to validate the idea that by enhancing the relations between frequent codes and rare codes, our method can achieve better performance on rare codes. Considering the insufficiency of distinguishing overall results and rare code performances, we select the codes with label frequency in the training set between 2-10 as rare codes, and report the results on predicting them. Since the macro- and micro- averaged results are very similar under this rare code setting (due to the similarity in label frequency), we only report their micro metrics.

4.4 Baselines

Our model is compared against the following SOTA models, chosen by their task settings:

CAML (Mullenbach et al., 2018), i.e. the Convolutional Attention network for Multi-Label classification, utilizes CNN as text encoder and propose a label attention for prediction. Meanwhile, they propose a **DR-CAML** version, where the ICD-9 code descriptions are used for regularization to improve the performance on rare codes, with a similar purpose to ours.

TransICD (Biswas et al., 2021) adopts an Transformer Encoder for discharge summaries.

MultiResCNN (Li and Yu, 2020) encodes the text with multi-filter residual CNN.

HyperCore (Cao et al., 2020a) takes also code co-occurrence and hierarchy into consideration. It embeds both code and text into hyperbolic space and calculates their similarities.

LAAT (Vu et al., 2020) applies also LSTM text encoder. A hierarchical joint structure, i.e. **JointLAAT**, is proposed to solve the imbalance label distribution, thus ameliorate the performances on rare codes.

MSMN (Yuan et al., 2022) enriches the code descriptions with synonyms from UMLS and encodes the clinical text with Bi-LSTM.

Baseline is designed by replacing the code-text attention and the relation-enhanced code encoder with label-wise attention (Vu et al., 2020).

4.5 Results

MIMIC-III-Full. Table 2 shows the results on MIMIC-III for all codes, where our model outperforms all the other baselines. By adopting the code descriptions with the synonyms from UMLS, our model performs better and outperforms MSMN in almost all metrics. With the simple code description, our method still produces comparable, even better results without the necessity of bringing external knowledge source. Moreover, the large margin between baseline and ours validates the idea that by enhancing the inter-code relations, model can produce better results than traditional label-wise attention method.

Rare Codes. We collect all the codes with an appearing times in the training set between 2 and 10, and observe specifically the model performance on these codes. As shown in Table 3, our model achieves the best results among all the baselines. With Sentence Permutation, the improvements are more significant. It is interesting to observe that CAML and DR-CAML mess up all the predictions for these rare codes. Though JointLAAT is proposed at the intention of improving its few-shot performances, where the final prediction is based on the prediction on the codes starting with the same first three characters, the results actually degrade compared with LAAT. It is because of a low recall, since excluding the first level codes affects the prediction on its child codes as well. Our model exploits the inter-code relations from co-occurrence and enhances them for rare codes via hierarchy, thus producing better results and leaving lower margin to the overall results.

4.6 Adjacency Matrix

It is commonly recognized that more training data can lead to a better performance. However, for clinical documents, collecting them and tagging the ICD codes can be quite difficult due to its privacy and difficulty of processing. In our model, we apply a GCN module with adjacency matrix (ADJ) based on co-appearing relations. The above experiments are conducted with ADJ sampled from training set, which is denoted "Training". We are wondering if the performances can be further ameliorated with ADJ of more samples, or derived from some prior medical knowledge. This idea is analysed by proposing another two ADJs. "Full" denotes the ADJ sampled from the full dataset, including training, validation and testing sets. Besides, we add the co-appearing relations in MIMIC-IV (Johnson et al., 2021) as well and denote this matrix "w/ MIMIC-IV".

Table 4 shows the results on full codes and rare codes with various adjacency matrices. Since the macro AUC is not very sensitive, we just list the macro/micro F1 and micro AUC. We notice that having a more complete and reasonable adjacency matrix can help the model prediction, since both of the F1 metrics get better.

Model	Micro AUC	Micro F1
CAML	50.0	0.00
DR-CAML	50.0	0.00
TransICD	79.9	5.54
MultiResCNN	81.7	0.65
LAAT	88.8	3.23
JointLAAT	87.2	0.79
Baseline	80.4	4.23
Ours	**94.5**	6.72
Ours w/ SP	**94.5**	**7.15**

Table 3: Results on rare codes, i.e. codes whose frequency in training set is between 2-10. Since this metric is not considered in the original papers for baselines, we select the models with released codes and re-implement their experiments.

MIMIC-III-Full			
	F1		AUC
	Macro	Micro	Micro
Training	10.3	58.0	**99.2**
Full	10.6	**58.2**	99.1
w/ MIMIC-IV	**10.7**	**58.2**	99.1
Rare Codes			
	F1		AUC
Training	6.72		94.6
Full	7.13		**94.8**
w/ MIMIC-IV	**7.35**		94.2

Table 4: Influence of different adjacency matrices.

4.7 Model Interpretability

Being able to explain the model decision with conformance to human understanding is an important criterion in healthcare. To prove the effectiveness of our model on rare code prediction, we provide in Figure 4 some spans from the same discharge summaries where the tokens with high attention scores locate for predicting *Pneumonia in other systemic mycoses* (484.7), an infrequent code with only 17 documents.

We notice that the baseline model puts high attentions to the words like "discharge", "date" and "marrow", whose relevance with *Pneumonia in other systemic mycoses* is hard to find because almost all the documents contain them. Our model can capture the closely related words like "pneumonia", "aspergillus" and "lobe", indicating a better code representation.

admission date dis harge date date of irth sex
m ser i e medi ine allergies no no n allergies
ad erse dr g rea tions attending rst name3
lf hief om laint fatig e shortness of reath
ma or s rgi al or in asi e ro ed re right i
la ement entral line la ement one marro
io sy left s la ian entral line la ement
one marro io sy left i la ement left
internal g lar entral line la ement one
marro io sy one marro io sy one
marro io sy ron hos o y one marro
io sy on er taneo s hole ystostomy t e
la ement i la ement on history of resent

er lo e ne monia rogressed from m lti le
lmonary nod les as des ri ed on t of as
stated on rior re ort these an e follo ed
ith a t hest ithin fo r ee s or the
ossi ility of io sy an e onsidered oronary
artery disease t hest im ression right a i al
onsolidation and t o left er lo e nod les
ha e not hanged sin e the most re ent s an t
right lo er lo e nod les ha e im ro ed o erall
a earan e is most onsistent ith an a te
infe tio s ro ess either f ngal e g as ergill s
or a terial in etiology ry togeni organi ing
ne monia may also ha e a similar imaigng
a earan e

Figure 4: The spans with high attentions (darker means higher) of the same text for predicting *Pneumonia in other systemic mycoses* (484.7 with 17 documents). The red shows the interpretation for the baseline and the green for ours.

5 Ablation Study

We conduct the ablation study concerning the effectiveness of the contextual transformation, code hierarchy and code co-occurrence modules inside the relation-enhanced code encoder. We measure the F1 metrics, which are more sensitive and representative to different models, between our ordinary version and those with a module removed. The results without the entire code encoder, which means using only the embeddings of code descriptions as query, are also listed.

As shown in Table 5, removing the transformation module causes a significant decrease, indicating that the gap we described between descriptions and discharge summaries do exist. Besides, we notice that removing the code hierarchy or the code co-occurrence module has also degraded the model performances, showing that the code hierarchy and co-occurrence are useful to the model. Since the three GCNs may bring extra training difficulties, further tuning the training process might still be helpful to improve the performances.

Model	F1	
	Macro	Micro
Ours	**10.3**	**58.0**
w/o transformation	9.8	57.2
w/o hierarchy	10.2	57.9
w/o co-occurrence	10.1	57.7
w/o code encoder	7.5	51.0

Table 5: Results of ablation study on all codes.

6 Conclusions

The multi-label clinical text classification is an important task in the domains of both healthcare and natural language processing. In this paper, we reveal the existing problem of traditional methods in capturing the inter-code relations for rare codes. Hereby, we propose to strengthen the relations, thus improving the model performance on rare code prediction. We exploit the inter-code relations by encoding code descriptions and incorporating co-occurrence under a code-organ-system hierarchical structure in order to enhance the connections for rare codes. Our model is then evaluated on the commonly used MIMIC-III dataset and outperforms the other baselines on both rare codes and full codes. The visualisations further demonstrate the advantage of our method on providing more human-understandable explanations. We conduct as well an analysis concerning the design of adjacency matrices and the ablation study to better understand the different components in our method.

Limitations

In this work, we adopt three GCNs to exploit the inter-code relations under different levels. However, this may bring extra training difficulty and the risk of over-parameterization to the model. Besides, during the preprocessing stage, we adopt a word-level tokenizer and CBOW to obtain their embeddings for MIMIC-III texts and code descriptions. However, this might not be enough to represent the words since medical documents have some special characteristics, but we do not take them into consideration. We tried in our work with other pretraining strategies, such as ClinicalBERT (Alsentzer et al., 2019), BERT (Devlin et al., 2019), BioBERT (Lee et al., 2020) and BioWordVec (Zhang et al., 2019). We added as well the BPE tokenizer (Sennrich et al., 2016) in order to capture the meaningful medical sub-word units. However, the results are all far from satisfactory.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Biplob Biswas, Thai-Hoang Pham, and Ping Zhang. 2021. Transicd: Transformer based code-wise attention model for explainable ICD coding. In *Artificial Intelligence in Medicine - 19th International Conference on Artificial Intelligence in Medicine, AIME 2021, Virtual Event, June 15-18, 2021, Proceedings*, volume 12721 of *Lecture Notes in Computer Science*, pages 469–478. Springer.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Aréchiga, and Tengyu Ma. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 1565–1576.

Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong. 2020a. HyperCore: Hyperbolic and co-graph representation for automatic ICD coding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3105–3114, Online. Association for Computational Linguistics.

Pengfei Cao, Chenwei Yan, Xiangling Fu, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong. 2020b. Clinical-coder: Assigning interpretable ICD-10 codes to Chinese clinical notes. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 294–301, Online. Association for Computational Linguistics.

Luciano R. S. de Lima, Alberto H. F. Laender, and Berthier A. Ribeiro-Neto. 1998. A hierarchical approach to the automatic categorization of medical documents. In *Proceedings of the 1998 ACM CIKM International Conference on Information and Knowledge Management, Bethesda, Maryland, USA, November 3-7, 1998*, pages 132–139. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ruiying Geng, Binhua Li, Yongbin Li, Jian Sun, and Xiaodan Zhu. 2020. Dynamic memory induction networks for few-shot text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1087–1094, Online. Association for Computational Linguistics.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Leo Anthony Celi, Roger Mark, and S Horng IV. 2021. Mimic-iv-ed. *PhysioNet*.

Alistair Johnson, Tom Pollard, and R Mark III. 2016. Mimic-iii clinical database. *Physio Net*, 10:C2XW26.

Byung-Hak Kim and Varun Ganapathi. 2021. Read, attend, and code: Pushing the limits of medical codes prediction from clinical notes by machines. In *Proceedings of the Machine Learning for Healthcare Conference, MLHC 2021, 6-7 August 2021, Virtual Event*, volume 149 of *Proceedings of Machine Learning Research*, pages 196–208. PMLR.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Bevan Koopman, Guido Zuccon, Anthony N. Nguyen, Anton Bergheim, and Narelle Grayson. 2015. Automatic ICD-10 classification of cancers from free-text death certificates. *Int. J. Medical Informatics*, 84(11):956–965.

Leah S. Larkey and W. Bruce Croft. 1996. Combining classifiers in text categorization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'96, August 18-22, 1996, Zurich, Switzerland (Special Issue of the SIGIR Forum)*, pages 289–297. ACM.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.*, 36(4):1234–1240.

Fei Li and Hong Yu. 2020. ICD coding from clinical text using multi-filter residual convolutional neural network. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8180–8187. AAAI Press.

Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. 2017. Meta-sgd: Learning to learn quickly for few shot learning. *CoRR*, abs/1707.09835.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2999–3007. IEEE Computer Society.

Yang Liu, Hua Cheng, Russell Klopfer, Matthew R. Gormley, and Thomas Schaaf. 2021. Effective convolutional attention network for multi-label clinical document classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5941–5953, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Carlos Medina, Arnout Devos, and Matthias Grossglauser. 2020. Self-supervised prototypical transfer learning for few-shot classification. *CoRR*, abs/2006.11325.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana. Association for Computational Linguistics.

Kimberly J O'malley, Karon F Cook, Matt D Price, Kimberly Raiford Wildes, John F Hurdle, and Carol M Ashton. 2005. Measuring diagnoses: Icd code accuracy. *Health services research*, 40(5p2):1620–1639.

Adler J. Perotte, Rimma Pivovarov, Karthik Natarajan, Nicole Gray Weiskopf, Frank D. Wood, and Noemie Elhadad. 2014. Diagnosis code assignment: models and evaluation metrics. *J. Am. Medical Informatics Assoc.*, 21(2):231–237.

John P. Pestian, Chris Brew, Pawel Matykiewicz, DJ Hovermale, Neil Johnson, K. Bretonnel Cohen, and Wlodzislaw Duch. 2007. A shared task involving multi-label classification of clinical free text. In *Biological, translational, and clinical language processing*, pages 97–104, Prague, Czech Republic. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Haoran Shi, Pengtao Xie, Zhiting Hu, Ming Zhang, and Eric P. Xing. 2017. Towards automated ICD coding using deep learning. *CoRR*, abs/1711.04075.

Congzheng Song, Shanghang Zhang, Najmeh Sadoughi, Pengtao Xie, and Eric P. Xing. 2020. Generalized zero-shot text classification for ICD coding. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 4018–4024. ijcai.org.

Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, 3(1):1–10.

Shang-Chi Tsai, Chao-Wei Huang, and Yun-Nung Chen. 2021. Modeling diagnostic label correlation for automatic ICD coding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4043–4052. Association for Computational Linguistics.

Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2016. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3630–3638.

Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2020. A label attention model for ICD coding from clinical text. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3335–3341. ijcai.org.

Aaron Sonabend W., Winston Cai, Yuri Ahuja, Ashwin N. Ananthakrishnan, Zongqi Xia, Sheng Yu, and Chuan Hong. 2020. Automated ICD coding via unsupervised knowledge integration (UNITE). *Int. J. Medical Informatics*, 139:104135.

Shanshan Wang, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Huasheng Liang, Qiang Yan, Evangelos Kanoulas, and Maarten de Rijke. 2021. Few-shot electronic health record coding through graph contrastive learning. *CoRR*, abs/2106.15467.

Pengtao Xie and Eric Xing. 2018. A neural architecture for automated ICD coding. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1066–1076, Melbourne, Australia. Association for Computational Linguistics.

Xiancheng Xie, Yun Xiong, Philip S. Yu, and Yangyong Zhu. 2019. EHR coding with multi-scale feature attention and structured knowledge graph propagation. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 649–658. ACM.

Yan Yan, Glenn Fung, Jennifer G. Dy, and Rómer Rosales. 2010. Medical coding classification by leveraging inter-code relationships. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, 2010*, pages 193–202. ACM.

Zheng Yuan, Chuanqi Tan, and Songfang Huang. 2022. Code synonyms do matter: Multiple synonyms matching network for automatic ICD coding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 808–814, Dublin, Ireland. Association for Computational Linguistics.

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. 2019. Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific data*, 6(1):1–9.

NewAgeHealthWarriors at MEDIQA-Chat 2023 Task A: Summarizing Short Medical Conversation with Transformers

Prakhar Mishra
IIIT Bangalore, India
prakhar.mishra@iiitb.org

Ravi Theja Desetty
IIIT Bangalore, India
ravi.theja@iiitb.org

Abstract

This paper presents the MEDIQA-Chat 2023 shared task organized at the ACL-Clinical NLP workshop. The shared task is motivated by the need to develop methods to automatically generate clinical notes from doctor-patient conversations. In this paper, we present our submission for *MEDIQA-Chat 2023 Task A: Short Dialogue2Note Summarization*. Manual creation of these clinical notes requires extensive human efforts, thus making it a time-consuming and expensive process. To address this, we propose an ensemble-based method over GPT-3, BART, BERT variants, and Rule-based systems to automatically generate clinical notes from these conversations. The proposed system achieves a score of 0.730 and 0.544 for both the sub-tasks on the test set (ranking 8th on the leaderboard for both tasks) and shows better performance compared to a baseline system using BART variants.

1 Introduction

Telecare has experienced an exponential increase in utilization since the onset of the COVID-19 pandemic, leading to the emergence of a vast network of healthcare providers and patients Garfan et al. (2021). We consider this to be a significant use case within the Telecare domain, medical personnel often need to provide a concise summary of the conversation they had with their patient in order to ensure that a colleague is able to follow up on the next consultation. Both patients and medical professionals can use these summaries to refer back to their interactions in the future. Unfortunately, manually creating conversation notes after each encounter consumes a significant amount of time, and energy, and also poses challenges when done at scale.

Recent advancements in Natural Language Processing (NLP) with large language models (LLMs) like GPT-3 have shown promising results in their ability to generate convincing natural language and successfully solve tasks including classification, answering questions, and summarization even in zero-shot and few-shot environments Brown et al. (2020). This makes them a popular choice as opposed to a pre-trained model, which needs to be adjusted separately for each downstream task. In this paper, we propose an ensemble of rule-based methods, traditional sequence models, large language models, and BERT-based models to develop an automated system for generating these notes from doctor-patient conversations. We also show that few-shot large language models outperform traditional sequence-to-sequence models in the setting of limited data.

2 Related Work

Summarization is a crucial task in NLP, particularly for extracting key information from multi-speaker conversations. Various approaches have been proposed for meeting summarization, such as DialogLM Zhong et al. (2022), a pre-trained neural encoder-decoder model. In the context of medical dialogues between doctors and patients, identifying symptoms, diagnoses, and treatments is essential for deriving a medical solution. Song et al. (2020) introduced the hierarchical encoder-tagger model (HET) to specifically identify important utterances in medical conversations for summarizing medical conversations. Krishna et al. (2021) introduced pointer generator networks for deep summarization of physician-patient dialogues. Joshi et al. (2020) introduced a variant of the pointer generator network that handles negations and imposes a penalty on the generator distribution and Zhang et al. (2021) fine-tuned BART models for summarizing doctor-patient interactions.

3 Task and Dataset Details

The MEDIQA-Chat 2023 Ben Abacha et al. (2023a) shared task has been developed to foster research in the field of automatic clinical note

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 414–421
July 14, 2023

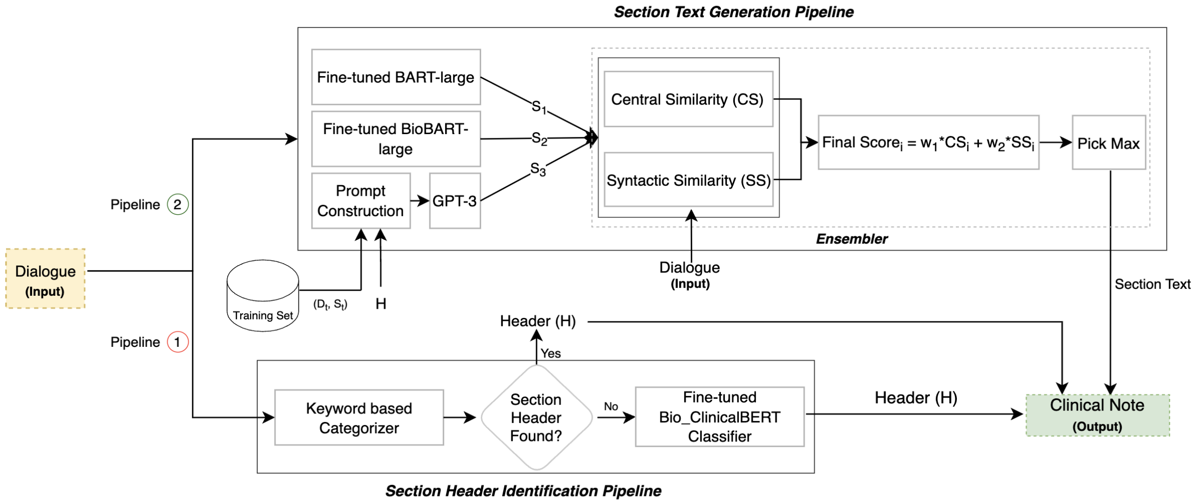


Figure 1: Pipeline for conversation to clinical note generation. Here, S_i, D_t, S_t, CS_i, and SS_i are section texts generated from independent models, dialogue from training data, section text from training data, central similarity score for S_i, syntactic similarity score for S_i respectively.

generation derived from doctor-patient conversations. It comprises three tasks[1], namely, Short Dialogue2Note Summarization (Task A) Ben Abacha et al. (2023b), Full Dialogue2Note Summarization (Task B) wai Yim et al. (2023), and Note2Dialogue Generation (Task C) wai Yim et al. (2023). Our work primarily focuses on Task A. This task requires us to create a section summary, encompassing both the section header and text, based on a short input of a doctor-patient conversation.

The dataset consists of 20 distinct section headers for each conversation, such as Medications, Review of Systems, Past Surgical History, Chief Complaint, etc. The training set contains 1,201 pairs of conversations, each accompanied by their relevant section headers and text, while the validation set is composed of 100 pairs of conversations and their respective summaries. Table 1 shows statistics around the train/val/test data splits. Table 3 shows a few snippets of actual training data containing Section Header, Section Text, and doctor-patient conversation.

4 System Description

In this section, we give a detailed explanation of our proposed system[2]. We propose two separate pipelines - one for section header identification and the other for section text generation. We also discuss our ensemble strategy and related intuition.

Data	Dialogue len	Sec. len	# Samples
Train	105.6	40.5	1201
Val	89.9	36.0	100
Test	100.0	-	200

Table 1: Dataset Statistics. Dialogue len, and Sec. len denotes the average number of words at the Dialogue and section level respectively.

Category	Coverage Text
Allergy	Incase of no allergies, reply with keyword 'no known allergies'
Fam/Sochx	Incase of no family medical history found, reply with keyword 'noncontributory'
Genhx	Don't forget to mention age and gender of the patient, if present.

Table 2: Examples of Coverage Text

Finally, we output results from both pipelines to generate final summaries.

Section Header Identification: The task of Section Header Identification involves categorizing a given doctor-patient conversation to the relevant header from a list of pre-defined headers. Table 4 lists down all the available headers along with their expanded form which we received as a part of the task description.

We developed a 2-step strategy for detecting the accurate section header for a given doctor-patient conversation. Figure 1 shows the inference flow

[1]Task Page: https://github.com/abachaa/MEDIQA-Chat-2023

[2]Code: https://github.com/prakhar21/MEDIQA-CHAT-2023-NewAgeHealthWarriors

section_header	section_text	dialogue
PASTMEDICALHX	Asthma.	**Doctor**: How's your asthma since you started using your inhaler again? **Patient**: Much better. I don't know why I didn't take it with me everywhere I went. **Doctor**: It's important to carry it with you, especially during times where you're exercising or walking more than usual. **Patient**: Yeah. I think I've learned my lesson. **Doctor**: Besides asthma, do you have any other medical problems?
CC	Burn, right arm.	**Doctor**: Hi, how are you? **Patient**: I burned my hand. **Doctor**: Oh, I am sorry. Wow! **Patient**: Yeah. **Doctor**: Is it only right arm? **Patient**: Yes.
FAM/SOCHX	His brothers had prostate cancer. Father had brain cancer. Heart disease in both sides of the family. Has diabetes in his brother and sister.	**Doctor**: Can you tell me about any diseases that run in your family? **Patient**: Sure, my brother has a prostate cancer. **Doctor**: Okay, brother. **Patient**: My father had brain cancer. **Doctor**: Okay, dad. **Patient**: Then on both sides of my family there are many heart related issues. **Doctor**: Okay. **Patient**: And my brother and sister both have diabetes. **Doctor**: Okay. **Patient**: Yes, that's it.

Table 3: Sample data from training set

of the Section Header Identification pipeline. In step-1, we categorize a given conversation to its section header using our Keyword lookup list. We refer to this as 'Keyword based Categorizer' in the diagram. We manually curated this list by going through many examples of section texts for every section header from the training data. If no section header is identified in this step, we pass the same conversation to the Bio_ClinicalBERT Alsentzer et al. (2019) model, variant of BERT Devlin et al. (2018), which we had fine-tuned on our dataset of conversation and section header pairs. Please refer to Section 5 and 6 for more details on model description, implementation, and results.

Section Text Generation: The task of Section Text Generation involves generating a summary of the given doctor-patient conversation. We propose an ensemble of 3 transformer-based models, i.e, BART-large, BioBART-large, and few-shot GPT-3 for the same. Here, we fine-tune BART-large Lewis et al. (2019) and BioBART-large Yuan et al. (2022) transformer models in a sequence-to-sequence paradigm on our training dataset. The models were fine-tuned with input as doctor-patient conversations and output as associated section text with the training objective of maximizing the likelihood of the generated summary.

For GPT-3, we adopt a few-shot prompt engineering-based Liu et al. (2023) approach for generating our section text. Few-shot prompting helps enable in-context learning for large language models like GPT-3. Figure 2 shows a detailed annotation of the GPT-3 prompt that we use for our purpose. In the figure, *<Dialogue Example>* is an example dialogue from the training dataset that we sample randomly based on the predicted section header(*<Section Header>*) on the *<Test Dialogue>* sequence. The intuition behind adding this extra knowledge to our prompt was to help our model learn the writing style of actual section text. We also experimented by giving multiple examples of dialogue, section header, and section text as a part of our prompt. Please refer to Section 6 for more details on experiments.

During our initial analysis of the training dataset, we observed that there were certain specific writing

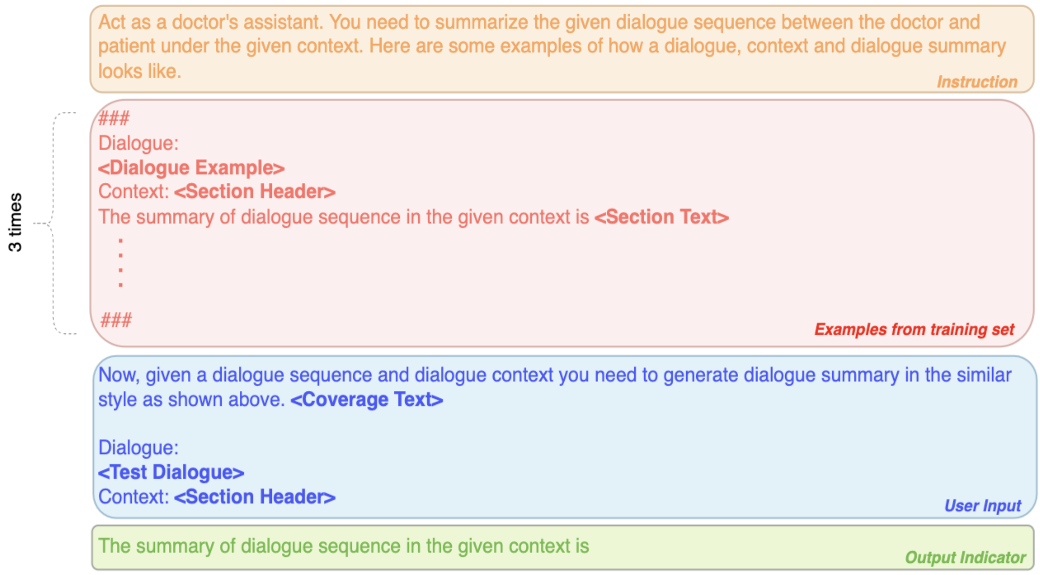


Figure 2: GPT-3 prompt structure

Header	Full Header
Fam/Sochx	Family History/Social History
Genhx	History of Present Illness
Pastmedicalhx	Past Medical History
Cc	Chief Complaint
Pastsurgical	Past Surgical History
Allergy	
Ros	Review of Systems
Medications	
Assessment	
Exam	
Diagnosis	
Disposition	
Plan	
Edcourse	Emergency Department Course
Immunizations	
Imaging	
Gynhx	Gynecologic History
Procedures	
Other_history	
Labs	

Table 4: List of all section headers and their full forms

patterns present in the section text. For example - in numerous cases, the section text for 'History of Present Illness' starts with the patient's age and gender. There were also cases where section text had words like 'Noncontributory', 'None', etc. To accommodate such a writing style in the model's output, we added an additional text called *<Coverage Text>* for some of the selected section headers. We have shown some examples of coverage text in Table 2.

Figure 1 shows the inference flow of the Section Text Generation pipeline. Once the section text from each of the models is generated, we score all of them and pick the one with the maximum score as our choice of final generated section text. For each section text, we calculate a final score based on a weighted scoring scheme that combines both central and syntactic similarity scores with weights w_1 and w_2 respectively. We found $w_1 = w_2 = 0.5$ to work best for our use case. For calculating central similarity for each of the generated section text, we implement the work done in Kobayashi (2018) and for syntactic similarity, we calculate the token-level Jaccard similarity between the section text and input dialogue. Jaccard similarity is defined as the ratio between the intersection of two sets and the union of two sets, and it is often used as a metric of similarity. Intuitively, section text that centrally captures the majority theme across all the generated section texts will have a high central similarity score, whereas, the syntactic similarity would help ensure faithfulness. Goel et al. (2021)

Dialogue	AS	GS	R1
Doctor: Do you drink alcohol or smoke cigarettes? **Patient**: No, I do not. **Doctor**: Are you sure? **Patient**: Yes.	Denies the use of alcohol or tobacco.	Denies the use of alcohol or tobacco.	1.0
Doctor: Have you ever had surgery? **Patient**: One too many times. **Doctor**: Which ones? **Patient**: I had my appendix taken out and glaucoma surgery fairly recently. I also had my gallbladder taken out ten years ago and a partial colon resection due to colon cancer in nineteen sixty one. **Doctor**: Any recurring episodes of colon cancer? **Patient**: No, thankfully.	Partial colon resection of colon carcinoma in 1961 with no recurrence, cholecystectomy 10 years ago, appendectomy, and glaucoma surgery.	Appendectomy and glaucoma surgery. Cholecystectomy 10 years ago and partial colon resection due to colon cancer in 1961.	70.3
Doctor: Any difficulty in hearing? **Patient**: No. **Doctor**: Difficulty swallowing? **Patient**: Um no. **Doctor**: Any double vision or blurred vision or difficulty seeing things properly? **Patient**: No, no problem at all. **Doctor**: Okay. **Doctor**: How about headaches or migraine? **Patient**: No headache. **Doctor**: Did you notice any change in your bowel moment? **Patient**: No, it is the same. **Doctor**: Any pain while urinating or change in frequency? **Patient**: No. **Doctor**: Okay.	No headaches. No visual, hearing, or swallowing difficulties. No changes in bowel or urinary habits.	NEUROLOGICAL: No difficulty in hearing, swallowing, double vision, blurred vision, headaches, or migraines. GASTROINTESTINAL: No change in bowel movements or difficulty urinating.	48.7
Doctor: It seems like you are not feeling very well today? **Patient**: Yeah. I have had diarrhea and pain in my stomach. **Doctor**: Have you experienced any vomiting? **Patient**: Yes. I threw up this morning."	Diarrhea, vomiting, and abdominal pain.	The patient states that he is not feeling very well today. He has had diarrhea and pain in his stomach this morning, and he has had vomiting this morning.	23.5
Doctor: I will do some examinations on you. I will check your chest and then I will talk to you as I move forward, okay? **Patient**: I'm okay with that. **Doctor**: So, let's see what we have here. Hm, Yeah, just looks good. I do not find anything abnormal.	CHEST: The chest examination is unremarkable.	Chest x-ray without any abnormality.	16.7
Guest_clinician: I did a review of her systems, and everything looks normal other than what was mentioned earlier. **Doctor**: Okay, thanks for your help. **Guest_clinician**: No problem.	The remaining ROS is unremarkable.	Review of Systems: Everything appears to be normal other than what was mentioned earlier.	0.0
Doctor: Are you allergic to anything, food or medicines? **Patient**: No allergies that I know of.	None.	No known drug allergies.	0.0

Table 5: Examples are arranged in decreasing order of R1. Here, AS, GS, and R1 refer to the actual, generated section texts, and Rouge-1 respectively.

defines a faithful summary to be one that contains minimal information outside the source text. Finally, output from both pipelines is used to report the final generated clinical note.

5 Model Background

In this section, we discuss in brief the background of various machine learning models that we have used in our implementation.

Bio_ClinicalBERT: The Bio_ClinicalBERT[3] model is initialized from BioBERT Lee et al. (2020) and trained on all notes from MIMIC III Johnson et al. (2016), a database containing electronic health records from ICU patients. The model was pre-trained on a GeForce GTX TITAN X 12 GB GPU, on a batch size of 32, a maximum sequence length of 128, and a learning rate of 5x10-5.

BART & BioBART: BART Lewis et al. (2019) [4] is a transformer encoder-decoder (seq2seq) model with a bidirectional (BERT-like) encoder and an autoregressive decoder to perform complex NLG tasks like summarization, translation, etc. BART is pre-trained to reconstruct the original text from the noisy text. BioBART Yuan et al. (2022) [5], a BART variant was pre-trained on PubMed abstracts to achieve biomedical domain adaption.

GPT-3: GPT-3 Brown et al. (2020) is an autoregressive language model with 175 billion parameters. It achieves strong performance on many NLP datasets, including translation, and question-answering, as well as several tasks that require on-the-fly reasoning or domain adaptation.

6 Experiments and Results

In this section, we discuss experiments, implementation details, and results obtained for both our tasks.

Section Header Identification: We trained Fasttext Joulin et al. (2016) along with multiple BERT variants like BERT-base Devlin et al. (2018), RoBERTa Liu et al. (2019), and Bio_ClinicalBERT Alsentzer et al. (2019) for identifying section headers from doctor-patient dialogues. We use simpletransformers Rajapakse (2019) python library for fine-tuning all our transformer models. Amongst all of them, Bio_ClinicalBERT gave us the best score on the validation dataset. Our final model also incorporates a Keyword-based categorizer in the pipeline giving us the best accuracy of 77% on the validation set. We use weighted cross-entropy as our loss function because of the skewed distribution of headers in the training data. The corresponding weights per header category were calculated using sklearn's *compute_class_weight* function on the training dataset and we train our best model for 10 epochs.

Section Text Generation: We fine-tuned BART, BioBART architectures and, also inferred GPT-3 model in a few-shot setting. Interestingly, GPT-3 in the few-shot setting outperforms all our fully supervised models by **1.2+** Rouge-Avg points (Refer Table 7). Rouge-Avg(RA) is the average score of Rouge-1 (R1), Rouge-2 (R2), Rouge-L (RL), and RougeLSum (RLS). We train both our BART-large and BioBART-large models for 5 epochs, set a beam size of 5 while decoding the sequence, and use the cross-entropy loss as our objective function. We tuned all the hyperparameters based on the performance of our model on the validation set. Before coming up with the final prompt structure for few-shot GPT-3, we experimented with a couple of things. We tested by keeping <*Section Header*> in their original form (as received in training data - acronymized) and also by replacing them to their full form, as received in the tasks description. Please refer to table 4 for a list of all headers and their expanded forms.

We experimented with the number of ground truth examples for our prompt. We tested with a maximum of 3 examples, because, for values higher than 3, the rate of getting the maximum token limit error from the GPT-3 API had increased significantly. Across 0, 1, 2, and 3, we found 3 to be giving the best results on the validation set. Finally, our ensemble of BART, BioBART, and GPT-3 outperforms all our individual models by **0.9+** Rouge-Avg points (Refer Table 7) on the validation dataset. Some sample dialogues and generated section text are shown in Table 5.

Table 6 and 7 show the result of our evaluation for the Section Header Identification and Section Text Generation pipeline on the validation and test datasets respectively. The evaluators report a few more metrics such as BERTScore Zhang

[3]Bio_ClinicalBERT Model:https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT

[4]BART Model: https://huggingface.co/facebook/bart-large

[5]BioBART Model: https://huggingface.co/GanjinZero/biobart-large

Split	Method	Accuracy
Val	Fine-tuned Bio_ClinicalBERT	0.75
	Fine-tuned Bio_ClinicalBERT + KW Classifier	0.77
Test	Fine-tuned Bio_ClinicalBERT + KW Classifier	0.73

Table 6: Section Header classification from Dialogue on Val set. Here, KW stands for Keyword-based.

Split	Method	R1	R2	RL	RLS	RA	BSF	BLEURT
Val	Fine-tuned BioBART-large	38.1	14.8	31.0	31.0	28.7	-	-
	Fine-tuned BART-large	39.0	14.6	31.6	31.4	29.2	-	-
	Few-shot GPT-3	40.3	16.5	32.4	32.2	30.4	-	-
	Ensemble	41.6	17.2	33.1	33.3	31.3	-	-
Test	Ensemble	39.8	17.17	33.14	33.13	30.81	69.82	53.5

Table 7: Section text generation from Dialogue on the validation set. Here, R1, R2, RL, RLS, RA, and BSF refer to Rouge-1, Rouge-2, Rouge-L, RougeLSum, Rouge-Avg, and BertScore-F1 respectively.

et al. (2019), a metric that focuses on computing semantic similarity between tokens of reference and hypothesis, and BLEURT Sellam et al. (2020), a learned evaluation metric based on BERT for evaluating the generated summaries. The default models used for calculating BERTScore and BLEURT were 'microsoft/deberta-xlarge-mnli' [6] and 'BLEURT-20'[7] respectively. We report the score for these metrics in Table 7 for the test datasets due to computing constraints.

7 Observations

Here we discuss some observations that we made on results as shown in Table 5.

- With reference to examples 1 and 2, our model was able to correctly capture the year, duration, and other diagnostic details.
- With reference to example 3, our model was able to capture more details and also attempted to categorize diagnosis under relevant categories, which was not originally present in the ground truth summary.
- With reference to examples 4 and 5, our model generated some made-up facts such as the duration of the day, and the chest examination being an x-ray.
- With reference to examples 6 and 7, our model was accurately able to generate text with the same findings. However, it wrote it in an elaborate manner pushing R1 to 0.0.

8 Conclusion and Future Work

We have presented a novel ensemble-based approach for the task of automatic short medical dialogue to note summarization. Our method effectively combines fully supervised transformer models, few-shot GPT-3, and rule-based systems, generating accurate and coherent summaries of doctor-patient conversations. The proposed system demonstrates competitive performance on the MEDIQA-Chat 2023 Task A, highlighting its potential to enhance telecare and healthcare services.

As part of future work, we plan to explore advanced pre-trained models and techniques to further improve our system's performance in the medical context. Additionally, we aim to investigate the applicability of our approach in handling more complex dialogues. We also plan to conduct an in-depth analysis of the generated summaries to identify areas for further fine-tuning.

[6]DeBERTa Model: https://huggingface.co/microsoft/deberta-xlarge-mnli

[7]BLEURT-20 Model: https://huggingface.co/lucadiliello/BLEURT-20

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Asma Ben Abacha, Wen wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen. 2023a. Overview of the mediqa-chat 2023 shared tasks on the summarization and generation of doctor-patient conversations. In *ACL-ClinicalNLP 2023*.

Asma Ben Abacha, Wen wai Yim, Yadan Fan, and Thomas Lin. 2023b. An empirical study of clinical note generation from doctor-patient encounters. In *EACL 2023*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Salem Garfan, Abdullah Hussein Alamoodi, BB Zaidan, Mohammed Al-Zobbi, Rula A Hamid, Jwan K Alwan, Ibraheem YY Ahmaro, Eman Thabet Khalid, FM Jumaah, Osamah Shihab Albahri, et al. 2021. Telehealth utilization during the covid-19 pandemic: A systematic review. *Computers in biology and medicine*, 138:104878.

Karan Goel, Nazneen Rajani, Jesse Vig, Samson Tan, Jason Wu, Stephan Zheng, Caiming Xiong, Mohit Bansal, and Christopher Ré. 2021. Robustness gym: Unifying the nlp evaluation landscape. *arXiv preprint arXiv:2101.04840*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. *arXiv preprint arXiv:2009.08666*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Hayato Kobayashi. 2018. Frustratingly easy model ensemble for abstractive summarization. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 4165–4176.

Kundan Krishna, Sopan Khosla, Jeffrey P. Bigham, and Zachary C. Lipton. 2021. Generating soap notes from doctor-patient conversations using modular summarization techniques.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

T. C. Rajapakse. 2019. Simple transformers. `https://github.com/ThilinaRajapakse/simpletransformers`.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Yan Song, Yuanhe Tian, Nan Wang, and Fei Xia. 2020. Summarizing medical conversations via identifying important utterances. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 717–729, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Wen wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. The aci demo corpus: An open dataset for benchmarking the state-of-the-art for automatic note generation from doctor-patient conversations. *Submitted to Nature Scientific Data*.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. Biobart: pretraining and evaluation of a biomedical generative language model. *arXiv preprint arXiv:2204.03905*.

Longxiang Zhang, Renato Negrinho, Arindam Ghosh, Vasudevan Jagannathan, Hamid Reza Hassanzadeh, Thomas Schaaf, and Matthew R. Gormley. 2021. Leveraging pretrained models for automatic summarization of doctor-patient conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3693–3712, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. Dialoglm: Pre-trained model for long dialogue understanding and summarization.

Storyline-Centric Detection of Aphasia and Dysarthria in Stroke Patient Transcripts

Peiqi Sui[1], **Kelvin K. Wong**[1], **Xiaohui Yu**[1], **John J. Volpi**[2], and **Stephen T. C. Wong**[1]
[1] T.T. and W.F. Chao Center for BRAIN & Houston Methodist Neal Cancer Center, Houston Methodist Hospital
[2] Eddy Scurlock Comprehensive Stroke Center, Department of Neurology, Houston Methodist Hospital and Weil College of Medicine
psui@houstonmethodist.org

Abstract

Aphasia and dysarthria are both common symptoms of stroke, affecting around 30% and 50% of acute ischemic stroke patients. In this paper, we propose a storyline-centric approach to detect aphasia and dysarthria in acute stroke patients using transcribed picture descriptions alone. Our pipeline enriches the training set with healthy data to address the lack of acute stroke patient data and utilizes knowledge distillation to significantly improve upon a document classification baseline, achieving an AUC of 0.814 (aphasia) and 0.764 (dysarthria) on a patient-only validation set.

1 Introduction

Published studies reported that about 30% of acute ischemic stroke patients are presented with aphasia as an initial symptom (Engelter et al., 2006), while around half of these patients exhibit some form of dysarthria (Urban et. al., 2001), with acute dysarthria specifically associated with small lacunar stroke primarily due to small vessel disease. The "cookie theft" picture description task shown in Figure 1 is commonly used for language assessment in the NIH stroke scale (NIHSS) to score the severity of aphasia and dysarthria among others. Currently, the scoring is done by a certified healthcare worker.

Figure 1. Cookie theft picture commonly used in patient description tasks for stroke assessment and aphasia and dysarthria diagnosis.

Recent research has demonstrated the feasibility of deep learning-based stroke detection, using facial expression and voice data gathered from the "cookie theft" storytelling task (Figure 1) that serves to differentiate mild/moderate stroke among stroke mimics in the emergency room (Cai et al., 2022). However, existing approaches to AI-enabled stroke prediction have only utilized the audio spectrum of patient recordings. The language content of these recordings is yet to be used for language assessment, even though the storytelling audio is often automatically transcribed. This motivates us to apply unstructured storytelling transcript and large-scale language model in order to predict the presence of aphasia or dysarthria in patients with stroke-like symptoms, using NIHSS subscores 9 and 10 as gold standard.

In this paper, we present a new storyline-centric pipeline that uses transcribed patient descriptions alone to detect aphasia and dysarthria in patients with stroke-like symptoms. Although no such study has been done in stroke to the best of our knowledge, these unlabeled patient transcripts are proven to be highly useful for language-related symptoms detection by a robust body of research in Alzheimer's disease (AD) prediction and monitoring (de la Fuente Garcia et al., 2020). Working with patient descriptions of the cookie theft picture amongst other transcripts, recent studies in AD discover that transformer-based language models that leverage a comprehensive language understanding (Guo et al., 2019; Qiao et al., 2021; Liu et al., 2021; Wang et al., 2023) tend to outperform models trained on syntactic, lexical, or pragmatic features alone (Fors et al, 2018; Ammar and Ayed, 2018). Moreover, models that depend on syntax and pre-defined lexicons are more prone to racial and educational biases that

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 422–432
July 14, 2023 ©2023 Association for Computational Linguistics

discriminate against patients who are non-native speakers or dialect users of English.

The relative lack of NLP-enabled stroke detection could be due to the lack of patient textual data, while similar studies in Alzheimer's disease could benefit from publicly available corpora of patient narratives, both on the cookie theft picture and otherwise (MacWhinney 2019). To tackle this challenge, we are interested in exploring how data from cloud-sourced healthy volunteers, which is easier and more cost-effective to obtain, could be used to improve clinical NLP models. We experiment with two approaches to enrich our training set with healthy subject data, by including them first directly as the texts themselves and then indirectly as metadata representations in the form of knowledge graphs. By circumventing the data bottleneck, we believe that it is possible to improve NLP-enabled detection of language symptoms in stroke patients.

Major Contributions. In this work, we present 1) a pair of ELECTRA-based models for detecting aphasia and dysarthria in patient documents by performing data distillation with storyline-encoded knowledge graphs extracted from both healthy and patient transcripts, 2) de-noised document-level knowledge graphs that represent the "correct" storyline as a consensus between healthy volunteers, which provides semantic emphasis that enriches document classification, and 3) a qualitative evaluation of our models' performance that examines their semantic and clinical limitations with error-based behavioral testing.

2 Data Enrichment and Baselines

2.1 Patient Data

To build our dataset, patients with stroke-like symptoms from the Houston Methodist Hospital System are instructed to verbally describe the "cookie theft" image for one minute while their audio and facial video were recorded. Ground truth labels for aphasia and dysarthria are respectively obtained from subscores 9 and 10 of the NIH stroke assessment. The voice recordings of patients describing the image in English are automatically transcribed with Assembly AI, resulting in a dataset of 268 patient transcripts (3 patient samples are dropped due to poor quality). We retrieve subscores 9 and scale 10 scores for aphasia and dysarthria respectively from these patient transcriptions (49/268 for aphasia, 74/268 for dysarthria). Notably, 44/49 of our aphasia patients and 60/74 of our dysarthria patients are diagnosed with stroke.

2.2 Data Enrichment with Crowdsourced Healthy Volunteer Transcripts

Data enrichment refers to the process of supplementing internal data with external data sources (Allen and Cervo, 2015). In the clinical domain, it could be applied to address the lack of available patient data by enlarging the training set with healthy subjects as negative labels. We leverage Amazon's Mechanical Turk (mTurk) to collect healthy volunteer voice data from native English speakers from the United States describing the same cookie-jar theft story (n=988). We conduct manual quality control and confidence score evaluations to filter the mTurk dataset (n=675) to ensure that it only consists of high-quality audio recordings and storylines that resemble that of healthy subjects.

The healthy subject transcripts are then used to enrich the patient transcripts, and both healthy and patient transcripts were separately split at an 80/20 ratio into an enriched training set (n=754, 214 patients), which utilizes both healthy and patient data for training, and a patient-only validation set (n=54), with the proportion of each label class (aphasia or dysarthria) preserved. We exclude all healthy data points from the validation set, to make sure that measurement metrics in upcoming sections would represent the classification performance on acute stroke patients alone. We choose to not include a hold-out test set due to the lack of patient data.

2.3 Enriched Baselines for Patient Document Classification

Transformer-based methods (Vaswani et al., 2017) have been credited with most recent progress in the area of text classification (Minaee et al., 2021). We experiment with fine-tuning various transformer-based language models, including BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2019), and ELECTRA (Clark et al., 2019), to implement binary classification on our patient documents for both aphasia and dysarthria. We choose not to implement any clinical NLP models because patient descriptions of the cookie theft picture themselves are not particularly relevant to the clinical domain.

Model	BERT-base	ALBERT-large	RoBERTa	ELECTRA-large
AUC (label='Aphasia')	0.533	0.595	0.512	**0.615**
AUC (label='Dysarthria')	0.421	0.416	0.403	**0.424**
AUC (label='Combined')	0.558	0.596	0.52	**0.627**

Table 1. Baseline patient document classification performance on aphasia, dysarthria, and combined (patients with either aphasia, dysarthria, or both) labels, after data enrichment.

We first experiment with only using patient data for model training, and their performance is evaluated with the area-under-curve (AUC) metric. Due to the small size of the patient training set (n=214), the validation performance of various baseline models is unsatisfactory when trained on patient data alone, with AUC between 0.43 to 0.46. To address this, an enriched training set is created by combining healthy and patient data, while the best model is selected using a patient-only validation set.

Baseline models are established when a significant improvement in performance is achieved with data enrichment. ELECTRA-large is the best performing model overall: after enrichment, its AUC rises to 0.615 for aphasia and 0.627 for either aphasia or dysarthria in the patient-only validation set. Notably, all models' inferences on aphasia outperform that of dysarthria, as shown in Table 1. This gap in performance could be attributed to the imprecise and unintelligible speech (Yorkston, 1996) that is common in dysarthria due to poor motor coordination. As a form of language impairment, dysarthria is more often manifested as difficulty articulating rather than semantic mismatch (Mitchell et al., 2017), which might not be directly visible to language models without domain-specific fine-tuning.

Our proposed methods in the next section aim to improve upon these baseline results, as stated in the metrics of Table 1.

3 Knowledge Distillation

This section reports our experiment designs aimed at testing the hypothesis that knowledge distillation with storyline-encoded knowledge graphs, extracted from both healthy and patient transcripts, would transfer semantic knowledge to the enriched document classification model and improve the performance of detecting aphasia and dysarthria.

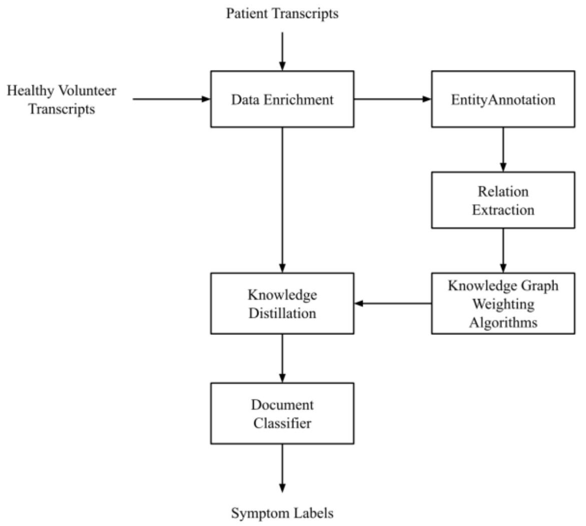


Figure 2. Schematic of the knowledge distillation pipeline.

As shown in Figure 2, the knowledge distillation pipeline has two components. First, we conduct unsupervised knowledge graph extraction with BERT-based entity annotation and relation extraction models. Second, we leverage these knowledge graphs to provide semantic reinforcement for the baseline ELECTRA document classification model defined in Section 2.3.

3.1 Knowledge Graph Extraction from Healthy Volunteer Transcripts as Representations of Ground Truth Storyline

Beyond data enrichment, we further leverage the mTurk dataset of healthy subjects to construct document-level knowledge graphs (KG) that represent the ground truth storyline of the "cookie theft" picture, which we will use to enrich the knowledge distillation learning in Section 3.3. The nodes of these KGs are key entities extracted from each transcript by a BERT-based annotator (Grootendorst, 2020), and the edges between them are semantic relations that describe a form of non-sequential storyline. For relation extraction, we implement a BERT-based model (Soares et al., 2019)[1] fine-tuned with the general domain relation labels of SemEval-2010 Task 8 (Hedrickx et al., 2010). However, the initial outputs lack coherence and contain excessive noise. To address this, we implement the following denoising strategies sequentially:

1. **Entity permutation on the sentence level.** Each transcript is tokenized into sentences, and key phrases (n=4) are extracted on a sentence level. Each key phrase is paired with each other's key phrase and then passed into the relation extractor with the rest of the sentence. We filtered out excessively short sentences (n<=5) from going into the classifier to avoid having trivial words annotated as key phrases, and the relations classified as "Other" from going into the output knowledge graph. This results in a significant improvement from the initial approach to pass the entire paragraph into the relation extractor, which creates many contrived situations where a word at the start of the paragraph might be paired with a word at the end.
2. **Domain-specific pre-training on mTurk and patient transcripts.** We include an additional pre-training step to the relation extractor so that the language model could have some exposure to our corpus before fine-tuning and produce more relevant results.
3. **Nodes and edges cleaning.** Entities that evidently describe the same entity, i.e., "mother" and "woman," are combined. Redundant relation labels are also removed from the fine-tuning stage. The Sem-Eval 2010 Task 8 dataset's relation classification dataset contains these following relation labels that are not relevant to the cookie theft picture: "Component-Whole," "Product-Producer," "Member-Collection," and "Message-Topic." Removing them significantly reduces the number of misplaced nodes and edges in the output KGs.
4. **Updating to lighter models.** We run relation extraction on ALBERT instead of BERT, since it performed better on both the patients and enriched set during the baseline testing.

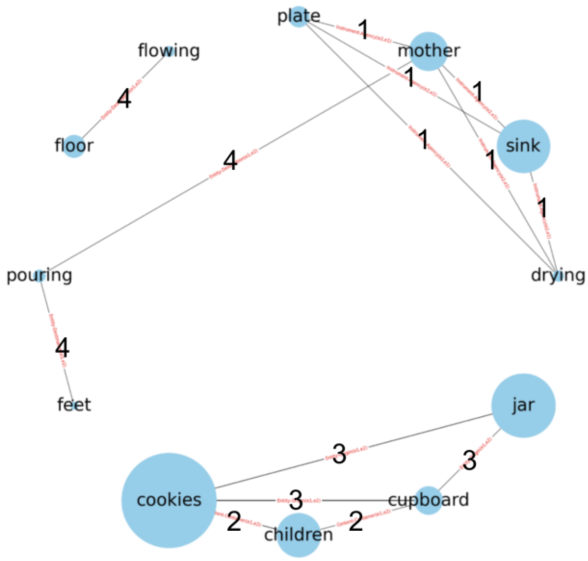


Figure 3. Example of a denoised knowledge graph. The nodes are key entities from the picture descriptions, with their size representing their phrase importance. The edges are semantic relations between the two nodes, including:

1. Instrument-Agency
2. Content-Container
3. Entity-Origin
4. Entity-Destination

An example of a typical document-level KG after denoising is shown in Figure 3. The nodes are then weighted using TF-IDF based on the collection of all entities extracted from the mTurk dataset, which indicates the relative importance of each entity.

3.2 Knowledge Distillation for Semantic Reinforcement in Aphasia and Dysarthria Document Classification

With 675 denoised KGs that represents the storyline as described by healthy subjects, we aim to conduct storyline-centric knowledge distillation learning to improve on the classification results of data enrichment alone (Section 3.1) in Table 1.

[1] We could not find the official code repository for Soares et al. (2019). Instead, we used a popular community implementation available at `https://github.com/plkmo/BERT-Relation-Extraction`.

Patient Passage	**Extracted Triples**

Okay. Oh, I've seen this before. There's a lot of things going on, and some of them aren't quite right. The **sink** is **overflowing**, and there getting some **cookies** from this cookie jar, and he's about to fall, but then she's acting like everything's okay, and it's not okay, but she must be on Prospect or something like that, because she does not even see that anything is happening over there. And he's about to **fall**, and he can have a **concussion** or something like that. And it's a nice day outside, and there's some other things in the **cabinet**, and there's nothing on this side of the **sink** or anything like that. And she's drying a dish, and there's **curtains**, and there's a walkway, and there's **sink**	sink, Cause-Effect(e2,e1), overflowing;
	cookie, Cause-Effect(e1,e2), overflowing;
	sink, Cause-Effect(e2,e1), fall;
	sink, Content-Container(e1,e2), cabinet;
	concussion, Cause-Effect(e2,e1), fall;
	curtains, Entity-Origin(e2,e1), drying;
	drying, Instrument-Agency(e2,e1), dish

Figure 4. Sample patient transcript and its extracted triples. Color scheme denotes sentence of origin, with entities bolded in the transcript.

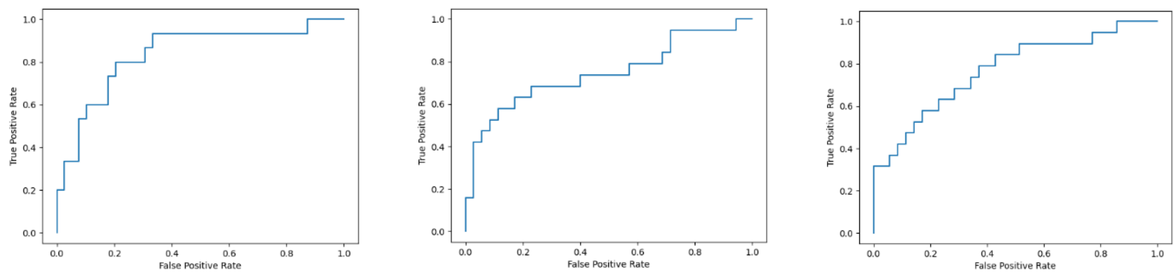


Figure 5. ROC curves of triples-distilled ELECTRA-large on aphasia (AUC=0.814), dysarthria (AUC=0.764), and combined (AUC=0.769) in patient only validation set

To achieve this, we experiment with two types of knowledge distillation: 1) triple classification, which combines all 675 KGs into one large KG, and 2) triple concatenation, which leverages the KGs individually on the document level.

To construct and extract more meaningful and accurate KGs from patient transcripts, we make the following adjustments to the methods described in Section 3.1 to increase its effectiveness in a clinical setting:

- **Joining sentences that are excessively short.** 41.8% of the patients with either aphasia or dysarthria and 37.4% of those without talk in very fragmented sentences with minimal syntax, usually in two-or-three-word sequences of "subject-verb," "subject-verb-object," or simply isolated words. To ensure the meaningfulness of the output entities, we join these sentences together to offer sufficient contexts for the BERT-based relation extractor.
- **Parameter tuning.** We evaluate the output from different key phrase counts per sentence and minimum sentence length, and qualitatively determine that four entities per sentence no shorter than 100 characters generates the best tradeoff between noise and the number of triples extracted.

A representative example of patient transcripts and their extracted triples is shown in Figure 4. While mostly accurate, the KG model is still limited by some inaccurate entity pairings as it permutates through each sentence, which we hope to address with further denoising steps outlined in Section 5.

Triple Classification: Triple classification is a KG completion task that identifies whether a triple could be a constitutive part of a certain given KG, as a naive approach to knowledge distillation. A triple is defined as a set of head, relation, and tail that is the basic constituting unit of a KG. We implement triple classification with LMKE (Wang et al., 2022), a BERT-based model that is representative of various highly similar recent triple classification models based on the same codebase proposed by Yao et al. (2019). We collapse all patient-level KGs generated in Section 3.1 into one large KG, use it to re-train LMKE as the ground truth training set, and then evaluate the model on all patient triples as the inference set. This is in essence a zero-shot learning approach, as the training set only has positive labels. We run inference on each triple extracted from the patient transcripts to determine if they belong to the ground truth KG. The triples extracted from patients with either aphasia or dysarthria are expected to output a negative label (not belonging to the ground truth KG), since our assumption is that their descriptions do not semantically fit the "no language symptom" storyline. Unfortunately, LMKE does not facilitate patient-level prediction, while its triple-level performance (AUC=0.612, label= 'combined') does not suggest potential improvement from the baselines in Section 2.3.

Triple Concatenation: Since LMKE does not perform well on the triple level, we seek to transfer and distill the knowledge that the model learned from triple classification to the patient level. One approach we successfully implement is to concatenate triples to the end of the transcript that they are generated from as a way of data distillation for semantic reinforcement. Since the triples are directly extracted from the transcripts and the two are too correlated to be considered statistically independent, we choose to concatenate them instead of training them as two separate features. This significantly increases the performance of ELECTRA on combined language symptoms detection (AUC=0.769), and the feature dependency that motivates the concatenation is further validated by an ablation that shows using the concatenated triples alone would perform poorly (AUC=0.427). Notably, the triples-enriched model demonstrates a significant improvement on dysarthria detection (AUC=0.764), which makes its performance more balanced between different language conditions (Figure 5). Excessive repetition (Mitchell et al., 2017) is a prominent sign of dysarthria, as recent studies (Mitchell et al., 2021; Kirshner, 2022) find a repetition test to be an effective metric for dysarthria diagnosis and examination. Our use of triples in aphasia and dysarthria detection could be considered as an AI-enabled automation of the repetition test. It puts semantic emphasis on key entities that dysarthria patients struggle to articulate, which would otherwise not be visible to ELECTRA or other language models from the word embedding space alone.

4 Behavioral Testing and Discussion

We conduct further qualitative testing to thoroughly evaluate the sensitivity of our aphasia and dysarthria models to different types of language errors and generalizability to external data. This is motivated by a recent surge in the behavioral testing of NLP models that challenge the effectiveness of common quantitative testing metrics (Ribeiro et al., 2020). For NLP models in the clinical domain, van Aken et al. (2022) highlight the need to simulate plausible real-life patient inputs to analyze model sensitivity directly.

Our main goals thus are to 1) verify that our models are in fact learning semantically, as expected by our methodology, 2) simulate "external" data to assess potential model overfitting to the specific clinical settings of the patient data, and 3) better understand the semantic limitations and boundary conditions of our models in order to make more accurate, informed, and measured claims about their clinical outcomes. For each patient transcript, we generate 9 versions of the original text that amplify types of language errors in both semantic and syntactic categories. Table 2 compares the aphasia and dysarthria models' performance on all categories of error-infused transcripts, with F1 as the evaluation metric.

4.1 Syntax Testing

On the word level, we manually saturate each patient description with subject-verb disagreement, verb tense, and pronoun errors. The dysarthria model's huge drop in F1, in particular, shows that it is particularly sensitive to word-level syntax errors.

On the sentence level, we experiment with sentence structures that could cause grammatical confusion: 1) run-on sentences with too few punctuations, and 2) overly fragmented sentences with too many punctuations. 1) has been directly identified as a symptom of aphasia by the NIH's

Error Type	Example	Aphasia (F1)	Dysarthria (F1)
Original	"There's a kid falling off a chair, trying to get a cookie. His sisters trying to take the cookie away from him. And moms washing dishes. The sinking is overflowing. She's looking out the window and the water is going all over the floor…"	0.86	0.72
Random Noise	"There's a kidh fallibng off a chair, trying touket a cookie.qHis sisters tryingh to take the cookie iway fromq him. jAnd mems wauhing dishes. Thepslinhking is overlflowing. Shme's lokking omt thg window and the watewr gis ghing all nver the floor…"	0.62	0.33
Excessive Grammatical Errors	"There have a kid falled off a chair, tryed to get a cookie. Her sisters tries take the cookie away from her. And moms wash dishes. The sinking are overflowing. Him's looked out the window and the water is going all over the floor…"	0.69	0.4
Run-on Sentences	"There's a kid falling off a chair trying to get a cookie his sister's trying to take the cookie away from him and moms washing dishes the sinking is overflowing she's looking out the window and the water is going all over the floor…"	**0.73**	0.62
Fragmented Sentences	"There's a. kid falling. off a chair trying. to get. a cookie. His. sisters trying. to. take. the. cookie away from him. And moms. washing dishes. The sinking is overflowing. She's looking out. the window. and the water. is going. all over the floor…"	0.67	**0.64**
Additional Object(s)	"There's radio a kid falling radio off a chair, trying radio to get radio a cookie. His sisters trying to take the cookie away from him. And moms washing dishes. The sinking is overflowing. She's looking out the window and the water is going all over the floor…"	0.7	0.6
Removed Key Object(s)	"There's a kid falling off a chair, trying to get. His sisters trying to take away from him. And moms washing dishes. The sinking is overflowing. She's looking out the window and the water is going all over the floor…"	0.67	0.54
Keeping First Sentence Only	"There's a kid falling off a chair, trying to get a cookie."	0.7	0.25
Randomly Deleting	"to His take from dishes. The sinking window and is pretty that's about getting soaking wet. a and the boy's cookie in his right grab another one. to"	0.56	0.37
Reversed Sentence Order	"Is that no. And the sisters reaching up, trying to get one of the cookies from me. The boy's holding cookie in his left hand as he's falling off the chair, and he's got his right hand in the cookie jar trying to grab another one. That's about all I see… There's a kid falling off a chair, trying to get a cookie."	0.68	0.56

Table 2. Examples of cookie theft picture descriptions infused with each category of language errors for behavioral testing

most up-to-date definition[2], while 2) has been linked to dysarthria as many studies find that dysarthria patients tend to be more effective at processing shorter sentences (Allison et al., 2019), especially when aided with pauses and verbal stress-making (Kuschmann and Lowit, 2021). Both observations are supported by our results: out of all error types, the aphasia model achieves the highest F1 on 1), validating it as a prominent feature of aphasia, and the dysarthria model performs the best on 2), which validates it as a prominent feature of dysarthria.

4.2 Semantics Testing

The aim of semantic pressure testing is to evaluate the extent that our models are making predictions based on semantic features, through observing their reaction to altered input descriptions with external or missing information.

- **Semantic Mismatching**: Objects in the patient descriptions are mismatched by both deleting key objects in the cookie theft picture or adding ones that are not in it. The key objects are selected from the TF-IDF ranking of knowledge graph entities extracted in Section 3.1. The performance of both models is a lot more affected by the removal of key objects that have been semantically reinforced by knowledge graphs, than the addition of external objects. This higher sensitivity verifies the effectiveness of knowledge distillation and triple concatenation.
- **Deleting**: The input text is experimented with two different degrees of deleting: 1) only keeping the first sentence, and 2) randomly deleting up to 70% of the text. The aphasia model is more affected by 2) and the dysarthria model is more affected by 1). The results are consistent with our findings in Section 4.1, as the aphasia model's strength with run-on sentences would be negated by random deleting's disruption of sentence structures, while the dysarthria model's strength with fragmented sentences would be irrelevant when there is only one sentence in the input. In addition, the dysarthria model is in general significantly more sensitive to missing texts.
- **Sentence reversal**: We also find that our models could be significantly impacted by reversing the sentence order alone, confirming numerous recent studies that demonstrate BERT-based models' sensitivity to the word or sentence order of the input (Hessel and Schofield, 2021; Pham et al., 2021).

5 Conclusion and Future Work

This work explores the under-researched area of applying NLP to unlabeled patient transcripts for supporting the triage and detection of stroke and stroke mimics. We introduce a storyline-centric approach that leverages data enrichment and knowledge distillation to overcome the lack of big clinical training datasets for automating aphasia and dysarthria detection. Our experiments show that our approach to knowledge distillation has the potential to significantly improve the performance of patient document classification. Nonetheless, we believe that it is possible to further enhance the results in Figure 5 by designing more robust and effective knowledge distillation techniques to integrate transcripts, triples, and graph-theoretic aspects of KGs.

Our ongoing work include: 1) using Sentence Transformers (Reimers and Gurevych, 2019) to further denoise the output of KG extraction, 2) developing solutions to incorporate both semantic knowledge embeddings and graph embeddings in clinical document classification, and 3) recruiting Spanish-speaking patients and healthy volunteers and expanding our storyline-centric pipeline to Spanish language models (Gutierrez-Fandino et al., 2021), to better serve the clinical needs of the Hispanic community in stroke triage and detection.

Limitations

Due to clinical and financial constraints, both the patient and the mTurk sample sizes of our study are still relatively small. This means that we cannot afford to set aside patient data as a hold-out test set, and have to use the validation set for model evaluation. As we work towards enrolling more patients and recruiting more healthy volunteers to

[2] https://www.nidcd.nih.gov/health/aphasia

improve model generalizability, we hope to expand the scope of our pipeline beyond English to serve non-native speaker patients.

One major limitation of the cookie theft picture description task is its lack of equitable assessment for an increasingly diverse patient population. Steinberg et al. (2022) identify gender as a particularly fraught aspect of the picture's expected response, as the rubrics of the initial NIHSS were established from a male-only corpus. Although there is no alternative picture or stroke patient corpus available to our study, we try to ensure the equity of our models by maintaining a gender balance in our patient set, with 136 female patients and 132 male patients. On our patient-only evaluation set, our aphasia model performs significantly better on female patients (AUC=0.909) compared to male patients (AUC=0.702), while the dysarthria model exhibits better performance on male patients (AUC=0.778) than female patients (AUC=0.719). At present, we are unable to draw any definitive conclusions about model equity due to the scale of our data. However, it will be a key area of focus for our future research.

Acknowledgments

We sincerely thank the workshop organizers and anonymous reviewers for their generous time, attention, and feedback. This work was supported by the T. T. & W. F. Chao Foundation and the John S. Dunn Research Foundation.

References

Mark Allen and Dalton Cervo. 2015. *Multi-domain master data management: Advanced MDM and data governance in practice*. Elsevier.

Kristen M. Allison, Yana Yunusov, and Jordan R. Greene. 2019. Shorter sentence length maximizes intelligibility and speech motor performance in persons with dysarthria due to amyotrophic lateral sclerosis." *American journal of speech-language pathology*, 28(1): 96-107.

Randa Ben Ammar and Yassine Ben Ayed. 2018. Speech processing for early Alzheimer's disease diagnosis: Machine learning based approach. In *IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA)*, pages 1-8, Aqaba, Jordan.

Tongan Cai, Haomiao Ni, Mingli Yu, Xiaolei Huang, Kelvin Wong, John Volpi, James Z. Wang, and Stephen T. C. Wong. 2022. *DeepStroke*: An efficient stroke screening framework for emergency rooms with multimodal adversarial deep learning. *Medical image analysis*, 80.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2019. ELECTRA: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv: 2003.10555*.

Sofia de la Fuente Garcia, Craig Ritchie, and Saturnino Luz. 2020. Artificial intelligence, speech and language processing approaches to monitoring Alzheimer's disease: A systematic review. *Journal of Alzheimer's disease*, 78(4): 1547-1574.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Stefan T. Engelter, Michal Gostynski, Susanna Papa, Maya Frei, Claudia Born, Vladeta Ajdacic-Gross, Felix Gutzwiller, and Phillipe A. Lyrer. 2006. Epidemiology of aphasia attributable to first ischemic stroke: Incidence, severity, fluency, etiology, and thrombolysis. *Stroke*, 37(6): 1379-1384.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquín Silveira-Ocampo, Casimiro Pio Carrino, Aitor Gonzalez-Agirre, Carme Armentano-Oller, Carlos Rodriguez-Penagos, and Marta Villegas. 2021. MarIA: Spanish Language Models. *arXiv preprint arXiv:2107.07253*.

Kristina Lundholm Fors, Kathleen Fraser, and Dimitrios Kokkinakis. 2018. Automated syntactic analysis of language abilities in persons with mild and subjective cognitive impairment. *Studies in health technology and informatics*, 247: 705-709.

Martin Grootendorst. 2020. KeyBERT: Minimal keyword extraction with BERT. https://doi.org/10.5281/zenodo.4461265.

Zhiqiang Guo, Zhenhua Ling, and Yunxia Li. 2019. Detecting Alzheimer's disease from continuous speech using language models. *Jourrnal of Alzheimer's disease*, 70(4): 1163-1174.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 Task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden. Association for Computational Linguistics.

Jack Hessel and Alexandra Schofield. 2021. How effective is BERT without word ordering? Implications for language understanding and data privacy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 204–211, Online. Association for Computational Linguistics.

Howard Kirshner. 2022. Dysarthria and Aphasia of speech. in Bradley and Daroff's *Neurology in Clinical Practice*, 8th Edition, 14: 149-151. United Kingdom, Elsevier.

Anja Kuschmann and Anja Lowit. 2021. Pausing and sentence stress in children with dysarthria due to cerebral palsy. *Folia phoniatrica et logopaedica*, 73(4): 298-307.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Ning Liu, Zhenming Yuan, and Qingfeng Tang. 2021. Improving Alzheimer's disease detection for speech based on feature purification network. *Frontiers in Public Health*, 9.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Brian MacWhinney. 2019. Understanding spoken language through TalkBank. *Behavioral research methods*, 51(4):1919-1927.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep learning based text classification: A comprehensive review. *ACM computing surveys*, 54(3): 1-40.

Claire Mitchell, Audrey Bowen, Sarah Tyson, Zoe Butterfint, and Paul Conroy. 2017. Interventions for dysarthria due to stroke and other adult-acquired, non-progressive brain injury. The Cochrane database of systematic reviews, 1(1).

Claire Mitchell, Matthew Gittins, Sarah Tyson, Andy Vail, Paul Conroy, Lizz Paley, and Audrey Bowen. 2021. Prevalence of aphasia and dysarthria among inpatient stroke survivors: describing the population, therapy provision and outcomes on discharge. *Aphasiology*, 35(7): 950-960.

Thang Pham, Trung Bui, Long Mai, and Anh Nguyen. 2021. Out of Order: How important is the sequential order of words in a sentence in Natural Language Understanding tasks?. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1145–1160, Online. Association for Computational Linguistics.

Yu Qiao, Xuefeng Yin, Daniel Wiechmann, and Elma Kerz. 2021. Alzheimer's disease detection from spontaneous speech through combining linguistic complexity and (dis)fluency features with pretrained language models. *arXiv preprint arXiv:2106.08689*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4902– 4912, Online. Association for Computational Linguistics.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Amy Steinberg, Patrick D. Lyden, and Arielle P. Davis. 2022. Bias in Stroke Evaluation: Rethinking the Cookie Theft Picture. Stroke, 53(6): 2123-2125.

P. P. Urban, S. Wicht, G. Vukurevic, C. Fitzek, S. Fitzek, P. Stoeter, C. Massinger, and H. C. Hopf. 2001. Dysarthria in acute ischemic stroke: Lesion topography, clinicoradiologic correlation, and etiology. *Neurology*, 56(8):1

Betty van Aken, Sebastian Herrmann, and Alexander Löser. 2022. What do you see in this patient? Behavioral testing of clinical NLP models. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 63-73, Seattle, Washington. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

Xintao Wang, Qianyu He, Jiaqing Liang, and Yanghua Xiao. 2022. Language models as knowledge embeddings. *arXiv preprint arXiv:2206.12167*.

Yi Wang, Jiajun Deng, Tianzi Wang, Bo Zheng, Shoukang Hu, Xunying Liu, and Helen Meng. 2023. Exploiting prompt learning with pre-trained language models for Alzheimer's disease detection. *arXiv preprint arXiv:2210.16539*.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. KG-BERT: BERT for knowledge graph completion. *arXiv preprint arXiv:1909.03193*.

Kathryn M. Yorkston. 1996. Treatment efficacy: Dysarthria. Journal of speech and hearing research, 39(5): 46-57.

Pre-trained language models in Spanish for health insurance coverage

Claudio Aracena[1,2], **Nicolás Rodríguez**[3], **Victor Rocco**[3], **and Jocelyn Dunstan**[2,4,5,6]
[1]Faculty of Physical and Mathematical Sciences, University of Chile
[2]Millennium Institute Foundational Research on Data, Chile
[3]Chilean Safety Association, Chile
[4]Department of Computer Science, Catholic University of Chile
[5]Institute for Mathematical and Computational Engineering, Catholic University of Chile
[6]Center for Mathematical Modeling, University of Chile
claudio.aracena@uchile.cl, {nrodrigueza,varoccoc}@achs.cl, jdunstan@uc.cl

Abstract

The field of clinical natural language processing (NLP) can extract useful information from clinical text. Since 2017, the NLP field has shifted towards using pre-trained language models (PLMs), improving performance in several tasks. Most of the research in this field has focused on English text, but there are some available PLMs in Spanish. In this work, we use clinical PLMs to analyze text from admission and medical reports in Spanish for an insurance and health provider to give a probability of no coverage in a labor insurance process. Our results show that fine-tuning a PLM pre-trained with the provider's data leads to better results, but this process is time-consuming and computationally expensive. At least for this task, fine-tuning publicly available clinical PLM leads to comparable results to a custom PLM, but in less time and with fewer resources. Analyzing large volumes of insurance requests is burdensome for employers, and models can ease this task by pre-classifying reports that are likely not to have coverage. Our approach of entirely using clinical-related text improves the current models while reinforcing the idea of clinical support systems that simplify human labor but do not replace it. To our knowledge, the clinical corpus collected for this study is the largest one reported for the Spanish language.

1 Introduction

Clinical text is one of the most comprehensive data types in electronic health records. Therefore, clinical natural language processing (NLP) has become relevant to extracting helpful information from clinical writing and supporting decision-making. The complexity of human languages makes it difficult to analyze unstructured text. Additionally, the clinical text is complicated because of the heavy use of jargon, unusual spellings, and abbreviations (Dalianis, 2018).

In this complex scenario, there are various tasks that clinical NLP aims to handle. These tasks might be anything from language-related ones like text categorization, relation extraction, and entity extraction to prediction-related ones like estimating patient mortality, length of hospital stay, unplanned readmissions, etc. Several publications have addressed these tasks that have produced specialized models (Dalianis, 2018).

However, since 2017, the NLP field has worked towards creating pre-trained language models (PLMs) that can be fine-tuned for any specific downstream task. These language models are built for a much simpler task, such as next-word or masked-word prediction in a massive amount of text. This process, known as pre-training, allows the language model to acquire language understanding that can be used for any text-related task (Tunstall et al., 2022).

As soon as the NLP field started to work in PLMs, clinical NLP introduced this type of model into its set of techniques to improve performance. Some examples of clinical PLMs are two different versions of ClinicalBERT (Alsentzer et al., 2019; Huang et al., 2020). These models show a significant improvement in language tasks and a moderate improvement in prediction tasks.

Most of the research in clinical NLP has been done for text written in English, but not so much for other languages (Névéol et al., 2018). In Spanish, some publicly available PLMs relevant to clinical NLP are bsc-bio-ehr-es (Carrino et al., 2022) and Spanish Clinical Flair (Rojas et al., 2022). These PLMs were pre-trained heavily in general and biomedical text with minor additions of clinical text. Despite this drawback, they outperform general and biomedical PLMs in language tasks.

In this context, an insurance and health provider aims to analyze their clinical text to apply in a labor insurance coverage process. This provider receives patients who have suffered from a labor-related accident. When a patient is admitted to one of their clinics, admitting staff writes a report detailing the

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 433–438
July 14, 2023 ©2023 Association for Computational Linguistics

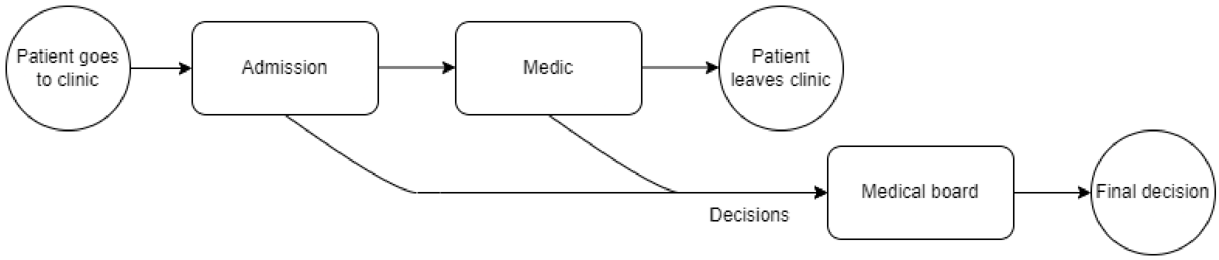


Figure 1: Flow diagram of patients and insurance coverage decisions.

accident. This report includes rich contextual information about what happened in the accident. After admission, a physician checks the patient and writes medical information. Every clinic's medical and administrative board decides a final coverage rating for each accident the next business day, considering both reports.

Currently, the provider has a model that gives a probability of not covering a patient given a specific diagnosis. This model serves as a ranking tool for the medical board to review cases with a high likelihood of no coverage. However, not all diagnoses are included in this model (considered as a categorical variable), and admission and medical reports are not considered to calculate the probability. Additionally, the model can calculate the probability of no coverage just after a physician diagnoses a patient.

This work aims to analyze clinical text from admission and medical reports to give a probability of no coverage. We try three approaches that use clinical PLMs in Spanish to carry out this goal. First, we use a clinical PLM. Second, we do continual pre-training of the previous PLM with text data from admission and medical reports. Finally, we pre-trained a LM from scratch using admission and medical reports. All outcomes will be compared to the current model performance.

2 Problem statement

In Chile, employers must hire an insurance and health provider specialized in labor accidents. These providers should cover all labor accidents. To decide if the insurance will cover a worker, the providers have clinics where they admit and check the workers to make a decision.

In this work, we use data from one of these providers. As this provider is specialized in labor accidents, the data that collects has some features. First, it has a high level of detail because that admission and medical reports are used to justify insurance coverage decisions. Second, many physicians can treat the same patient, requiring information in clinical records to be as complete as possible so that any medical staff can give better continuity to patient care over time. These features make the data cleaner compared to general health provider records.

The Asociación Chilena de Seguridad (ACHS), Chilean Safety Association in English, is a pre-eminent non-profit insurance and health provider. Its principal objective is to conceptualize and administer risk prevention programs alongside providing comprehensive coverage for occupational accidents. As evidence of its influence, ACHS accommodates more than 2.6 million affiliated workers and over 73,000 affiliated employing entities nationwide. Moreover, with a record of the lowest average accident rate, ACHS unequivocally operates as the largest mutual association in Chile.

The stringent regulations under Law No. 16,744 mandate that all Chilean employing entities, regardless of their operational scale, must be affiliated with a Social Security Administration agency. This agency is responsible for safeguarding against the risks of Occupational Accidents and Diseases. As one of three private administrative bodies, ACHS is tasked with formulating risk prevention programs. It also offers health coverage and compensation for occupational accidents, transport mishaps, and professional illnesses.

The type of labor accidents can be of two types, work-related and commuting accidents. Work-related accidents happen at the workplace or as a result of work. Commuting accidents happen on the way to or from work with no stops in between (direct trips). The staff writes an admission report when the patient is admitted in both cases. Later, when a physician receives the patient, a medical report is written.

The medical report is based on three sources. The first includes the patient's anamnesis. The second information is from the physical examination performed on the patient. The third is the medical indication for treatment. Each time a new or old patient passes through this healthcare provider and needs to be seen by a doctor, a new medical report entry is generated.

The admitting staff and the physician give a label (covered or uncovered) classification on the reports. The final classification is made the next business day after the patient is seen by the medic. A board of physicians and administrative heads from each clinic determines a final coverage rating for each case. This final rating takes into account the medical and admitting staff reports.

Most admitting staff's labels will state that patients will be covered, and medics, after clinical examination, have a more robust filter to say whether a patient will be covered or not. The committee of physicians and administrative heads has a reviewer role, and some decisions are finally changed. Figure 1 shows the flow diagram of the described process.

The current model employed by the healthcare provider only makes predictions in 72.1% of cases, basing its predictions on structured diagnoses alone. Unfortunately, this approach results in a lack of predictions for less frequently observed diagnoses. However, the majority of cases come with either admission records or medical reports, making it possible to improve coverage by utilizing these additional resources.

We expect that the use of admission and medical reports can help to take better coverage classification compared to the current model that only uses diagnosis as a categorical variable with the most common ones. Moreover, the classification prediction with the admission report can help the physician consider more information that may have been overlooked.

3 Datasets

For this study, three different types of datasets were built, for fine-tuning, continual pre-training, and pre-training from scratch. Here we list the details of these datasets:

1. For the fine-tuning process, three datasets were created. An admission dataset, which only contains text from admission reports, and a label with the final decision if that case was covered (coverage decision). A medical report, which only contains text from medical reports, and coverage decision labels. Finally, an admission and medical dataset, which concatenate text from admission and medical reports, and coverage decision labels.

2. For the continual pre-training process, also three datasets were created (admission, medical, and admission-medical datasets) similar to the fine-tuning datasets. We do not need a label in this case since these datasets are only used to continue pre-training a pre-existing PLM.

3. For the pre-training process from scratch, only one dataset was created, combining all admission and medical reports available. This dataset does not include a coverage decision label, as it is used for pre-training. However, it is bigger than previous datasets because it is used to pre-trained a PLM from scratch. According to our knowledge, this is the biggest corpus containing only clinical-related text in Spanish.

Table 1 shows details for every dataset.

Datasets	documents	tokens
Fine-tuning		
Admission	300 k	22.5 M
Medical	300 k	26.3 M
Admision+Medical	300 k	57.2 M
Continual Pre-training		
Admission	1.5 M	112.6 M
Medical	1.2 M	154.0 M
Admision+Medical	855 k	164.6 M
Pre-training		
Admision+Medical	7.1 M	1.03 B

Table 1: Number of documents and tokens in every dataset.

4 Methods

This section described the processes of pre-training and fine-tuning using the datasets described in the previous section.

4.1 Fine-tuninig of bsc-bio-ehr-es

Bsc-bio-es and bsc-bio-ehr-es are the first PLMs trained with exclusively biomedical and clinical text in Spanish (Carrino et al., 2022). These PLMs have a RoBERTa architecture and contain around 130 million parameters. Two corpora were built for

Model	bsc-bio-ehr-es	continual PLM	custom PLM
Admission	93.2 ± 0.9	**93.5 ± 0.9**	92.8 ± 0.7
Medical	94.4 ± 0.6	**94.9 ± 0.6**	94.9 ± 0.7
Admission+medical	95.9 ± 0	96.1 ± 0.1	**96.3 ± 0.2**

Table 2: Results in the test set (AUC) for all fine-tuned models.

this purpose, biomedical and clinical. The biomedical corpus consists of 2.5 million documents and 1.1 billion tokens, and the clinical corpus consists of 514k documents and 95 million tokens.

A biomedical corpus refers to medical text from academic sources, such as scientific publications or clinical trials. On the contrary, a clinical corpus is a collection of documents collected from the medical practice. In other words, it is what clinicians write during and/or after the examination of a patient.

Bsc-bio-es was pre-trained only with the biomedical corpus and bsc-bio-ehr-es with the biomedical and clinical corpora. The reason behind this design decision is two-fold; the clinical corpus is too small to create a functional PLM by itself, and to assess if adding a small clinical corpus to a large biomedical corpus positively impacts clinical NLP tasks.

As a first step, fine-tuning processes were carried out with the three fine-tuning datasets using bsc-bio-ehr-es as PLM. As a result, three fine-tuned models were built.

4.2 Fine-tuninig of a continual pre-training of bsc-bio-ehr-es

As a second step, continual pre-training processes were implemented using bsc-bio-ehr-es as a base PLM. For continuing the pre-training, the second type of datasets were used. One T4 GPU (16 GB) was used, and the processes lasted 42 hours for each. After this step, three PLM were built (admission, medical, admission+medical). Then, like the previous step, fine-tuning processes were carried out, and three more fine-tuned models were obtained.

4.3 Fine-tuninig of a PLM pre-trained from scratch

Finally, a pre-training process from scratch was implemented. This process used the same configuration as bsc-bio-ehr-es (RoBERTa) and our clinical-related corpus. Four T4 GPU (16GB) were used, and pre-training lasted 96 hours in 2 epochs. After this process, a new custom PLM was built.

With this new PLM, similar to the previous steps, fine-tuning processes were carried out, and three more fine-tuned models were obtained.

5 Results

Table 2 shows test results for every fine-tuned model. The test set only contains data not included in the fine-tuning or pre-training datasets.

We can notice that the continual PLMs and the custom PLMs are the best performers, but all the models are close performance-wise. Also, as expected, medical models are better than admission models, given that medical models capture more clinical information than admission models. The admission+medical models are the best performers since they combine admission and medical information.

As the metrics of all admission+medical models are close, we could select the least expensive and time-consuming when implementing it. In the case of this task, this process is the fine-tuning of the publicly available PLM, bsc-bio-ehr-es. However, this evidence should not be generalized for other types of tasks like named entity recognition or question answering, which are more complex and may benefit from lexical specificity. In those tasks, a PLM pre-trained with more clinical-related text could be better than a PLM trained with a mix of biomedical and clinical text.

Interestingly, admission models perform 1 to 2% worse than medical models. Therefore, there is an opportunity to make a coverage prediction before physicians check patients, helping the physicians review more medical details when their coverage decision does not match the predictions. Moreover, the admitting staff can have a stronger opinion on a coverage decision. Another benefit of providing a coverage prediction prior to the medical checkup is the possibility that the patient can manage his or her case more effectively. Depending on the likelihood of coverage the model provides, the patient may seek resources to help better justify his or her accident.

Finally, implementing a pre-trained language model could help the healthcare provider increase

Model	AUC	coverage
Current model	95.8	72.7
Custom PLM: admission+medical	**97.7**	**96.1**

Table 3: Results calculated on the accidents that both models have in common in January 2023 (AUC-Coverage).

savings from the correct classification of accidents. Table 3 shows the results of a shadow deployment (a method for simulating the new model's performance in the production environment) comparing the performance of the custom PLM with both administrative and medical reports against the current model in January of 2023. We estimated that with the implementation of this new model into production, more cases would be covered by a model, increasing between 20 to 24%. Considering the increase in coverage and in addition to the increase in the predictive metrics of the continual model, it is estimated that the health care provider could save between 1.5 to 2.5MM US annually. The saving will come from correctly classified cases where the administrative and medical cases were classified as covered, but in reality, they should not be covered.

6 Related work

NLP has been used in applications of the insurance industry in recent years. NLP techniques can be used to analyze vast amounts of unstructured data, such as customer interactions and policy documents, to gain insights and make informed decisions. In several areas within the insurance industry, NLP is being used, including customer service, claims processing, and fraud detection (Ly et al., 2020).

One of the most significant uses of NLP in the insurance sector is customer service (Quarteroni, 2018). To ascertain a client's wants and preferences, NLP techniques can be utilized to evaluate customer interactions such as phone calls and chat chats. Customers may receive more individualized help and recommendations thanks to the utilization of this data. Additionally, regular customer support operations like responding to frequently requested queries, have been automated using NLP algorithms.

Another area where NLP is employed in the insurance sector is claims processing (Popowich, 2005). By automating the analysis and classification of claims, NLP approaches can cut down on the time and resources needed to process claims. To make more educated judgments about claims, NLP algorithms have been employed, for instance, to extract information from claim documents, such as the type of injury and the reason for the accident.

Fraud detection is another area where NLP is being used in the insurance industry (Wang and Xu, 2018). Huge amounts of unstructured data, including policy documents and customer interactions, can be analyzed using NLP approaches to spot probable fraud cases.

7 Conclusion

This work studied the performance of clinical PLMs in a coverage prediction task. Three approaches were implemented, and the best model was compared to the current model used by the health provider. A PLM from scratch was the best-performing model but the most expensive and time-consuming.

Clinical natural language processing has great potential to impact the insurance industry, not only because of the great predictive power they offer but also because it is unnecessary to implement expensive training in the models. As there are no significant differences in performance between the pre-trained model and the fine-tuning with the admission+medical data, by just fine-tuning a PLM we can obtain good results at a lower cost for this downstream task. However, the situation might differ in other NLP tasks that benefit from lexical specificity.

8 Limitations

Some limitations of this work are listed below:

- The architecture and configuration for the custom PLM are the same as bsc-bio-ehr-es. Another architecture and configuration could obtain better results.
- The textual data come from just one provider. Using data from several providers could help with generalization.
- The custom PLM has not been compared with other PLMs in language tasks such as named entity recognition or question answering. This

comparison can help to understand if the custom PLM can outperform available PLMs in other types of tasks.

Ethics Statement

The ethical considerations of this work are related to the data that we used and the models we built. The data was extracted from administrative and clinical records from an insurance and health provider that specialized in labor accidents. Within this data, it is possible to find personal and sensitive information such as personal and company names, addresses, health information, pre-existing conditions, and diagnoses, among others. An anonymization process was not carried out since the model will be used for internal purposes and will not be released. As a process of memorization can occur in the PLM, we believe it is best to keep the model private because privacy attacks can extract personal and sensitive information.

We did not test the models for any bias under any protected field. Therefore, the trained models could benefit certain patients or accidents over others in the insurance decision. If a biased model is deployed in this provider's systems, it could harm patients with their insurance coverage decisions.

Acknowledgements

This work was funded by ACHS and ANID Chile: Basal Funds for Center of Excellence FB210005 (CMM); Millennium Science Initiative Program ICN17_002 (IMFD) and ICN2021_004 (iHealth), Fondecyt grant 11201250, and National Doctoral Scholarships 21211659 (Claudio Aracena).

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Casimiro Pio Carrino, Joan Llop, Marc Pàmies, Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Joaquín Silveira-Ocampo, Alfonso Valencia, Aitor Gonzalez-Agirre, and Marta Villegas. 2022. Pretrained Biomedical Language Models for Clinical NLP in Spanish. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 193–199, Dublin, Ireland. Association for Computational Linguistics.

Hercules Dalianis. 2018. *Clinical Text Mining: Secondary Use of Electronic Patient Records*. Springer International Publishing.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission.

Antoine Ly, Benno Uthayasooriyar, and Tingting Wang. 2020. A survey on natural language processing (nlp) and applications in insurance. *arXiv preprint arXiv:2010.00462*.

Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. 2018. Clinical Natural Language Processing in languages other than English: opportunities and challenges. *Journal of Biomedical Semantics*, 9(1):12.

Fred Popowich. 2005. Using text mining and natural language processing for health care claims processing. *ACM SIGKDD Explorations Newsletter*, 7(1):59–66.

Silvia Quarteroni. 2018. Natural language processing for industry: Elca's experience. *Informatik-Spektrum*, 41(2):105–112.

Matías Rojas, Jocelyn Dunstan, and Fabián Villena. 2022. Clinical Flair: A Pre-Trained Language Model for Spanish Clinical Natural Language Processing. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 87–92, Seattle, WA. Association for Computational Linguistics.

Lewis Tunstall, Leandro von Werra, and Thomas Wolf. 2022. *Natural Language Processing with Transformers*. O'Reilly Media, Inc.

Yibo Wang and Wei Xu. 2018. Leveraging deep learning with lda-based text analytics to detect automobile insurance fraud. *Decision Support Systems*, 105:87–95.

Utterance Classification with Logical Neural Network: Explainable AI for Mental Disorder Diagnosis

Yeldar Toleubay
IBM Research, Japan
yeldar.toleubay@gmail.com

Don Joven Agravante
IBM Research, Japan
don.joven.r.agravante@ibm.com

Daiki Kimura
IBM Research, Japan
daiki@jp.ibm.com

Baihan Lin
Columbia University, USA
baihan.lin@columbia.edu

Djallel Bouneffouf
IBM Research, USA
djallel.bouneffouf@ibm.com

Michiaki Tatsubori
IBM Research, Japan
mich@jp.ibm.com

Abstract

In response to the global challenge of mental health problems, we proposes a Logical Neural Network (LNN) based Neuro-Symbolic AI method for the diagnosis of mental disorders. Due to the lack of effective therapy coverage for mental disorders, there is a need for an AI solution that can assist therapists with the diagnosis. However, current Neural Network models lack explainability and may not be trusted by therapists. The LNN is a Recurrent Neural Network architecture that combines the learning capabilities of neural networks with the reasoning capabilities of classical logic-based AI. The proposed system uses input predicates from clinical interviews to output a mental disorder class, and different predicate pruning techniques are used to achieve scalability and higher scores. In addition, we provide an insight extraction method to aid therapists with their diagnosis. The proposed system addresses the lack of explainability of current Neural Network models and provides a more trustworthy solution for mental disorder diagnosis.

1 Introduction

A mental disorder is a significant deterioration of human thinking, emotional control, or behavior, which is diagnosed clinically and can affect key areas of life. Due to the COVID-19 pandemic, the number of people who suffer from anxiety and depressive illnesses greatly increased in 2020. Initial projections indicate a 26% and 28% increase in anxiety and major depressive disorders respectively during the first year of the pandemic (who). Moreover, every year, 703 000 people commit suicide, with many more attempting to do so. Although people of all ages commit suicides, it is alarming that in 2019 suicide was one of the leading causes of death among young people worldwide (Sui). Furthermore, around 24 million people, or 1 in 300 persons (0.32%), globally suffer from schizophrenia. Although it is not common as other mental disorders, schizophrenia produces psychosis, is associated with significant disability, and may have an impact on all aspects of life, including personal, family, social, educational, and occupational functioning (Sch).

Diagnosis of mental disorders is accomplished through a clinical interview, where a therapist evaluates the mental health of the patient and identifies possible disorders based on symptoms. However, although many mental health issues may be properly treated at low cost, there is still a wide gap between those who need care and those who have access to it. Despite the progress in some countries, there is still a severe lack of effective therapy coverage. Therefore, there is a need for an AI solution that can assist therapists with a diagnosis of mental disorders.

Although current Neural Network (NN) models are powerful and can operate in a wide range of tasks, their effectiveness in mental disorder classification is questionable due to their black-box nature. In this regard, the model explainability is a vital property, which is required to make a diagnosis of mental disorders. While Neural Network models can achieve high scores, therapists may be hesitant to trust such tools and accept classification results if proper explanations are not provided. Because of NN nature, it is impossible to tell whether their predictions are the result of robust features or some spurious clues (Ribeiro et al., 2020). There are attempts to provide interpretable insights in mental disorder diagnosis, such as using topic modeling to extract concepts (Lin et al., 2023a) or inferring psychological properties such as working alliance (Lin et al., 2023b). Although such approaches can enable explainable AI systems for passive assistance (Lin et al., 2023c; Lin, 2022b) or interventional recommendations (Lin et al., 2023e,d) to the therapists, applying these insights directly to the classification problem yields suboptimal performance (Lin et al., 2022). Furthermore, despite

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 439–446
July 14, 2023 ©2023 Association for Computational Linguistics

being able to provide global explanations for the prediction (Mowery et al., 2017), traditional ML models lack scalability and they are not generalizable for broader tasks.

In this regard, Logical Neural Network (Riegel et al., 2020) might be a good solution to the problem. It is a Neuro-Symbolic AI method (NSAI), that combines the learning capabilities of neural networks with the reasoning capabilities of classical logic-based AI. The LNN is a Recurrent neural network architecture in which neurons represent a precisely defined notion of weighted real-valued logic. It has a 1-to-1 relationship to a system of logical formulae. The main problem related to this approach is that it has not been implemented for the supervised learning utterance classification task. Therefore, this work proposes an LNN-based explainable NSAI utterance classification method for mental disorder diagnosis. The model was trained with different predicate pruning techniques to achieve scalability and higher scores. The advantages of the proposed system can be summarized via the following points:

- We propose design of the supervised NSAI method for utterance classification task, where input to the model is predicates from clinical interviews and output is a mental disorder class. After the training the system outputs weighted logical rule to make classifications.
- We propose a predicate pruning methods to improve scalability and generalizability of the model.
- We propose an insight extraction methods which can aid therapists with their mental disorder diagnosis.

1) This paper is organized as follows: Section II details the proposed system, Section III contains experiment results , Section IV provides discussions and future work, and the paper ends with a Conclusion.

2 Supervised learning with LNN

Although NSAI supports data driven training of the network, it encodes knowledge into logic rules with predicates as inputs, where predicates represents a property or a relation. Therefore, NSAI method requires special preprocessing of the dataset to generate predicates and data samples for training and testing purposes. The proposed system consists of two parts Abstract Meaning Representation (AMR) (Zhou et al., 2021) semantic parser and LNN. Fig. 1 shows overall pipeline of the system, first part containing AMR parser is used to convert raw text into classifier input data, and second part is an LNN model which performs rule-based classification.

2.1 Dataset preperation and preprocessing

Counseling and Psychotherapy Transcripts (ale) is a unique and fully anonymized online series of clinical interviews that allows students and researchers to dive deeply into the patient-therapist relationship and track the progress and setback of patients over multiple therapy sessions. These materials bring the mental disorder diagnosis process to life and provide unprecedented levels of access to the widest possible range of clients. Therefore, transcripts of 4 types of mental disorders, which are anxiety, depression, suicidal thoughts, and schizophrenia, from this dataset are used in our training and evaluation of the model. Table 1 shows the details of the dataset; in our simulations, only 12 sessions of clinical interviews have been used due to the computational constraints of semantic parsing. An example from the transcript has shown in Fig 2. In our experiments, a transcript is a full clinical interview between a patient and a therapist, while an utterance represents a full response of the patient to a specific question from the therapist.

Class	**Number of total sessions**	**Number of used session**
Anxiety	498	12
Depression	377	12
Suicidal	12	12
Schizophrenia	71	12

Table 1: Details of the dataset.

As mentioned before, LNN requires a special data structure to function. AMR parser is used for generation of predicates by extracting the semantics of the utterance and converting semantics into a graph, where nodes (keys) represent concepts and edges (values) represent relations to concepts. Example of AMR Representation is shown in Fig 1. AMR Representation keys and values are combined to generate predicates as shown in Table 2.

Moreover, a training and a testing sample is input to the model and is obtained by using AMR parser over an utterance. Furthermore, a sample contains all predicates that has been mined from dataset and

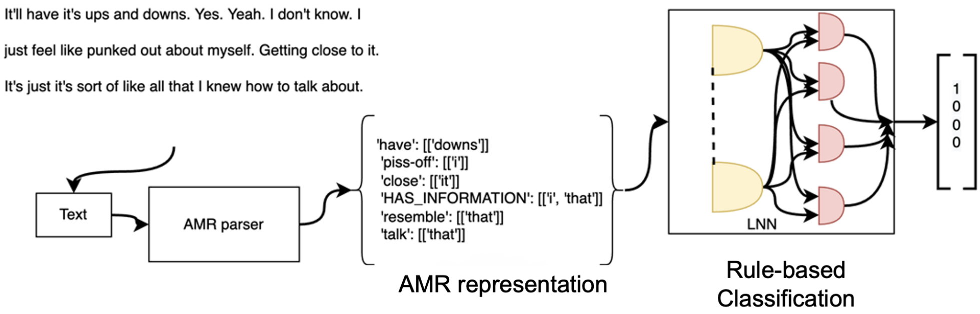


Figure 1: The overview of the proposed system.

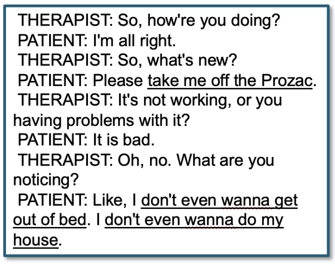


Figure 2: Examples of a dataset transcript.

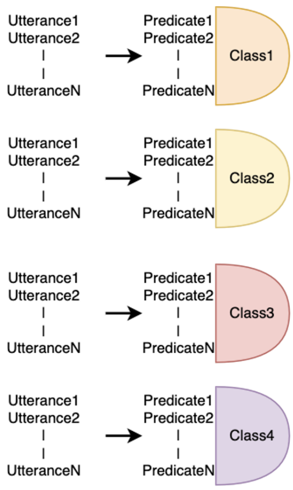


Figure 3: Proposed LNN architecture for mental disorder diagnosis.

the corresponding output of parser as groundings. The values of the groundings are assigned according to the presence of the particular predicates in the parsed utterance, which means only predicates that results from that particular utterance assigned with *TRUE* grounding for that particular sample. In this regard, certain combinations of predicates might repeat over multiple classes and the proposed design takes into account this issue.

2.2 Proposed system details

LNN is a core of the model, which has only few differences from regular neural network. The main difference of LNN is that its neural parameters are limited such that the truth functions of the relevant logical gates govern the behavior of the neurons. Moreover, LNN neuron has more parameters compared to dense neuron, since it keeps both upper and lower bounds to the corresponding subformula or predicate.

The proposed LNN architecture has 4 *AND* logic gates that act as binary classifiers for each mental disorder class. Predicates are inputs to the logic gates, while model is trained by samples generated from utterances. Those samples show truth values for formulae. After the training model outputs set of weight for each predicate and outputs a tensor of lower and upper bounds as a score for a particular input. In our experiments the each logic gate is evaluated as a binary classifier that classifies according to some threshold, thus the upper and lower bounds are averaged to obtain a single score. The score S for each class is obtained via following equation:

$$S = w_1\ P_1\ (x_1) + w_2\ P_2\ (x_2) + .. + w_N\ P_N\ (x_N) \tag{1}$$

where P is a predicate, w is a weight obtained from training and x_i is a grounding for each predicate in a sample.

The proposed system will be evaluated as a separate binary classification models for each gate by True Positive Rate (TPR) and False Positive Rate (FPR) metrics. The TPR indicates the proportion of all available positive samples that contain correct positive results. In contrast, FPR quantifies the pro-

Input	Predicates			Output
	AMR Representation			
	Keys	Values	Output of Parser	Class
sample0	HAS_POSSESSION	your medication	TRUE	Depression
	HAS_POSSESSION	any details	FALSE	
	HAS_POSSESSION	downs	FALSE	
	HAS_POSSESSION	just awkward thing	FALSE	
	have	your medications	FALSE	
	have	any details	FALSE	
	have	downs	TRUE	
	have	just awkward thing	FALSE	
	talk	your medication	FALSE	
	talk	any details	FALSE	
	talk	downs	FALSE	
	talk	just awkward thing	FALSE	
sample1	HAS_POSSESSION	your medications	TRUE	Anxiety
	HAS_POSSESSION	any details	TRUE	
	HAS_POSSESSION	downs	FALSE	
	HAS_POSSESSION	just awkward thing	TRUE	
	have	your medications	FALSE	
	have	any details	FALSE	
	have	downs	FALSE	
	have	just awkward thing	FALSE	
	talk	your medications	FALSE	
	talk	any details	TRUE	
	talk	downs	FALSE	
	talk	just awkward thing	FALSE	

Table 2: LNN for supervised learning inputs and outputs – predicates, data samples and class

portion of available negative samples that contain incorrectly positive results. Moreover, the receiver operating characteristic (ROC) curve is created by plotting the TPR against the FPR at various threshold values.

$$TPR = \frac{True\ Positives}{True\ Positives\ + False\ Negatives} \tag{2}$$

$$FPR = \frac{False\ Positives}{True\ Negatives\ + False\ Positives} \tag{3}$$

2.3 Predicate pruning methods

Predicates play a crucial role in LNN training and can greatly affect the accuracy of the model. Table 4 shows that 48 transcripts result in more than 19000 predicates. However, according to Table 3 a preliminary simulation results show that for a linear

# of predicates	Training time (s)
710	4.49
1415	16.54

Table 3: Results of training time with different number of predicates for an LNN model with 2 Logic gates.

increase in number of predicates, LNN requires exponential increase in training time. Therefore, there is a need for predicate pruning methods, which will help to chose predicates that contribute the most towards the correct diagnosis. Thus, similarity, exclusivity and frequency based predicate pruning methods has been proposed to reduce number of predicates.

Similarity pruning. Simulations has shown that AMR Parser returns multiple variants of values per one key. Often, those values contain repeating phrases. Thus, it is possible to group all those

Class	Original predicates	Similarity pruning	Exclusive Pruning	F=1	F=2	2<F<10	F>9
Anxiety	5529	2773	245	3152	216	150	14
Depression	7227	3532	472	2174	454	133	12
Suicidal	6067	3213	230	2839	197	160	17
Schizophrenia	3746	1914	96	1718	102	87	7

Table 4: Number of original predicates and number of predicates after similarity, exclusive and frequency pruning methods.

lookalike predicates by taking a predicate that contains possible repetitions, e.g. instead of taking both "HAS_POSSESSION_my sister's birthday" and "HAS_POSSESSION_sister's birthday", one can take only the first one.

Frequency pruning. In traditional ML word count can show the importance of some features for a specific class. Using the same logic, it was assumed that predicates that are encountered frequently in sessions will have higher impact on model training. Thus, predicates has been analyzed in terms of repetitions across sessions and have been grouped according to the specified frequencies.

Exclusive pruning. Since transcripts are conversations between patients and therapists, there are many repeating predicates between classes. Thus, it was suggested that predicates belonging only to a class will avoid contradictions in the model as well as will have higher correlation to a specific class. Therefore, predicates repeating between classes predicates that are repeated only once have been removed.

3 Experiment results

In this section experiment results for predicate pruning and LNN model evaluation will be provided. Table 4 shows number of predicates for a particular pruning method. Similarity pruning method prunes almost half of the original predicates. Furthermore, Exclusive and Frequency pruning methods have been applied on top of the similarity pruning method. Results for the Exclusive pruning shows that Depression class has twice of Anxiety and Suicidal predicates and 5-times of Schizophrenia predicates. Moreover, results for the different frequencies show that majority of the predicates (43%) repeat just once, while the higher frequency rates have fewer predicates.

The LNN models have been trained using different pruning methods and have been compared with Deep Learning (DL) and LNN baselines. The number of predicates and training samples are shown in the Table 5. The LNN models have been trained with supervised loss, which targets the labels with learning rate of 0.05, for 50 epochs. The main difference between LNN models is in the predicates. The details of each model are summarized below:

- *DL baseline*. As a DL baseline pre-trained BERT (Devlin et al., 2018) model and Bert tokenizer with a maximum sequence length of 256 inputs have been selected for finetuning. The model has been trained for 10 epochs using Adam optimizer with a learning rate of 10^{-5}.
- *LNN baseline*. The predicates for the LNN baseline have been selected randomly from Similarity predicates. The number of predicates for each class varies from 340 to 380 predicates.
- *Frequency pruning models*. Several models with different frequencies have been trained to examine the effectiveness of the frequency pruning methods. $F > Threshold$ stands for the model with predicates repeating with a frequency higher than the threshold value. The $F > 5$ *balanced* ensures that classes are balanced in terms of predicates. The remaining predicates have been chosen from a lower frequency.
- *Exclusive pruning models*. The exclusive pruning method is used in combination with similarity and frequency pruning methods. In the simulations, Frequency pruning prunes predicates that repeat only once. Then the exclusive pruning removes all the repeating predicates between classes.

According to the Table 5, Frequency predicates does not have a significant effect on model performance when they are applied alone, since the are

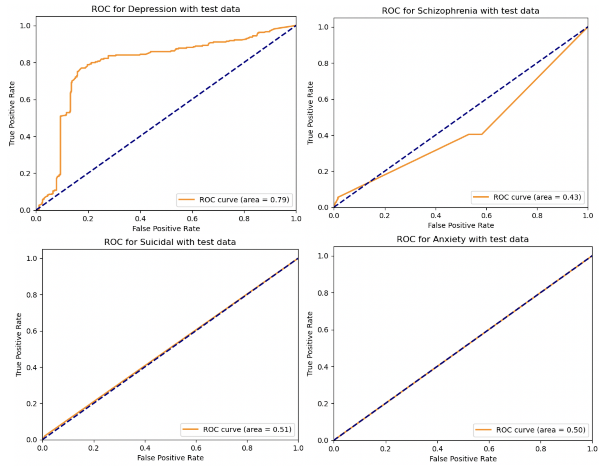


Figure 4: AUC ROC curves for each class in testing.

under (AUC) the ROC curve is around 0.5, which is close to the random classifier. Moreover, LNN baseline with 1000 predicates and 10000 training samples performed surprisingly well for the Anxiety class, achieving AUC of 0.76. Baseline DL model has AUC scores higher than 0.72 for all classes when treated as a binary classifier for each class. However, since therapist cannot use this data explicitly, the accuracy for the multi-class classification has higher importance for this case and DL model can provide only 58% accuracy in such setting. Exclusive predicates model have shown a good performance overall. It reached AUC of 0.79 for the depression class and 0.57 for schizophrenia.

4 Discussions and Future Work

Scaling of the LNN is a significant issue which requires selection of the right predicates. Pruning of the predicates essentially limits the knowledge base of the LNN, thus it is important to understand the effect of the predicates on model performance. Frequency predicate models have not shown great results, the possible explanation for that behavior can be found in predicate analysis. The analysis shows that predicates with higher frequencies also tend to be inclusive for several classes. Such predicates might be extracted from common dialogue phrases, that are common to regular conversations. Thus, it is more difficult to learn for LNN in such circumstances and it might lead to a behavior similar to the random classifiers'. Moreover, variation in the frequency thresholds did not affect the overall performance of the LNN model. Thus, it can be concluded that frequency predicates cannot provide a quality selection of the predicates when they are applied alone. Furthermore, in the case of exclusive predicates, the model has learned depression class better than others. It can be explained by the depression class possessing more exclusive predicates compared to other classes. Interestingly, the model has learned to identify non-schizophrenia samples better than schizophrenia samples. Possible reasoning for that is fewer predicates for schizophrenia compared to other classes. Furthermore, some mental disorders have the same symptoms, and exclusive pruning eliminate such predicates from the training, which might lead to limited diagnostic abilities. Thus, exclusive predicates should be combined with other methods to provide trade-off between generalization and exclusivity of predicates.

Another challenge of this line of work is the usage of AI for mental disorder diagnosis. As pointed out in (Lin, 2022a), one significant challenge is related to the privacy and security of patient data. To train the model, the system requires access to sensitive patient data, which must be protected from unauthorized access or misuse. There is also a concern that the use of AI in mental health diagnosis may lead to the stigmatization of individuals with mental disorders. In this work, we have deidentified all the sessions and all the transcripts are obtained under proper license and consent. We would also like to point out that the system may not work for all individuals, which could lead to misdiagnosis or lack of diagnosis, leading to harm to the patient. Therefore, the ethical challenge lies in ensuring the system's reliability, fairness, and transparency and balancing the use of AI with the need for human involvement in mental health diagnosis and treatment, as part of the future work.

The main advantage of the LNN over DL is in its explainability. It is possible to extract predicates with high weights for the each class and to examine which predicates contribute to the result significantly. Table 6 shows the predicate semantics analysis for each class after the training. Predicates of depression and anxiety suicidal classes are mostly related to the first-person and third-person actions respectively, while people with anxiety tend to talk about feelings more. In addition, predicates of the schizophrenia class tend to relate to the medical terms. This overlaps with overall content of the transcripts and predicates that posses high weights can be used to give insights to therapist during the diagnosis of the patients.

	# of training samples	# of predicates	Suicidal (AUC)	Depression (AUC)	Anxiety (AUC)	Schizophrenia (AUC)
BaselineLNN	10000	1000	0.50	0.50	0.76	0.52
BaselineDL	10000	N/A	0.73	0.83	0.81	0.72
			Accuracy for multiclass = 0.58			
F >5	3947	87	0.55	0.58	0.55	0.52
F>5. Balanced	3947	141	0.55	0.59	0.52	0.55
F >3	3605	349	0.54	0.56	0.53	0.52
F>6	3947	81	0.55	0.53	0.56	0.50
Exclusive predicates and F>1	3947	981	0.51	0.79	0.50	0.43

Table 5: AUC ROC scores for DL baseline, LNN baseline and proposed pruning methods .

	Grouping of predicates	Top 1 weight	Top 2 weight	Top 3 weight
Depression	Related to first-person actions	Do_i	Come _they	Resemble _what
Anxiety	Related to feelings	get_it	look_it	have-rel-role_my mom
Schizophrenia	Related to medical terms	give_me	resemble _things	do_it
Suicidal	Related to third-person actions	have-manner _sense	put_it	do_ everything

Table 6: Analysis of the semantics of the predicates for the each class.

4.1 Future work

Overall, it is evident that predicates are too specific from the number of predicates with a frequency of 1, which might be a possible explanation for the poor performance of the model overall. Therefore, they might require some generalization of the predicates. One of the promising methods for that is to use synonym-based predicates. By using thesaurus dictionaries, it is possible to cluster all the keys and values of the AMR representations and use only one variants for the synonyms. That way, it might be possible to reduce the number of predicates significantly and achieve their generalization.

Another possible way to enhance the model is the explore LNN and DL hybrid approach. By using LNN scores it is possible to train some dense layers with SoftMax to predict classes in the multiclass setting. In a such way it will be more convenient to compare LNN results with DL solutions while keeping the explainability of the LNN.

5 Conclusion

Mental disorders are a significant issue that is affecting more people every year. Therefore, explainable AI mental disorder diagnosis through utterance classification can aid the therapist in their practice. In this work, a supervised learning setting for the LNN has been proposed to address this issue. Moreover, predicate pruning methods based on the similarity, frequency, and exclusivity of the predicates are analyzed in terms of training performance. Overall, the model trained with exclusive predicates shows the best results among the pruning methods, and acheived AUC ROC of 0.79 for the depression disorder. Finally, explainability of the LNN diagnosis has been shown by analyzing significant predicates for each class and extracting the predicates with high weights.

References

Counseling and psychotherapy transcripts series.

Mental health and covid-19: Early evidence of the pandemic's impact: Scientific brief.

Schizophrenia - key facts.

Suicide - key facts.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Baihan Lin. 2022a. Computational inference in cognitive science: Operational, societal and ethical considerations. *arXiv preprint arXiv:2210.13526*.

Baihan Lin. 2022b. Voice2alliance: automatic speaker diarization and quality assurance of conversational alignment. In *INTERSPEECH*.

Baihan Lin, Djallel Bouneffouf, Guillermo Cecchi, and Ravi Tejwani. 2023a. Neural topic modeling of psychotherapy sessions. In *International Workshop on Health Intelligence*. Springer.

Baihan Lin, Guillermo Cecchi, and Djallel Bouneffouf. 2022. Working alliance transformer for psychotherapy dialogue classification. *arXiv preprint arXiv:2210.15603*.

Baihan Lin, Guillermo Cecchi, and Djallel Bouneffouf. 2023b. Deep annotation of therapeutic working alliance in psychotherapy. In *International Workshop on Health Intelligence*. Springer.

Baihan Lin, Guillermo Cecchi, and Djallel Bouneffouf. 2023c. Helping therapists with nlp-annotated recommendation. In *Joint Proceedings of the ACM IUI Workshops*.

Baihan Lin, Guillermo Cecchi, and Djallel Bouneffouf. 2023d. Psychotherapy AI companion with reinforcement learning recommendations and interpretable policy dynamics. In *Proceedings of the Web Conference 2023*.

Baihan Lin, Guillermo Cecchi, and Djallel Bouneffouf. 2023e. SupervisorBot: NLP-Annotated Real-Time Recommendations of Psychotherapy Treatment Strategies with Deep Reinforcement Learning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*. International Joint Conferences on Artificial Intelligence Organization.

Danielle L. Mowery, Hilary Smith, Tyler Cheney, Greg Stoddard, Glen A. Coppersmith, Craig J. Bryan, and Mike Conway. 2017. Understanding depressive symptoms and psychosocial stressors on twitter: A corpus-based study. *Journal of Medical Internet Research*, 19.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist.

Ryan Riegel, Alexander Gray, Francois Luus, Naweed Khan, Ndivhuwo Makondo, Ismail Yunus Akhalwaya, Haifeng Qian, Ronald Fagin, Francisco Barahona, Udit Sharma, Shajith Ikbal, Hima Karanam, Sumit Neelam, Ankita Likhyani, and Santosh Srivastava. 2020. Logical neural networks.

Jiawei Zhou, Tahira Naseem, Ramón Fernandez Astudillo, Young-Suk Lee, Radu Florian, and Salim Roukos. 2021. Structure-aware fine-tuning of sequence-to-sequence transformers for transition-based AMR parsing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6279–6290, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Survey of Evaluation Methods of Generated Medical Textual Reports

Yongxin Zhou Fabien Ringeval François Portet
Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000, Grenoble, France
yongxin.zhou@univ-grenoble-alpes.fr,
{fabien.ringeval, Francois.Portet}@imag.fr

Abstract

Medical Report Generation (MRG) is a subtask of Natural Language Generation (NLG) and aims to present information from various sources in textual form and synthesize salient information, with the goal of reducing the time spent by domain experts in writing medical reports and providing support information for decision-making. Given the specificity of the medical domain, the evaluation of automatically generated medical reports is of paramount importance to the validity of these systems. Therefore, in this paper, we focus on the evaluation of automatically generated medical reports from the perspective of automatic and human evaluation. We present evaluation methods for general NLG evaluation and how they have been applied to domain-specific medical tasks. The study shows that MRG evaluation methods are very diverse, and that further work is needed to build shared evaluation methods. The state of the art also emphasizes that such an evaluation must be task specific and include human assessments, requesting the participation of experts in the field.

1 Introduction

Medical Report Generation (MRG)[1] (Chen et al., 2022) is a subfield of Natural Language Generation (NLG) and aims to present information from various sources in textual form to synthesize salient information, with the goal of reducing the time spent by domain experts in writing medical reports and providing supporting information for decision-making.

MRG includes all systems used to generate medical documentation, such as the generation of radiology reports (Chen et al., 2021), discharge summaries or SOAP notes (Krishna et al., 2021), etc.

The evaluation of automatically generated texts is important for the validity of the systems, especially in the era of ChatGPT and its increased usage in medical domain (Ma et al., 2023). However, it has often been reported that the two main approaches of NLG evaluation – human-based and automatic-based – need to be improved (Reiter and Belz, 2009; van der Lee et al., 2019). On the one hand, the use of automatic metrics for system quality assessment has been criticized for two main points: they are uninterpretable and do not correlate well with human evaluations (Qader et al., 2018; van der Lee et al., 2019). On the other hand, the deployment of human evaluation is sometimes too complex. Indeed, crowdsourcing solutions are not always reliable or adequate while using experts, when available, can lead to high costs. Furthermore, there is a lack of unified framework/criteria (van der Lee et al., 2019).

Although there have been many surveys on NLG evaluation (Gkatzia and Mahamood, 2015; Amidei et al., 2018; van der Lee et al., 2019; Sai et al., 2022), there is no systematic study on report generation evaluation in the medical domain. It is worth mentioning Messina et al. (2021), which has done survey work in the area of automatic report generation from medical images, including an analysis of evaluation methods.

In this paper, we focus on MRG tasks and their evaluations. We include more than 20 papers in this study, classified into two broad categories: text-to-text and data-to-text. We summarize the evaluation methods currently in use and make recommendations for future evaluation of MRG systems.

2 Medical Report Generation

2.1 Paper Search and Selection

To perform a literature review of MRG evaluation, we first followed the paper list introduced in a survey paper on dialogue summarization (Feng et al., 2022). We then extended our search to search engine such as Google Scholar. Papers reviewed in this study were primarily from the major NLP con-

[1] It is also called 'Clinical Note Generation'.

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 447–459
July 14, 2023 ©2023 Association for Computational Linguistics

ferences, including ACL, EMNLP, NAACL, etc. In addition, we also included some articles from journal and domain-specific workshops, including NLPMC, ClinicalNLP, etc. [2]

2.2 MRG Systems and Applications

Since the seminal work of Kukich (1983), there have been several kinds of medical report systems and applications, such as the generation of psychiatric case notes (Kazi and Kahanda, 2019), the generation of consultation notes from transcripts (Papadopoulos Korfiatis et al., 2022), the generation of radiology reports (Chen et al., 2021), nurse-patient summaries (Liu et al., 2019), counseling (conversation) summarization (Srivastava et al., 2022), discharge summaries or clinical notes (Krishna et al., 2021), and even data augmentation for other medical tasks (Kocabiyikoglu et al., 2021).

Joshi et al. (2020) provided a general definition of a medical report in the case of medical dialogue summarization: "*the medical report captures and summarizes the important parts of the medical conversation necessary for clinical decision-making and subsequent follow-ups.*"

Despite the diversity of their tasks, structures and audiences, the main characteristics of MRG remain similar, namely the use of the documentation and the subsequent use of the diagnosis, which can also be used for administration and by institutions, subsequently referenced by clinicians and retained by patients. The main objectives of such systems in clinical practice are to reduce the time spent by clinicians on manual writing and facilitate medical decision-making.

2.3 Main NLG approaches to MRG

According to the different types of input sources, MRG can be divided into two categories: 1) Text-to-text, e.g. summarizing medical dialogues; 2) Data-to-text, e.g. automatically generating reports from medical images or other data. After classifying the different approaches according to the input, we then further categorized different works in the literature according to their tasks, with examples in Table 1.

2.3.1 Text-to-text

There are three main tasks in the Text-to-Text category: 1) summarizing medical dialogues/conversations, including spoken conversations and online medical conversations; 2) summarizing hospital stays/hospitalizations; and 3) summarizing medical reports, where the original reports may come from different domains, such as radiology reports or general clinical reports.

The most common work in text-based MRG is that of **summarizing medical conversations**, where the input source can be either transcripts of clinician-patient spoken conversations (Kazi and Kahanda, 2019; Enarvi et al., 2020; Liu et al., 2019; Krishna et al., 2021; Zhang et al., 2021; Molenaar et al., 2020; Srivastava et al., 2022; Lacson et al., 2006; Moramarco et al., 2022; Yim and Yetisgen, 2021); or online medical conversations (Chintagunta et al., 2021; Nair et al., 2021; Joshi et al., 2020; Song et al., 2020; Chen et al., 2022).

Regarding the **summarization of hospital stays**, some work (Di Eugenio et al., 2014; Acharya et al., 2016) used both physician discharge notes (free text) and the structured nursing documentation (such as nursing plans of care) to generate a unique summary. Others generated summaries from long-form hospital admissions (Adams et al., 2023).

Moreover, work has also been carried out on **summarizing medical reports**. For example, Moramarco et al. (2021) used the *MTSamples* dataset to fill automatically the '*description field*' of a medical report based on the information present in the overall report. In addition, radiology report summarization (Zhang et al., 2020; Karn et al., 2022) is intended to produce a concise and easily comprehensible '*IMPRESSIONS*' section from the rest of the radiology report. The '*IMPRESSIONS*' section of a radiology report is considered a summary of the radiologist's reasoning and conclusions, which helps the referring physician confirm or exclude certain diagnoses (Karn et al., 2022).

2.3.2 Data-to-text

As presented in Table 1, there are different tasks in the Data-to-Text category: 1) generation of reports from medical images, such as radiology images, brain image data.; 2) generation of text summaries from intensive care data; and 3) generation of medical reports from multimodal inputs; 4) other applications such as the generation of tailored smoking cessation letters based on responses to a smoking questionnaire (Reiter et al., 2003).

In order to meet the growing demand of image-based diagnosis from patients using artificial in-

[2]The list of papers about MRG is available at https://github.com/yongxin2020/Medical-Report-Generation-Papers

Category	Task	Description	Examples
Text-to-text	Medical dialogue/conversation summarization	Transcribed conversations	Srivastava et al. (2022); Molenaar et al. (2020); Zhang et al. (2021); Krishna et al. (2021); Liu et al. (2019); Enarvi et al. (2020); Kazi and Kahanda (2019); Lacson et al. (2006); Moramarco et al. (2022); Yim and Yetisgen (2021)
		Online medical conversations	Nair et al. (2021); Chintagunta et al. (2021); Song et al. (2020); Joshi et al. (2020); Chen et al. (2022)
	Summarization of hospital stays/hospitalizations	Sources from physician discharge notes and nursing plans of care	Acharya et al. (2016); Di Eugenio et al. (2014)
		Long-form hospital admissions	Adams et al. (2023)
	Medical report summarization	Radiology report	Zhang et al. (2020); Karn et al. (2022)
		Clinical reports	Moramarco et al. (2021)
Data-to-text	Report Generation from medical images	Radiology	Miura et al. (2021); Chen et al. (2021); Yan et al. (2021); Chen et al. (2020); Lovelace and Mortazavi (2020); Nooralahzadeh et al. (2021); Qin and Song (2022)
		Brain imaging data	Jordan et al. (2014)
	Clinical Data Summarization	Intensive care data	Portet et al. (2009); Reiter et al. (2008)
	Automated Medical Reporting	Sources from multimodal inputs: audio, video and sensor data from medical consultations	Maas et al. (2021)
	Other applications	Generation of tailored smoking cessation letters	Reiter et al. (2003)

Table 1: Categorization of MRG according to system inputs and tasks.

telligence and applying image captioning to the medical field, radiology report generation is the subject of continuous work and growing interest from researchers, which aims to describe radiology images with professional quality reports (Chen et al., 2020, 2021; Lovelace and Mortazavi, 2020; Nooralahzadeh et al., 2021; Yan et al., 2021; Qin and Song, 2022; Miura et al., 2021; Liu et al., 2021). Such research has also been applied to generation of clinician reports from brain imaging data (Jordan et al., 2014).

Besides images, the summarization of physiological data as also been the subject of research. In the *BabyTalk* project, Portet et al. (2009) presented a prototype that generates textual summaries of about 45 minutes of continuous physiological signals and discrete events. Their evaluation with physicians showed that text summaries could be an effective decision-support aids for clinicians.

To cope with the high workload due to the time required for proper documentation, Maas et al. (2021) presented a real-time automated report of the interaction between care provider and patient, taking multimodal inputs that include audio, video, and sensor data from medical consultations, and introducing knowledge graphs – the Patient Medical Graph. They used speech and action recognition technology to first transform multimodal inputs into text before formally representing them and generating reports.

3 Evaluation in NLG

In this section, we will briefly introduce automatic evaluation in NLG and then will look at human evaluation with some current practices.

3.1 Automatic Evaluation

Automatic evaluation is popular because it is cheap and fast and it is widely used in benchmarking activity and for system development.

There is a wide range of automatic evaluation metrics used in NLG (Sai et al., 2022) and we will restrict to the two most popular: 1) the corpus-based metrics and 2) the trainable metrics. The corpus-based metrics rely on a set of reference texts (i.e. gold standard outputs) to which system outputs are compared. For instance, it can be based on n-grams, such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004); or on edit distance: WER (Woodard and Nelson, 1982), TER (Snover et al., 2006), etc.

Most automatic metrics require gold references, but these are not always available. Reference-less metrics, where neural models are trained to predict human ratings from texts (e.g. regression models trained on ratings data), are getting more and more attention recently. For example, BLEURT (Sellam et al., 2020) is a learned evaluation metric for English to predict human judgments. It relies on the BERT model using unsupervised techniques with millions of synthetic examples. There are also metrics based on question-answer pairs on a given source document (Scialom et al., 2021; Rebuffel et al., 2021), for example QuestEval (Scialom et al., 2021) uses pre-trained models to assess if two different inputs contain the same information. Note that QuestEval can also be used with references.

To summarize, ROUGE scores assess the similarity between candidates and references based on the overlap of unigrams, bigrams, and the longest common sequence, likewise for BLEU; while BLEU focuses on precision, ROUGE focuses on recall. BERTScore evaluates the similarity between candidates and references at token level, using contextual embeddings from BERT, while QuestEval assesses whether a summary contains all the relevant information from its source document and BLEURT attempts to model human judgments.

However, automatic evaluation metrics have their limitations and do not sufficiently reflect human judgments of system performance (Novikova et al., 2017).

3.2 Human Evaluation

Human evaluation is considered the most informative form of evaluation of NLG systems, but it can be expensive and time-consuming since qualified human evaluators have to be recruited. Hence, human evaluation is difficult to scale up unless using crowd sourcing approaches but these are difficult to apply in medicine for expertise and privacy reasons.

There are several commonly used methods for human evaluation, including the Likert scale scoring and pairwise comparison for general text generation, as well as Pyramid and binary factuality evaluation specifically designed for summarization (Gao et al., 2023). Some other methods consist in evaluating how much information can be extracted back from the text in a formal form (A. Baez Miranda et al., 2015).

It has been argued that human evaluation approaches are difficult to compare (van der Lee et al., 2019; Belz et al., 2020) since different tasks and criteria are used (with different names). Furthermore, only a small number of papers provide full details of human evaluation experiments (Belz et al., 2020). Howcroft et al. (2020) concluded that due to a pervasive lack of clarity in reports and extreme diversity in approaches, human evaluation in NLG presents as extremely confused in 2020, and that the field is in urgent need of standard methods and terminology.

In addition, van der Lee et al. (2019) provided an overview of best practices in human evaluation of automatically generated text based on papers published at INLG (N=51) and ACL (N=38) in 2018, and released a list of best practices on 7 different topics: general, criteria, sampling, annotation, measurement, design and statistics.

4 Evaluation for Text-based Medical Report Generation Systems

In the following subsections, we summarize the automatic measures and human evaluation used in the literature in Table 2.

4.1 Automatic Metrics

We divide automatic metrics into two categories: text quality and medical concept correctness, for medical correctness there are two subcategories: those based on reports and those for auxiliary or intermediate tasks.

4.1.1 Text Quality

For automatic text quality assessment, there are word-overlap-based metrics like ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002), embedding-based metrics such as BERTScore (Zhang* et al., 2020); learned evaluation metrics like BLEURT (Sellam et al., 2020); and evaluation metrics which rely on question answering like QuestEval (Scialom et al., 2021).

ROUGE (Lin, 2004) has been widely used in MRG tasks: some papers reported only ROUGE-L F1 score (Joshi et al., 2020; Enarvi et al., 2020; Nair et al., 2021), while some others reported ROUGE-1, ROUGE-2, and ROUGE-L scores (Song et al., 2020; Yim and Yetisgen, 2021). Yim and Yetisgen (2021) reported BLEU (Papineni et al., 2002) in addition to ROUGE performances across different note sections.

Category	Subcategory	Metric or evaluation		Used by papers
Automated Metrics	Text quality	ROUGE (-1, -2, -L)		Srivastava et al. (2022); Zhang et al. (2021); Chintagunta et al. (2021); Krishna et al. (2021); Song et al. (2020); Joshi et al. (2020); Enarvi et al. (2020); Liu et al. (2019); Nair et al. (2021); Yim and Yetisgen (2021); Chen et al. (2022); Ben Abacha et al. (2023)
		BLEU		Yim and Yetisgen (2021)
		BERTScore		Ben Abacha et al. (2023)
		QuestEval (QAE) score		Srivastava et al. (2022)
		Bleurt Score (BS)		Srivastava et al. (2022); Ben Abacha et al. (2023)
	Medical correctness (report based)	Medical Concept Coverage	-	Joshi et al. (2020); Chintagunta et al. (2021); Nair et al. (2021); Chen et al. (2022)
			Factual correctness (F1)	Enarvi et al. (2020)
			Concept-based (F1/R/P)	Zhang et al. (2021)
			Items included (P, R, false positives)	Molenaar et al. (2020)
			Fact-based (Fact-Core + Fact-Full)	Ben Abacha et al. (2023)
		Negation Correctness	Negex (Harkema et al., 2009) is used to determine negated concepts	Joshi et al. (2020); Chintagunta et al. (2021); Nair et al. (2021)
		Disease diagnosis	Regex-based Diagnostic Accuracy (RD-Acc)	Chen et al. (2022)
	Medical correctness (auxiliary tasks)	Classification of electronic health record (EHR) categories	AUROC (Area Under the ROC Curve) scale	Kazi and Kahanda (2019)
		Utterances classification	Multilabel classification of noteworthy utterances (Accuracy, Ma-AUC, Ma-F1, Mi-AUC, Mi-F1)	Krishna et al. (2021)
			Utterance tags classification (PD/DT/OT labels) (precision, recall, and F scores)	Song et al. (2020)
			Dialogue turns classification (precision, recall, and F measure)	Lacson et al. (2006)
Human Evaluation	Intrinsic (Text quality)	Relevance		Srivastava et al. (2022); Zhang et al. (2021)
		Consistency		Srivastava et al. (2022)
		Fluency		Srivastava et al. (2022); Zhang et al. (2021); Ben Abacha et al. (2023)
		Coherence		Srivastava et al. (2022)
		Missing		Zhang et al. (2021)
		Hallucination		Zhang et al. (2021)
		Repetition / Non-redundancy		Zhang et al. (2021); Ben Abacha et al. (2023)
		Contradiction		Zhang et al. (2021)
		Extent of verbatim copying from conversation		Krishna et al. (2021)
		Comprehensiveness		Krishna et al. (2021)
		Sentence-level (factually correct, incoherent, irrelevant, redundant, or placed under an inappropriate section)		Krishna et al. (2021)
		Categories relevancy, factual accuracy, writing-style, completeness, and overall		Yim and Yetisgen (2021)
	Intrinsic (Medical Correctness)	Factually correct and medically relevant information		Joshi et al. (2020); Chintagunta et al. (2021)
		Critical Omissions, Hallucinations, Correct Facts, Incorrect Facts based on fact extraction		Ben Abacha et al. (2023)
	Extrinsic	Clinical acceptability framework (Sekhon et al., 2017)		Srivastava et al. (2022)
		List of key questions based on topics that commonly arise between hemodialysis patients and caregivers		Lacson et al. (2006)
		Post-editing (Post-edit times, errors into "incorrect statements" and "omissions")		Moramarco et al. (2022)

Table 2: Summary of evaluation methods used in the articles reviewed.

Commonly used automated metrics, such as ROUGE and BLEU, have their limitations and are known to correlate poorly with human evaluations (van der Lee et al., 2019). Therefore, other measures such as QuestEval (Scialom et al., 2021) and BLEURT (Sellam et al., 2020) which can correlate better with human judgements are used, Srivastava et al. (2022) used these two scores in addition to ROUGE. In addition to ROUGE and BLEURT, Ben Abacha et al. (2023) also reported BERTScore.

4.1.2 Medical Correctness: Report-based

In the study by Joshi et al. (2020) two measures are defined: *Medical Concept Coverage* and *Negation Correctness*. The former captures the coverage of medical terms in the predicted summaries to the gold standard reference, while the latter identifies the negated status of medical concepts. In the healthcare domain, it is crucial to ensure high-quality results in terms of accurate usage of medical terms and capturing negation.

The evaluation of *Concept* involves using specific and in-house extractors and Named Entity Recognition (NER) models. They refer to domain-specific knowledge and compare the match of extracted concepts to standardized health and biomedical vocabularies, such as the Unified Medical Language Systems (UMLS). Several studies have utilized concept correctness measures, such as F1-score, precision, recall, and false positives, at various levels of granularity, including the report level and section level.

For instance, Joshi et al. (2020) used an in-house medical entity extractor to match concepts in the summary to UMLS, and they used Negex (Harkema et al., 2009) to determine negated concepts. Medical concepts in the predicted summary that were not present in the original conversation would be false positives, and vice versa for false negatives. Among the concepts present in the predicted summary, they assessed precision and recall to see whether the predicted negation was accurate for the decoded concepts and computed a Negation F1. The set of automatic metrics proposed was then used in several works (Chintagunta et al., 2021; Nair et al., 2021).

If in-house entity extractor to match concepts in the summary to UMLS have been frequent (Soldaini and Goharian, 2016; Joshi et al., 2020; Zhang et al., 2021), entity extraction using machine learning has appeared recently, which is even more specific to the task. For instance, Enarvi et al. (2020) employed a machine learning-based clinical fact extractor to measure factual correctness by extracting medical facts from both the predicted reports and the ground-truth reports, such as conditions and medications, as well as their attributes such as body part, severity, or dosage, then calculating the F1 score from these two sets. To compute Concept-F1, Chen et al. (2022) used the medical entity extractor – BERT-CRF (Devlin et al., 2019) trained on their NER task to match entities in the predicted summary to the reference summary.

Similarly, Ben Abacha et al. (2023) used "fact-based metrics (Fact Scores)", which is a machine learning-based medical fact extraction system. The Fact Score metric measures the F1-score of medically relevant facts extraction, is used to assess the factual consistency of the generated summaries. The Fact-based metrics consist of two variants: Fact-Core, which relies on the extraction of seven core fact attributes, and Fact-Full, which combines these core facts and five additional attributes.

In addition, there is also work combining the two approaches to extract concepts: Zhang et al. (2021) extracted medical relevant concepts via one of two systems: their in-house rule-based system and quickUMLS (Soldaini and Goharian, 2016) – a Python implementation of UMLS. Their rule-based system was found to be effective in capturing symptom-related findings in clinical reports, and quickUMLS is capable of extracting a wide scope of medical findings such as symptoms, diseases, medication and procedures.

Moreover, Molenaar et al. (2020) measured the quality of the dialogue summarization pipeline for healthcare reporting by establishing the number of items included in the generated and gold standards, using precision, recall and false positives (FPs) as metrics. They followed the SOEP/SOAP format – Subjective (S), Objective (O), Evaluation (E) / Assessment (A) and Plan (P) – commonly used by general practitioners in the Netherlands. It appears that they manually calculated the number of items included in each section of the SOEP format for the eight reports generated.

However, concept-based evaluation can have its own limitations, particularly with regard to false positives errors, Zhang et al. (2021) employed filtering rules to attempt to mitigate this issue.

Additionally, Chen et al. (2022) reported Regex-based Diagnostic Accuracy (RD-Acc), which measures the model's ability to diagnose the disease. Their reference reports written by annotators contain six parts, RD-Acc is calculated using the regex-based approach based on the *diagnosis* part. They calculated for what percentage of the generated reports, the content of their *diagnosis* part contains the actual disease text or key concepts.

4.1.3 Medical Correctness: Auxiliary or Intermediate Tasks

Another subcategory of automatic measures of concept correctness is those that evaluate auxiliary or intermediate tasks, there are two types: classification of electronic health record (EHR) categories and utterances classification.

To generate case notes from digital transcripts of doctor-patient conversations, Kazi and Kahanda (2019) divided the task into two subtasks: (1) predict semantic topics for segments of the transcripts (EHR categories) and then (2) generate a more formal version of the text that goes into the corresponding section of the EHR form. They used the AUROC (Area Under the ROC Curve) scale (Bewick et al., 2005) to assess their first task of predicting EHR categories, which could be any of the following: Client Details, Chief Complaint, Family History, Social History, Medical History and Others. Correct prediction of EHR categories could be useful for subsequent formal text generation.

For utterances classification, there are different types of classification such as classifying noteworthy utterance (Krishna et al., 2021), label prediction for medical conversation utterances (Song et al., 2020), and dialogue turn classification (Lacson et al., 2006).

In detail, Krishna et al. (2021) evaluated the multi-label classification of noteworthy utterances that are relevant to each summary section before clustering related utterances and generating one summary sentence per cluster. Their modular summarization technique outperforms its purely abstractive counterpart, producing much more factual and coherent sentences. Besides, Song et al. (2020) first identified two types of utterances (problem statements and treatment recommendations) and then generated summaries, they showed that for the particular dataset used, high-quality summaries can be generated by extracting these two types of utterances. Thus, in addition to reporting ROUGE scores, they also reported the precision, recall, and F scores of the predicted labels for utterances of medical conversations, compared to the standard labels. In addition, Lacson et al. (2006) also measured precision, recall, and F measure of dialogue turns classification.

4.2 Human Evaluation

The differences between intrinsic and extrinsic evaluation lie in the fact that the former aims to evaluate the properties of the system's output (Graham et al., 2018; Ji et al., 2022), while the latter aims to examine the extent to which the system accomplishes the overarching task for which it was developed. Of the 19 articles reviewed on text-based MRG, 9 included human evaluation (47%), and only 3 of them (16%) included extrinsic evaluation.

4.2.1 Intrinsic Approaches

The intrinsic human evaluation of generated reports comprises two categories as for automated metrics: text quality and medical correctness.

Text quality is important in MRG as in general NLG output. For text quality, a wide variety of properties can be considered and various linguistic parameters can be used, e.g. relevance, consistency, fluency, coherence, missing, hallucination, repetition and contraction. As an example, Srivastava et al. (2022) used four standard linguistic parameters: relevance (selection of relevant content), consistency (factual alignment between the summary and the source), fluency (linguistic quality of each sentence), and coherence (structure and organization of summary). In addition to these commonly used and well-studied criteria, the evaluation of MRG also concludes other medical correctness criteria, such as factually correct and medically relevant information (Joshi et al., 2020; Chintagunta et al., 2021), which are specific to MRG tasks. As another example, Ben Abacha et al. (2023) performed expert-based manual evaluation using NLG criteria such as Fluency and Non-redundancy, and medical criteria such as Critical Omissions, Hallucinations, Correct Facts, Incorrect Facts based on fact extraction.

Furthermore, depending on whether evaluators assess the output directly or by comparing different texts, intrinsic human evaluation can be classified into direct and relative evaluation. As for the articles involving human evaluation, they all used at least direct evaluation, i.e. the evaluators judged the generated texts directly on a defined scale. Some authors also performed relative evaluation in addition to direct evaluation: Joshi et al. (2020) and Chintagunta et al. (2021) performed a comparison task in which, given two summaries generated by different models and the associated dialogue, annotators had to choose which summary was better, they could also choose "both" and "none". Yim and Yetisgen (2021) ranked the four systems against each other, with 1 being the best, in addition to evaluating each system independently with a score from 1 to 5 for the categories relevancy, factual accuracy, writing-style, completeness, and overall.

In general, MRG outputs are evaluated at the report level, however, depending on the design of the model, some are additionally evaluated at the sentence/section/part level. For example, Krishna et al. (2021) divided SOAP notes into several subsections: *Family Medical History*, *Past Surgical History*, *Chief Complaint*, etc. Therefore, they evaluated the generated SOAP notes in two ways: 1) SOAP note sentence level and 2) SOAP note level.

4.2.2 Extrinsic Approaches

As for extrinsic human evaluation: to evaluate generated summaries, a team of mental health experts used clinical acceptability framework (Sekhon et al., 2017), which includes six parameters: affective attitude, burden, ethicality, coherence, opportunity costs, and perceived effectiveness (Srivastava et al., 2022). In addition, to perform a task-based evaluation and measure the usefulness of summaries for preserving important information in the medical setting, Lacson et al. (2006) asked physicians and nurses to create a list of key questions based on topics that commonly arise between hemodialysis patients and caregivers, and then asked five physicians to answer each of the six "yes/no" questions using each of 40 dialogues. Furthermore, in a study evaluating the correlation between human evaluation and automatic metrics in consultation note generation, Moramarco et al. (2022) asked 5 clinicians to post-edit generated notes and extract all errors.

4.2.3 Presence of Domain Experts

Most of the articles we reviewed that included a human evaluation involved domain experts, such as doctors serving patients on their telehealth platform (Chintagunta et al., 2021; Joshi et al., 2020), five licensed physicians (Lacson et al., 2006), three general practice physicians (Moramarco et al., 2021), an annotator with a medical degree (Yim and Yetisgen, 2021), etc. Sometimes, the expertise of the annotators is not specified, e.g. "trained human annotators" (Krishna et al., 2021).

We also note that of the 9 articles including human evaluation, 5 of them reported Inter-Evaluator Agreement: three of the medical dialogue summarization articles (Zhang et al., 2021; Moramarco et al., 2022; Ben Abacha et al., 2023), and two medical (report) summarization articles (Moramarco et al., 2021; Karn et al., 2022). It would be preferable to indicate Inter-Evaluator Agreement in the presence of several annotators.

5 Conclusion

Automating medical report generation can save time for experts and provide crucial information for decision-making. However, the evaluation process is necessary for validation and adoption of MRG systems in the real world. Due to the specificity of domain-specific NLG tasks like MRG, their evaluation requires more investigation and subtlety.

MRG evaluation shares similarities with general NLG evaluation, but it differs in its focus on domain knowledge and task-specific concerns, especially in the assessment of conceptual accuracy of medical concepts. However, the question of which medical facts to pay attention to (correlation, consensus, etc) is an open question, requiring close collaboration with experts in the field.

In addition, the evaluation of MRG systems requires both intrinsic and extrinsic evaluation. Intrinsic evaluation focuses on properties of the system's output, while extrinsic evaluation involves professional experts in the design acceptability process, developing a list of key questions, and post-editing. Future research should prioritize extrinsic evaluation, particularly in scenarios where references are unavailable, and developing efficient, medical task-specific automated measures.

Limitations

In this work, we studied only the evaluation of textual medical report generation from both automatic and human evaluation perspectives, but we did not study the evaluation of data-to-text medical report generation, which has its own specificities.

Acknowledgements

This research was carried out as part of the THERADIA project (funded by BPI) and was partially supported by MIAI@Grenoble-Alpes (ANR-19-P3IA-0003) and the COST action Multi3Generation (CA18231). We would also like to thank the anonymous reviewers for their insightful comments.

References

Belén A. Baez Miranda, Sybille Caffiau, Catherine Garbay, and François Portet. 2015. Towards a computational generation of récit: evaluating the perception of the récit plan. In *1st international Workshop on Data to Text Generation (D2T)*.

Sabita Acharya, Barbara Di Eugenio, Andrew D. Boyd, Karen Dunn Lopez, Richard Cameron, and Gail M Keenan. 2016. Generating summaries of hospitalizations: A new metric to assess the complexity of medical terms and their definitions. In *Proceedings of the 9th International Natural Language Generation conference*, pages 26–30, Edinburgh, UK. Association for Computational Linguistics.

Griffin Adams, Jason Zucker, and Noémie Elhadad. 2023. A meta-evaluation of faithfulness metrics for long-form hospital-course summarization.

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018. Evaluation methodologies in automatic question generation 2013-2018. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 307–317, Tilburg University, The Netherlands. Association for Computational Linguistics.

Anya Belz, Simon Mille, and David M. Howcroft. 2020. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 183–194, Dublin, Ireland. Association for Computational Linguistics.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, Dubrovnik, Croatia. Association for Computational Linguistics.

Viv Bewick, Liz Cheek, and Jonathan Ball. 2005. Statistics review 13: Receiver operating characteristic curves. *Critical care (London, England)*, 8:508–12.

Wei Chen, Zhiwei Li, Hongyi Fang, Qianyuan Yao, Cheng Zhong, Jianye Hao, Qi Zhang, Xuanjing Huang, Jiajie Peng, and Zhongyu Wei. 2022. A benchmark for automatic medical consultation system: Frameworks, tasks and datasets.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5904–5914, Online. Association for Computational Linguistics.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449, Online. Association for Computational Linguistics.

Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2021. Medically aware GPT-3 as a data generator for medical dialogue summarization. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 66–76, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Barbara Di Eugenio, Andrew Boyd, Camillo Lugaresi, Abhinaya Balasubramanian, Gail Keenan, Mike Burton, Tamara Goncalves Rezende Macieira, Jianrong Li, Yves Lussier, and Yves Lussier. 2014. PatientNarr: Towards generating patient-centric summaries of hospital stays. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, pages 6–10, Philadelphia, Pennsylvania, U.S.A. Association for Computational Linguistics.

Seppo Enarvi, Marilisa Amoia, Miguel Del-Agua Teba, Brian Delaney, Frank Diehl, Stefan Hahn, Kristina Harris, Liam McGrath, Yue Pan, Joel Pinto, Luca Rubini, Miguel Ruiz, Gagandeep Singh, Fabian Stemmer, Weiyi Sun, Paul Vozila, Thomas Lin, and Ranjani Ramamurthy. 2020. Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, pages 22–30, Online. Association for Computational Linguistics.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2022. A survey on dialogue summarization: Recent advances and new frontiers. In *IJCAI*.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt.

Dimitra Gkatzia and Saad Mahamood. 2015. A snapshot of NLG evaluation practices 2005 - 2014. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, pages 57–60, Brighton, UK. Association for Computational Linguistics.

Yvette Graham, George Awad, and Alan Smeaton. 2018. Evaluation of automatic video captioning using direct assessment. *PLOS ONE*, 13(9):1–20.

Henk Harkema, John N. Dowling, Tyler Thornblade, and Wendy W. Chapman. 2009. Context: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *J. of Biomedical Informatics*, 42(5):839–851.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad

Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Tianbo Ji, Yvette Graham, Gareth Jones, Chenyang Lyu, and Qun Liu. 2022. Achieving reliable human assessment of open-domain dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6416–6437, Dublin, Ireland. Association for Computational Linguistics.

Pamela Jordan, Nancy Green, Chistopher Thomas, and Susan Holm. 2014. TBI-doc: Generating patient & clinician reports from brain imaging data. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, pages 143–146, Philadelphia, Pennsylvania, U.S.A. Association for Computational Linguistics.

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3755–3763, Online. Association for Computational Linguistics.

Sanjeev Kumar Karn, Ning Liu, Hinrich Schuetze, and Oladimeji Farri. 2022. Differentiable multi-agent actor-critic for multi-step radiology report summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1542–1553, Dublin, Ireland. Association for Computational Linguistics.

Nazmul Kazi and Indika Kahanda. 2019. Automatically generating psychiatric case notes from digital transcripts of doctor-patient conversations. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 140–148, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Nazmul Kazi, Matt Kuntz, Upulee Kanewala, and Indika Kahanda. 2020. Dataset for automated medical transcription.

Ali Can Kocabiyikoglu, Jean-Marc Babouchkine, François Portet, and Raheel Qader. 2021. Neural medication extraction: A comparison of recent models in supervised and semi-supervised learning settings. In *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*, pages 148–152.

Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary C. Lipton. 2021. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4958–4972, Online. Association for Computational Linguistics.

Karen Kukich. 1983. Design of a knowledge-based report generator. In *21st Annual Meeting of the Association for Computational Linguistics*, pages 145–150, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Ronilda C. Lacson, Regina Barzilay, and William J. Long. 2006. Automatic analysis of medical dialogue in the home hemodialysis domain: Structure induction and summarization. *Journal of Biomedical Informatics*, 39(5):541–555. Dialog Systems for Health Communications.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Ping Zhang, and Xu Sun. 2021. Contrastive attention for automatic chest X-ray report generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 269–280, Online. Association for Computational Linguistics.

Zhengyuan Liu, Angela Ng, Sheldon Lee, Ai Ti Aw, and Nancy F. Chen. 2019. Topic-aware pointer-generator networks for summarizing spoken conversations. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 814–821.

Justin Lovelace and Bobak Mortazavi. 2020. Learning to generate clinically coherent chest X-ray reports. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1235–1243, Online. Association for Computational Linguistics.

Chong Ma, Zihao Wu, Jiaqi Wang, Shaochen Xu, Yaonai Wei, Zhengliang Liu, Xi Jiang, Lei Guo, Xiaoyan Cai, Shu Zhang, Tuo Zhang, Dajiang Zhu, Dinggang Shen, Tianming Liu, and Xiang Li. 2023. Impressiongpt: An iterative optimizing framework for radiology report summarization with chatgpt.

Lientje Maas, Adriaan Kisjes, Iman Hashemi, Floris Heijmans, Fabiano Dalpiaz, Sandra Van Dulmen, and Sjaak Brinkkemper. 2021. Automated medical reporting: From multimodal inputs to medical reports through knowledge graphs. In *HEALTHINF*.

Ganeshan Malhotra, Abdul Waheed, Aseem Srivastava, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 735–745, New York, NY, USA. Association for Computing Machinery.

Pablo Messina, Pablo Pino, Denis Parra, Alvaro Soto, Cecilia Besa, Sergio Uribe, Marcelo Andía, Cristian Tejos, Claudia Prieto, and Daniel Capurro. 2021. A survey on deep learning and explainability for automatic report generation from medical images. *ACM Comput. Surv.* Just Accepted.

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. 2021. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5288–5304, Online. Association for Computational Linguistics.

Sabine Molenaar, Lientje Maas, Verónica Burriel, Fabiano Dalpiaz, and Sjaak Brinkkemper. 2020. Medical dialogue summarization for automated reporting in healthcare. *Advanced Information Systems Engineering Workshops*, 382:76 – 88.

Francesco Moramarco, Damir Juric, Aleksandar Savkov, and Ehud Reiter. 2021. Towards objectively evaluating the quality of generated medical summaries. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 56–61, Online. Association for Computational Linguistics.

Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Damir Juric, Jack Flann, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. Human evaluation and correlation with automatic metrics in consultation note generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5739–5754, Dublin, Ireland. Association for Computational Linguistics.

Varun Nair, Namit Katariya, Xavier Amatriain, Ilya Valmianski, and Anitha Kannan. 2021. Adding more data does not always help: A study in medical conversation summarization with PEGASUS. *CoRR*, abs/2111.07564.

Farhad Nooralahzadeh, Nicolas Perez Gonzalez, Thomas Frauenfelder, Koji Fujimoto, and Michael Krauthammer. 2021. Progressive transformer-based generation of radiology reports. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2824–2832, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. PriMock57: A dataset of primary care mock consultations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 588–598, Dublin, Ireland. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

François Portet, Ehud Reiter, Albert Gatt, Jim Hunter, Somayajulu Sripada, Yvonne Freer, and Cindy Sykes. 2009. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, 173(7):789–816.

Raheel Qader, Khoder Jneid, François Portet, and Cyril Labbé. 2018. Generation of Company descriptions using concept-to-text and text-to-text deep models: dataset collection and systems evaluation. In *11th International Conference on Natural Language Generation*, Tilburg, Netherlands.

Han Qin and Yan Song. 2022. Reinforced cross-modal alignment for radiology report generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 448–458, Dublin, Ireland. Association for Computational Linguistics.

Clement Rebuffel, Thomas Scialom, Laure Soulier, Benjamin Piwowarski, Sylvain Lamprier, Jacopo Staiano, Geoffrey Scoutheeten, and Patrick Gallinari. 2021. Data-QuestEval: A referenceless metric for data-to-text semantic evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8029–8036, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Ehud Reiter, Albert Gatt, François Portet, and Marian van der Meulen. 2008. The importance of narrative and other lessons from an evaluation of an NLG system that summarises clinical data. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 147–156, Salt Fork, Ohio, USA. Association for Computational Linguistics.

Ehud Reiter, Roma Robertson, and Liesl M. Osman. 2003. Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, 144(1):41–58.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Comput. Surv.*, 55(2).

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mandeep Sekhon, Martin Cartwright, and Jill J. Francis. 2017. Acceptability of healthcare interventions: an overview of reviews and development of a theoretical framework. *BMC Health Serv Res*, 17(88):114–133.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Luca Soldaini and Nazli Goharian. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction. In *Medical Information Retrieval (MedIR) Workshop, SIGIR*.

Yan Song, Yuanhe Tian, Nan Wang, and Fei Xia. 2020. Summarizing medical conversations via identifying important utterances. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 717–729, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Aseem Srivastava, Tharun Suresh, Sarah P. Lord, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Counseling summarization using mental health knowledge guided utterance filtering. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3920–3930, New York, NY, USA. Association for Computing Machinery.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

J.P. Woodard and J.T. Nelson. 1982. An information theoretic measure of speech recognition performance. In *Workshop on standardisation for speech I/O technology, Naval Air Development Center, Warminster, PA*.

An Yan, Zexue He, Xing Lu, Jiang Du, Eric Chang, Amilcare Gentili, Julian McAuley, and Chun-Nan Hsu. 2021. Weakly supervised contrastive learning for chest X-ray report generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4009–4015, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wen-wai Yim and Meliha Yetisgen. 2021. Towards automating medical scribing : Clinic visit Dialogue2Note sentence alignment and snippet summarization. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 10–20, Online. Association for Computational Linguistics.

Longxiang Zhang, Renato Negrinho, Arindam Ghosh, Vasudevan Jagannathan, Hamid Reza Hassanzadeh, Thomas Schaaf, and Matthew R. Gormley. 2021. Leveraging pretrained models for automatic summarization of doctor-patient conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3693–3712, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D. Manning, and Curtis Langlotz. 2020. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5108–5120, Online. Association for Computational Linguistics.

A Appendix

A.1 Review Structure

We reviewed the MRG articles following the criteria described in van der Lee et al. (2019), the attributes examined are presented in Table 3.

A.2 Resources for Text-based MRG

We identified six datasets used in text-based MRG and summarized them in Table 4. Song et al. (2020) collected medical conversations from online platforms. MTSamples (Moramarco et al., 2021) were also collected from a community platform website. Srivastava et al. (2022) extended data from the publicly available counseling conversation dataset - HOPE (Malhotra et al., 2022), which takes place between therapist and patient. We observed that mock conversations were mentioned several times: Kazi et al. (2020) used transcripts from two different sources to generate audio recordings of simulated doctor-patient conversations, Papadopoulos Korfiatis et al. (2022) simulated primary care consultations. Recently, Ben Abacha et al. (2023) introduced a new collection of simulated doctor-patient conversations from publicly available clinical notes and corresponding clinical summaries.

	Criteria
Task	Sub-task(s) of MRG
Uses Automated Metrics	YES/NO
What kind of Automated Metrics	NLP metrics and/or other specific metrics
Uses Intrinsic (Human) Evaluation	YES/NO
What kind of Intrinsic (Human) Evaluation	Fluency, naturalness, quality, meaning preservation, etc.
Scale	Likert (5-point), preference, rank-based magnitude estimation, etc.
Number of participants	Number of annotators for the Human Evaluation task (including details on annotators)
Uses Extrinsic (Human) Evaluation	YES/NO
What kind of Extrinsic (Human) Evaluation	Task success, etc.
Number of examples	Number of samples evaluated for each system
Examples per participant	Number of examples that each participant is asked to evaluate
Details about design (order, groups)	Methods for selecting human evaluation samples from the original test set and how they are distributed to each annotator
Inter-Annotator Agreement (IAA)	Presence of inter-annotator agreement statistics

Table 3: Attributes studied and their descriptions in our structured review, adapted from van der Lee et al. (2019). *MRG* means *Medical Report Generation*.

Dataset	Language	Description	Domain	Size
Medical Conversation (ChiCCo) (Song et al., 2020)	Chinese	The summary has two parts: SUM1 describes the patient's medical problem; SUM2 summarizes the doctor's diagnosis or treatment recommendations.	Medical (online platforms conversation)	44,983 cases, 855,403 utterances
Automated Medical Transcription (Kazi et al., 2020)	English	Used transcripts from two different sources to generate audio recordings of enacted doctor-patient conversations	Medical, psychiatric consultations	71 recordings with transcripts and case notes
MTSamples (Moramarco et al., 2021)	English	From a community platform website, 40 medical specialties. Reports are free text with headings –> to generate the description field of a report	Medical summaries	5,000 sample medical transcription reports
MEMO (Srivastava et al., 2022)	English	Extend data collected from the publicly available counseling conversation (between therapist and patient) dataset - HOPE (Malhotra et al., 2022) to annotate psychotherapy elements and counseling summary	Mental health, Counseling	12.9K utterances, 212 conversations
PriMock57 (Papadopoulos Korfiatis et al., 2022)	English	Mocked primary care consultations, including audio recordings, their manual utterance level transcriptions, and the associated consultation notes	Medical, Primary Care Mock Consultations	57
MTS-Dialog (Ben Abacha et al., 2023)	English	A collection of 1.7k doctor-patient conversations and corresponding clinical notes/summaries.	Doctor-Patient Encounters	1.7k

Table 4: Text-based Medical Report Generation related datasets.

UMASS_BioNLP at MEDIQA-Chat 2023: Can LLMs generate high-quality synthetic note-oriented doctor-patient conversations?

Junda Wang * **Zonghai Yao** * **Avijit Mitra** **Samuel Osebe** **Zhichao Yang** **Hong Yu**

CICS, University of Massachusetts, Amherst, MA, USA

jundawang@umass.edu **zonghaiyao@umass.edu**

Abstract

This paper presents UMASS_BioNLP team participation in the MEDIQA-Chat 2023 shared task for Task-A and Task-C. We focus especially on Task-C and propose a novel LLMs cooperation system named a doctor-patient loop to generate high-quality conversation data sets. The experiment results demonstrate that our approaches yield reasonable performance as evaluated by automatic metrics such as ROUGE, medical concept recall, BLEU, and Self-BLEU. Furthermore, we conducted a comparative analysis between our proposed method and ChatGPT and GPT-4. This analysis also investigates the potential of utilizing cooperation LLMs to generate high-quality datasets. [1]

1 Introduction

The issue of the growing burden of clinical documentation has become a critical concern in healthcare, resulting in increased job dissatisfaction and burnout rates among clinicians and adversely affecting patient experiences. Nevertheless, timely and accurate documentation of patient encounters is crucial for safe, effective care and communication between specialists. Consequently, there is a growing interest in automating assisting doctors in diagnosis based on Large Language Models (LLMs) due to its remarkable advancement in the field of artificial intelligence (AI), being highly sophisticated systems that have been extensively trained on massive amounts of textual data. (Brown et al., 2020; Sanh et al., 2021; Chowdhery et al., 2022; Longpre et al., 2023; OpenAI, 2023)

The swift progress of AI and its extensive influence on various fields have garnered considerable attention from the research community. One notable area is the creation of instruction-following LLMs (Touvron et al., 2023; Taori et al., 2023; Chiang et al., 2023; Zhu et al., 2023), which demonstrate extraordinary ability in understanding instructions and producing human-like responses. These auto-regressive LLMs undergo a two-step process: they are initially pre-trained on web-scale natural languages through next-token prediction and subsequently fine-tuned to comply with extensive human instructions (Dale, 2021). This method leads to impressive performances across a broad range of natural language processing (NLP) tasks and generalizes to unseen tasks, underscoring their potential as a comprehensive solution for diverse challenges, including natural language understanding, text generation, and conversational AI (Floridi and Chiriatti, 2020). Many auto-regressive LLMs, such as ChatGPT, have further training with RLHF to align with human preference and finally allow these models to generate content that most people prefer. In the biomedical domain, many researchers have attempted to apply auto-regressive models to medical tasks such as patient triage (Levine et al., 2023), automatic disease coding (Yang et al., 2022), and doctor-chatbot (Yunxiang et al., 2023; Xu et al., 2023).

However high-quality dialogue datasets featuring doctor-patient interactions are a task that is inherently complex. One major difficulty in constructing such a dataset is the sensitive nature of the content, as healthcare conversations often involve private and confidential patient information (Kelly et al., 2019; Rindfleisch, 1997; Annas, 2003). Ensuring privacy protection and adhering to strict data regulations, such as HIPAA, becomes crucial in the development process. Consequently, the compilation of authentic doctor-patient dialogues requires careful consideration of privacy and data protection measures to prevent potential ethical and legal concerns. Recent work (Ben Abacha et al., 2023; Yim et al., 2023) attempts to synthesize data by letting humans play the roles of doctor and patient in a conversation, but the huge cost makes

* indicates equal contribution

[1] Our codes are released at https://github.com/believewhat/Dr.NoteAid

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 460–471
July 14, 2023 ©2023 Association for Computational Linguistics

the research community seek the help of LLMs or Chatbot models to simulate such role-playing game (doctor and patient) for data augmentation. However, recent Chatbot models (Yunxiang et al., 2023; Zeng et al., 2020b) are only based on single-turn or multi-turn question-and-answer repositories rather than real conversations between patients and doctors. Question-and-answer datasets lack logical coherence, whereas real conversations can help the model understand the proper order of questions (Drew et al., 2001), thereby guiding patients to describe their symptoms accordingly and ultimately aiding in disease diagnosis. So we cannot simply use the recent medical chatbot models to generate high-quality note-oriented doctor-patient conversations. On the other hand, the success of these LLM models remains heavily reliant on human input to guide conversations in the right direction. This dependence necessitates users to provide relevant and precise prompts based on their intentions and the chat agent's feedback, which can be challenging, time-consuming, and occasionally unfeasible. In healthcare contexts, individuals without medical expertise may struggle to generate appropriate prompts for directing communicative agents to deliver accurate medical advice or diagnoses (Tang et al., 2023; Liao et al., 2023).

All of this raises a crucial question: How can we try role-playing games to guide conversations toward clinical note completion in healthcare settings without a large number of human annotations? To address these issues, we propose a cutting-edge cooperative agent framework, Doctor-Patient Loop. This approach involves two main ChatGPT agents taking on the roles of doctor and patient in dialogue, with additional ChatGPT agents responsible for fact-checking, ensuring conversations remain focused on the provided notes, determining when the dialogue should be terminated, and refining the conversation to enhance its coherence and fluency. The collaboration among multiple ChatGPT agents leads to the creation of more realistic doctor-patient dialogue datasets, which in turn can be utilized for training models that better mimic genuine healthcare communication scenarios.

In this paper, we conducted a series of experiments with the help of the data set of the MEDIQA-Chat competition shared task. Specially, we present our entry for Task-A and Task-C. We explored a new approach to solve For Task-A. We trained BioMedLM [2] on the dataset of Task-A and designed prompts for different section headers. For Task-c, we explored the potential for creating scalable methods that promote autonomous cooperation among communicative agents in medical settings. We construct a doctor-patient loop to generate high-quality clinical dialogue. Our paper's contributions can be summarized as follows:

1) We propose innovative approaches to foster autonomous cooperation among communicative agents in medical settings, highlighting their cognitive processes and collaborative capabilities.

2) We concentrate on the generation and utilization of continuous doctor-patient dialogue datasets, which serve as valuable resources for developing AI systems that can better understand and address context-sensitive inquiries in healthcare communication.

3) We finetuned BioMedLM on a collection of data sources to obtain the FLAN-BioMedLM model and then finetuned this model on the Task-A dataset on the classification task. It achieved good performance in the task of section header classification and was used to assist ChatGPT in generating clinical notes.

2 Related Work

The MEDIQA-Chat 2023 tasks [3] (Abacha et al., 2023a; Ben Abacha et al., 2023; Ben Abacha et al., 2023; Yim et al., 2023) focused on both Dialogue2Note Summarization and Note2Dialogue Generation tasks. The researchers constructed a novel dataset comprising 1,700 doctor-patient conversations (16k turns and 18k sentences) and their summarized clinical notes (6k sentences). They also proposed an investigation of standard evaluation metrics, domain-specific metrics, and expert judgments for the task, including the calculation of the correlation between the automatic and manual scores for the evaluation of the generated clinical notes. In this paper, we use a cooperative agent framework to generate the conversation data sets.

For the **section header and content classification** (Task-A), the SOAP (Subjective, Objective, Assessment, and Plan) structure is commonly used

[2]https://github.com/stanford-crfm/BioMedLM
[3]https://sites.google.com/view/mediqa2023

by providers (Podder et al., 2021). The Subjective section is a detailed report of the patient's current conditions, such as source, onset, and duration of symptoms, mainly based on the patient's self-report. This section usually includes a history of present illness and symptoms, current medications, and allergies. The Objective section documents the results of physical exam findings, laboratory data, vital signs, and descriptions of imaging results. The Assessment section typically contains medical diagnoses and reasons that lead to medical diagnoses. The assessment is typically based on the content from the subjective and objective sections. The Plan section addresses treatment plans based on the assessment. Previous work focused on identifying these four general SOAP sections (Kwon et al., 2022a). In this paper, we focused on predicting the specific subsections.

For the **Dialogue2Note Summarization** task (Task-A&B), there are many solutions already in the industry (Krishna et al., 2021; Song et al., 2020; Yim and Yetisgen-Yildiz, 2021; Krishna et al., 2021; Schloss and Konam, 2020). This process generally follows a similar pipeline. Clinical conversations are initially recorded and then transcribed. Subsequently, the utterances are classified into various medical sections, and clusters of utterances containing medically relevant information for each section are predicted from the transcript. Finally, a section-conditioned summarization model is employed to generate concise summaries for each cluster of utterances associated with their respective sections. However, the size of their private training data is larger than the training data of this competition. Thus, we applied LLM for this competition. Instruction finetuning helps LLM generalize into unseen tasks where training data is limited (Longpre et al., 2023). For example, an instruction-finetuned 11B-param model outperforms the 60B-param model without instruction-finetuning in the BIG-Bench dataset (Chung et al., 2022). Thus, we instructed finetuned BioMedLM model for Task-A Dialogue2Note Summarization task.

For the **Note2Dialogue Generation** task (Task-C), MEDIQA-Chat 2023 treats it as a data augmentation task. Recent investigations into utilizing LLMs for data augmentation have produced notable results. Li et al. (2023) explored the possibility of using LLMs to generate training data for tasks such as code summarization, code translation, and code generation. In a similar vein, Dai et al. (2023) suggested employing LLMs to tackle low-resource scenario model training by augmenting data to enhance performance. Moreover, Gilardi et al. (2023) and Ding et al. (2022) studied the effectiveness and accuracy of LLMs for data annotation, respectively, achieving promising outcomes even when compared with Crowd-Workers. Bonifacio et al. (2022) utilized LLMs to create positive sample pairs for training downstream models. At the same time, Zhou et al. (2022) focused on generating appropriate prompts with LLMs to improve performance further. Lastly, Dai et al. (2022) mainly targeted few-shot retrieval tasks, combining LLMs with a limited number of samples to produce additional training data for retrieval models. In the biomedical field, Tang et al. (2023) investigated the potential of LLMs in clinical text mining and introduced a novel training paradigm to address sub-optimal performance and privacy concerns. Liao et al. (2023) examined responsible and ethical Artificial Intelligence Generated Content (AIGC) in medicine, analyzing differences between human-authored and LLM-generated medical texts and developing machine learning workflows for efficient detection and differentiation. In this paper, we explored our cooperative agent framework's performance in Task-C.

3 Methods

3.1 MEDIQA-Chat Tasks

The competition proposed two new shared tasks, namely Dialogue2Note and Note2Dialogue, which aim to facilitate clinical note creation through the summarization of medical conversations and the generation of synthetic doctor-patient dialogues for data augmentation purposes, respectively.

1) Dialogue2Note Summarization: This task entails generating a clinical note that succinctly summarizes a conversation between a doctor and a patient. The resulting clinical note may contain one or multiple sections, such as Assessment, Past Medical History, and Past Surgical History. Task-A focuses on generating specific note sections from the doctor-patient conversation: first, predicting the section heading and then generating the content of the specific section.

2) Note2Dialogue Generation: This task involves creating a synthetic doctor-patient conversation based on the information provided

Domain	Dataset	Reference
Medical	MeQSum	(Ben Abacha and Demner-Fushman, 2019)
	Primock57	(Papadopoulos Korfiatis et al., 2022)
	EmrQA	(Pampari et al., 2018)
	DiSCQ	(Lehman et al., 2022)
	MEDIQA-AnS	(Savery et al., 2020)
	Pubmed-ccdv	(Cohan et al., 2018)
	Medal	(Wen et al., 2020)
	Diagnoise-me	(Zeng et al., 2020a)
	Medmcqa	(Pal et al., 2022)
	Ebm_pico	(Nye et al., 2018)
	Pubhealth	(Kotonya and Toni, 2020)
	Pmc_patients	(Zhao et al., 2022)
General	Multiwoz	(Zang et al., 2020)
	Taskmasters	(Byrne et al., 2019)
	Dart	(Nan et al., 2021)
	WebNLG	(Shimorina and Gardent, 2018)

Table 1: Datasets used to train FLAN-BioMedLM

in a full clinical note. Participants are required to generate a dialogue that effectively captures the context and content of the original clinical note, thereby contributing to relevant data creation and augmentation.

These shared tasks, as presented at the ACL conference, are expected to significantly contribute to the development of cutting-edge methodologies and techniques in the realm of automatic clinical note generation, ultimately improving the overall efficiency and quality of healthcare documentation.

3.2 Conversation to Clinical Note

Similar to the general training pipeline of FLAN, we first instruction-finetuned BioMedLM (2.7 billion parameters model pre-trained on PubMed articles) on a collection of data sources to obtain the FLAN-BioMedLM model and then finetuned this model on Task-A dataset. Our approach differed from FLAN in the specific model and the data collection tailored for Task-A. Since this is a medical-domain conversation task, we selected related data sources including 12 medical-domain generation datasets and 4 general-domain conversation/controlled generation datasets as shown in table 1. Medical-domain generation tasks include long-form question answering between doctors and patients, squad-like question answering from medical notes, medical notes summarization, research article summarization, and abbreviation disambiguation. This task collection contains 110 prompt templates and 400 million tokens.

We then finetuned FLAN-BioMedLM on Task-A. Specifically, we built a pipeline to classify section heading first, and then used this heading to generate section content. When the generated heading string did not match to the ground truth class name, we used fuzzy string matching to find its nearest valid header. We finetuned FLAN-BioMedLM on these two subtasks separately. Our prompts are shown in table 2.

We also explored this task using ChatGPT. We found that ChatGPT has a lower accuracy in classifying section headings, and its performance in generating notes is highly dependent on the given examples. Therefore, we first use FLAN-BioMedLM for headings classification and then provide ChatGPT with corresponding examples based on the section headings.

3.3 Clinical Note to Conversation

3.3.1 Segmentation

In MEDIQA-Chat Task-C, the training set consists of comprehensive and extensive clinical notes. There are 20 validation samples and 40 test samples. We try to apply GPT3.5-turbo to generate the dialogue. However, due to the maximum token limitation imposed by the GPT3.5-turbo API, it is infeasible to input the entire dialogue when providing a prompt. Consequently, we dissect the clinical note into several section headings as shown in the heading subtask of Table 2. For each section heading, we leverage the dataset from Task-A to construct a prompt that assists the model in generating a dialogue segment. Ultimately, the conversation fragments corresponding to different section headings are concatenated to form a complete dialogue.

3.3.2 Doctor-Patient Loop

Language models often lack sufficient medical knowledge to help them accomplish the target tasks (Sung et al., 2021; Yao et al., 2022a,b). So we employed the MedSpaCy library to extract relevant CUI codes from clinical notes, aiming to guide subsequent conversations around these key terms. Such a checklist can help our pipeline improve factuality (Tang et al., 2022; Abacha et al., 2023b; Chang et al., 2023), and can be changed very flexibly for other purposes, like information retrieval (Khattab et al., 2022), entity linking (Yao et al., 2020), medical jargon extraction (Kwon et al., 2022b), causulity (Yuan et al., 2023), and rules or knowledge injection (Fei et al., 2021; Yao and Yu, 2021; Yao et al., 2023). Upon extraction, we initiated a doctor-patient loop involving multiple rounds of dialogue to generate comprehensive conversations. In each round, one ChatGPT instance played the role of a doctor while the other acted as a patient. The doctor, incorporating the case details and identified keywords, would select and focus on up to four key terms to pose questions to the patient. The ChatGPT representing the patient would then respond to the inquiries based on the clinical

Heading subtask	Dialogue: dialogue. Given the dialogue above, select a section of the medical note from the options below. Options: history of present illness; review of systems; past medical history; medications; chief complaint; past surgical history; disposition; diagnosis; emergency department course; plan; labs; assessment; allergy; gynecologic history; exam; other history; procedures; imaging; immunizations; family history social history.
Content subtask	Dialogue: dialogue. Generate section heading of the medical note from dialogue.

Table 2: FLAN-BioMedLM prompt template for Task-A, where colored words will be replaced with actual content.

Model	R-1	R-2	R-L	R-L-Sum	bertscore_f1	bleurt
FLAN-BioMedLM	0.3283	0.1351	0.2743	0.2743	0.6699	0.4757
ChatGPT	0.3828	0.1828	0.3158	0.3166	0.7015	0.5405

Table 3: Synthetic data quality evaluation on Task-A.

notes. In each dialogue round, the conversation history from previous rounds was incorporated as a prompt input to ensure context continuity and coherence throughout the interaction, finally completing a loop in the process. Furthermore, we also construct a factuality-checking module to ensure the comprehensiveness of our conversation. We employed a ChatGPT-based approach to monitoring whether the dialogue encompasses all essential information.

3.4 Evaluation

All methods are evaluated with Rouge-1, Rouge-2, Rouge-L, Rouge-Lsum (Lin, 2004) on both Task-A and Task-C. For Task-A, we also apply BERTScore-F1 (Zhang et al., 2019) and BLEURT (Sellam et al., 2020) to test our result. For Task-C, we also use BLEU score (Post, 2018). For Task-C, to measure the generated text including all the important information such as symptoms or medication from the clinical domain, we used another metric named Concept-Recall, which evaluates the overlap of clinical keywords present in the two texts. We first extracted all Unified Medical Language System (UMLS) (Bodenreider, 2004) entities from text using MedSpaCy (Eyre et al., 2021). We further refined this list of entities by selecting only those that were clinically important. Specifically, we included entities whose semantic groups are diseases, drugs, devices, and procedures as defined in (Bodenreider and McCray, 2003), and exclude other semantic types such as fish, bird, and other conceptual entities. Finally, we calculated the overlap of entities from generated text and reference text by recall scores. We also evaluated the text diversity in Task-C. Zhu et al. (2018) proposed a benchmarking platform for text generation models that is fully open-sourced. We followed their work and evaluated the diversity of the generated conversation based on their proposed Self-BLEU score.

4 Experiment

In this section, we discuss our proposed methods' performance on MEDIQA-Chat 2023 Task-A and Task-C. All the detailed experiment settings can be found on our GitHub.

4.1 Task-A

We compared FLAN-BioMedLM and ChatGPT in the two subtasks. In the heading classification task, FLAN-BioMedLM achieved an accuracy of 0.705, and ChatGPT scored an accuracy of 0.355. However, ChatGPT outperformed FLAN-BioMedLM in the content generation task, as shown in Table 3.

4.2 Task-C

As of the end of the competition, the results of our method in the competition ROUGE family are shown in Table 4. After the end of the competition, we further did follow-up prompt engineering and saw a significant improvement in the results. In this and the next section, our discussions are all based on new results. In order to be fair, we can't compare the new results with other teams in the competition, so our baseline is mainly ChatGPT and GPT4.

	R-1	R-2	R-L	R-L-Sum	C-R	BLEU	SBLEU ↓	Len
	MEDIQA-CHAT-2023-RESULTS Task-C							
1. Cadence	54.36	23.81	20.64	47.45	-	-	-	-
2. UMass_BioNLP	42.36	11.96	15.96	40.46	-	-	-	-
3. NUSIDS	40.63	14.18	17.24	39.45	-	-	-	-
	Additional Experiments and Results (done after competition)							
ChatGPT-short	48.31	17.43	19.33	50.74	35.42	4.00	0.018	46.5
GPT-4-short	53.16	**19.49**	**23.10**	50.39	44.95	6.13	0.016	42.0
Ours-short	**54.18**	17.43	19.33	**50.74**	**47.19**	**6.62**	**0.013**	45.1
ChatGPT-long	48.56	16.74	22.41	46.36	35.75	4.93	0.017	62.8
GPT-4-long	53.29	20.20	**24.06**	50.81	45.69	5.92	0.019	58.1
Ours-long	**56.48**	19.74	20.03	**53.41**	**51.23**	**6.12**	**0.017**	62.5

Table 4: Synthetic data quality evaluation on MEDIQA-Chat using auto-metrics.

In our study, we found that ChatGPT and GPT-4 are highly sensitive to the choice of the prompt. To achieve optimal performance, we experimented with various prompts and categorized them into two groups: one for generating short-length conversations with an average length of around 40 utterances and the other for generating long conversations exceeding 50 utterances. We also observed that the length of the conversation has a significant impact on the Rouge score and concept score, as shown in the table 4.

Due to the API's maximum token limit, ChatGPT and our method (based on ChatGPT) could not generate long conversations. We found that the length of the conversation has a significant impact on the score, and scores tend to improve when the conversation length approaches that of human conversations. Therefore, we optimized the combined prompt to only concatenate the next conversation segment with the one generated from the previous topic. This allowed us to generate longer conversations within the maximum token limit. As a result, our Rouge total scores have further improved.

In addition, we found that ChatGPT and GPT-4 are suitable for generating conversations of moderate length. When we forced them to generate very long conversations, GPT4 will generate highly repetitive sentences and diverge significantly from real conversations. ChatGPT will divide long utterances into several short utterances. Hence, both ChatGPT and GPT-4 struggle to cover all the essential information even if we force them to generate longer conversations, and their concept recall scores were lower than our model's. Even in their longer versions of conversations, the amount of information covered was less than that of our shorter version because in the experiment result our shorter version model's concept is 47.19 indicating that our model can include most information and the Self-BLEU score is 0.013 which demonstrate the diversity of our model. For the longer version, our model sacrificed a small amount of diversity but gained a significant improvement in concept recall (51.23) and Rouge score. Therefore, the experiment result can demonstrate that segmentation can guide ChatGPT to cover all the essential medical information. In the segmentation module, we provide separate prompts for each different section header to guide the model's attention to the corresponding important information. Furthermore, the doctor-patient loop can make the generated conversations more logical, and the maximum turn setting ensures that the model covers all the key phrases.

4.3 Case Study

In this section, we provide examples of conversations in Table 7 generated by our model and some prompts (Table 5) to demonstrate that our approach can produce more human-like conversations. Our system mainly consists of the following prompts:

Doctor Prompt is utilized to instruct the model to assume the role of a physician, asking logically coherent questions based on the patient's clinical note and previous dialogue for the purpose of generating dialogue datasets.

Patient Prompt is designed to guide the model to play the role of a patient, answering the doctor's questions based on their own medical history. We set the patient's level of education to be low to ensure that ChatGPT's language style is more similar to that of an actual patient in daily conversation.

Doctor Prompt
Clinical Note: Note
Please role-play as a doctor and further ask a question based on the above dialogue to follow up the history conversation. The treatment plan, medication, and dosage you give to the patient must also be consistent with the clinical note. Your question should be around these keywords, and you cannot modify these keywords or use synonyms.
Key Words: $key_1, key_2, ...$
Patient Prompt
Clinical Note: Note
Please act as a patient and answer my question or follow up on the conversation. Your answer must be consistent with the clinical note and cannot include information that is not in the clinical note. Your responses should be more colloquial.
Polish Prompt
Please rewrite all the conversations based on the notes to become fluence and more colloquial, like a normal conversation between the doctor and patient based on the clinical notes. Now you should rewrite the following conversations, and your conversation should include all the information and all the keywords. The keywords must be used directly instead of using synonyms when using them in the conversation
Key Words: $key_1, key_2, ...$
The conversation:" Conversation
Clinical Note: Note
The conversation between the doctor and the patient should involve multiple rounds, with each question and answer being relatively short. You should try to ensure that the dialogue is smooth.
Hallucination Prompt
Check whether the information of the conversation is consistent with the clinical note. If there is some information that you cannot find on the clinical note, please eliminate it. You also should delete the duplicate part. The conversation should include all the key words: $key_1, key_2, ...$
Clinical Note: Note
Conversation: Note
Postediting Prompt
The above two paragraphs were extracted from a complete conversation. Please concatenate the two dialogues together. It means that your generation should include all the information such as the dosage of the medication which is mentioned in the clinical note. You should try to ensure that the dialogue is smooth. The conversation must include these key words: $key_1, key_2, ...$ and you should also eliminate the repeat parts.
History Conversation: Conversation 1
Generated Conversation: Conversation 2

Table 5: Prompts for Task-C.

Polish Prompt is utilized to modify previously generated conversations, enhancing their coherence and naturalness to resemble real-life conversations between doctors and patients. It also ensures that no important information is omitted during the conversation.

Hallucination Prompt is used to remove information generated by the model that is inconsistent with the clinical note. This ensures that the content generated by the model is more aligned with the clinical note.

Postediting Prompt is utilized to merge conversations from two different sections into a single smooth dialogue, ensuring that the conversation occurs at the same time and place.

By comparing our model to the conversations generated by GPT-4 (Table 8), we observed that our generated conversations are more logical and resemble human-to-human conversations more closely. For instance, in the GPT-4 generated conversation, the doctor did not ask any questions regarding symptoms such as fevers, chills, dizziness, nausea, vomiting, or diarrhea, as highlighted in the red utterance, but the patient answered that they did not have those symptoms. Regarding the conversations generated by ChatGPT (Table 6), we observed that the sentence structure of the questions posed by the doctor was highly uniform and the patient responses were quite mechanical, failing to resemble human-to-human conversations. Although GPT-4 somewhat addresses this issue, it still exhibits similar limitations. By contrast, our proposed method generates non-repetitive questions and replies, resulting in a more realistic dialogue dataset.

5 Conclusion

In conclusion, our work proposes innovative approaches to promote autonomous cooperation among communicative agents in medical settings to generate the doctor-patient dialogue datasets as valuable resources for developing AI systems that can better understand and address context-sensitive inquiries in healthcare communication. To validate the quality of our generated datasets, we establish a hybrid evaluation system consists of Rouge scores, concept recall, Sarebleurt and Self-bleurt. We use our evaluation system to compare our model with GPT4 and ChatGPT. We demonstrate that our method contributes to advancing the development of AI systems in healthcare communication and paves the way for more effective doctor-patient interactions.

6 Limitation

Although our method delivers optimal results, it doesn't comply with data protection regulations

ChatGPT
Doctor: Good morning, Brian. How have you been feeling lately?
Patient: I have been feeling out of sorts lately, not sure if it is due to the change in seasons or due to performing lots of projects and some construction on my home.
Doctor: That sounds exhausting. Have you been experiencing any symptoms?
Patient: Yes, I have been feeling fatigued and lightheaded for about five weeks now.
Doctor: Have you noticed any shortness of breath or chest discomfort while exerting energy?
Patient: Yes, I have experienced some shortness of breath and chest cramps.
Doctor: Have you been coughing lately?
Patient: Yes, I have a slight cough, but I am not sure if it's due to the change in seasons.
Doctor: How has your diet been lately?
Patient: My diet has been a little bit of a struggle. We have been eating less healthy food due to the construction work in the kitchen.
Doctor: How have you been doing with your salt intake?
Patient: I have been pretty good with my salt intake.
Doctor: How has your heart failure been? **Patient:** I have been pretty good with my diet since last year and have been staying on top of that as much as possible. I have continued to utilize Lasix daily.
Doctor: How has your hypertension been?
Patient: It has been well controlled with lisinopril 20 mg a day. I have been monitoring my blood pressure regularly.
...

Table 6: ChatGPT data sample for Task-C.

like HIPAA, even though Azure offers a HIPAA-compliant option. From a privacy standpoint, deploying a local model such as LED might be preferable, but our findings indicate that further work is needed for this method to achieve satisfactory performance. Regardless, when creating automated conversation-generation systems, healthcare providers and developers must ensure that the entire system—including text-to-dialogue, data transmission and storage, and model inference—complies with privacy and security standards to maintain trust and avoid privacy breaches in clinical environments. Hence, developing an automated conversation generation system from clinical note entails several ethical considerations.

Firstly, obtaining informed consent is vital: patients should be informed about their recordings, and data ownership must be emphasized. Equitable access is also crucial; the system must accommodate patients from various backgrounds, including those with disabilities, limited technical literacy, or language barriers. Lastly, continuous evaluations are required to ensure that the system's performance does not deteriorate and adversely affect the quality of care.

Our:
Doctor: Good morning, Brian. How are you feeling today?
Patient: Good morning, Doctor. I'm feeling a little out of sorts lately. I'm not sure if it's because of the change in seasons or because I've been doing a lot of projects around the house.
Patient: Sure. I've been feeling fatigued and lightheaded for about five weeks now. And when I'm active, I've been experiencing some shortness of breath and chest cramps. I've also had a slight cough, but I'm not sure if it's just because of the change in seasons.
Doctor: Have you noticed any bloating or changes in your diet?
Patient: Yeah, I've been feeling bloated every once in a while. And my diet has been a little bit of a struggle because we've had construction going on in our kitchen, so we've been eating less healthy food.
Doctor: Okay, let's talk about your medical history. You have a history of congestive heart failure, correct?
Patient: Yes, that's correct.
Doctor: And how have you been managing that?
Patient: I've been pretty good with my salt intake, and I've been keeping up with my diet since the last year. And I've been taking Lasix daily.
Doctor: Okay, good to know that. And what about your hypertension?
Patient: That's been well controlled with lisinopril 20 mg a day. I've been monitoring my blood pressure regularly.
Doctor: Great. Did you experience any other symptoms that you haven't mentioned yet, such as swelling in the lower extremities, fever, chills, dizziness, nausea, vomiting, weight gain, or diarrhea?
Patient: No, I didn't.
Doctor: Okay, thanks for letting me know. Based on what you've told me, I'm going to order some tests, and we'll follow up after those come back. In the meantime, try to get some rest and eat healthier if possible.
Patient: Alright, thank you.
...

Table 7: Our pipeline sample for Task-C.

GPT4
Doctor: Good morning, Mr. White. How have you been feeling since your last visit?
Patient: Good morning, doctor. I've been feeling a bit out of sorts lately. I'm not sure if it's due to the change in seasons or because I've been doing a lot of projects and construction on my home. **Doctor:** Can you be more specific about how you've been feeling?
Patient: Sure, I've been experiencing fatigue and lightheadedness for about 5 weeks now. When I exert energy, I sometimes get short of breath and have chest cramps.
Doctor: That's concerning. Have you noticed any other symptoms, like a cough or fever?
Patient: I've had a slight cough, but I'm not sure if it's just the change in seasons. No fever though.
Doctor: Have you experienced any changes in your diet or eating habits recently?
Patient: Yes, actually. We started doing construction on our kitchen over Labor Day weekend, and since then we've been eating less healthy food.
Doctor: And how have you been managing your heart failure and hypertension? Have you been monitoring your salt intake and taking your medications regularly?
Patient: I've been pretty good with my salt intake and have been watching my diet for the past year. I take Lasix daily for my heart failure, and my hypertension is controlled with lisinopril 20 mg a day. I monitor my blood pressure regularly.
Doctor: That's good to hear. You mentioned earlier that you filled out a review of the systems sheet when you checked in. Did you experience any weight gain, swelling in the lower extremities, or other symptoms?
Patient: No, I didn't experience any weight gain or swelling in my lower extremities. Also, no fevers, chills, dizziness, nausea, vomiting, or diarrhea.
Doctor: Alright, let's move on to the physical examination. I'm going to check your neck for any jugular venous distention.
Patient: Sure, go ahead.
...

Table 8: GPT4 data sample for Task-C.

References

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023a. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2283–2294.

Asma Ben Abacha, Wen-wai Yim, George Michalopoulos, and Thomas Lin. 2023b. An investigation of evaluation metrics for automated medical note generation. *arXiv preprint arXiv:2305.17364*.

George J Annas. 2003. Hipaa regulations: a new era of medical-record privacy? *New England Journal of Medicine*, 348:1486.

Asma Ben Abacha and Dina Demner-Fushman. 2019. On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2228–2234, Florence, Italy. Association for Computational Linguistics.

Asma Ben Abacha, Wen-wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen. 2023. Overview of the mediqa-chat 2023 shared tasks on the summarization and generation of doctor-patient conversations. In *ACL-ClinicalNLP 2023*.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, Dubrovnik, Croatia. Association for Computational Linguistics.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Olivier Bodenreider and Alexa T McCray. 2003. Exploring semantic groups through visual approaches. *Journal of biomedical informatics*, 36(6):414–432.

Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. Inpars: Data augmentation for information retrieval using large language models. *arXiv preprint arXiv:2202.05144*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4516–4525, Hong Kong, China. Association for Computational Linguistics.

Haw-Shiuan Chang, Zonghai Yao, Alolika Gon, Hong Yu, and Andrew McCallum. 2023. Revisiting the architectures like pointer networks to efficiently improve the next word distribution, summarization factuality, and beyond. *arXiv preprint arXiv:2305.12289*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Haixing Dai, Zheng Liu, Wenxiong Liao, Xiaoke Huang, Zihao Wu, Lin Zhao, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. 2023. Chataug: Leveraging chatgpt for text data augmentation. *ArXiv*, abs/2302.13007.

Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B Hall, and Ming-Wei Chang. 2022. Promptagator: Few-shot dense retrieval from 8 examples. *arXiv preprint arXiv:2209.11755*.

Robert Dale. 2021. Gpt-3: What's it good for? *Natural Language Engineering*, 27(1):113–118.

Bosheng Ding, Chengwei Qin, Linlin Liu, Lidong Bing, Shafiq Joty, and Boyang Li. 2022. Is gpt-3 a good data annotator? *arXiv preprint arXiv:2212.10450*.

Paul Drew, John Chatwin, and Sarah Collins. 2001. Conversation analysis: a method for research into interactions between patients and health-care professionals. *Health Expectations*, 4(1):58–70.

Hannah Eyre, Alec B Chapman, Kelly S Peterson, Jianlin Shi, Patrick R Alba, Makoto M Jones, Tamara L Box, Scott L DuVall, and Olga V Patterson. 2021. Launching into clinical space with medspacy: a new clinical text processing toolkit in python. In *AMIA Annual Symposium Proceedings*, volume 2021, page 438. American Medical Informatics Association.

Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. 2021. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in bioinformatics*, 22(3):bbaa110.

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. 2019. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17:1–9.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. *arXiv preprint arXiv:2212.14024*.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Kundan Krishna, Amy Pavel, Benjamin Schloss, Jeffrey P Bigham, and Zachary C Lipton. 2021. Extracting structured data from physician-patient conversations by predicting noteworthy utterances. In *Explainable AI in Healthcare and Medicine*, pages 155–169. Springer.

Sunjae Kwon, Zhichao Yang, and Hong Yu. 2022a. An automatic soap classification system using weakly supervision and transfer learning. *ArXiv*, abs/2211.14539.

Sunjae Kwon, Zonghai Yao, Harmon S Jordan, David A Levy, Brian Corner, and Hong Yu. 2022b. Medjex: A medical jargon extraction model with wiki's hyperlink span and contextualized masked language model score. *arXiv preprint arXiv:2210.05875*.

Eric Lehman, Vladislav Lialin, Katelyn Edelwina Legaspi, Anne Janelle Sy, Patricia Therese Pile, Nicole Rose Alberto, Richard Raymund Ragasa, Corinna Victoria Puyat, Marianne Katharina Taliño, Isabelle Rose Alberto, Pia Gabrielle Alfonso, Dana Moukheiber, Byron Wallace, Anna Rumshisky, Jennifer Liang, Preethi Raghavan, Leo Anthony Celi, and Peter Szolovits. 2022. Learning to ask like a physician. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 74–86, Seattle, WA. Association for Computational Linguistics.

David Michael Levine, Rudraksh Tuwani, Benjamin Kompa, A. Varma, Samuel G. Finlayson, Ateev Mehrotra, and Andrew L. Beam. 2023. The diagnostic and triage accuracy of the gpt-3 artificial intelligence model. *medRxiv*.

Zongjie Li, Chaozheng Wang, Pingchuan Ma, Chaowei Liu, Shuai Wang, Daoyuan Wu, and Cuiyun Gao. 2023. On the feasibility of specialized ability stealing for large language code models. *arXiv preprint arXiv:2303.03012*.

Wenxiong Liao, Zhengliang Liu, Haixing Dai, Shaochen Xu, Zihao Wu, Yiyang Zhang, Xiaoke Huang, Dajiang Zhu, Hongmin Cai, Tianming Liu, and Xiang Li. 2023. Differentiate chatgpt-generated and human-written medical texts.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

S. Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. *ArXiv*, abs/2301.13688.

Linyong Nan, Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, Yangxiaokang Liu, Nadia Irwanto, Jessica Pan, Faiaz Rahman, Ahmad Zaidi, Mutethia Mutuma, Yasin Tarabar, Ankit Gupta, Tao Yu, Yi Chern Tan, Xi Victoria Lin, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. 2021. DART: Open-domain structured data record to text generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 432–447, Online. Association for Computational Linguistics.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207, Melbourne, Australia. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering. In *ACM Conference on Health, Inference, and Learning*.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium. Association for Computational Linguistics.

Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. PriMock57: A dataset of primary care mock consultations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 588–598, Dublin, Ireland. Association for Computational Linguistics.

V Podder, V Lew, and S Ghassemzadeh. 2021. Soap notes.[updated 2021 sep 2]. *StatPearls [Internet]. StatPearls Publishing. Available from: https://www. ncbi. nlm. nih. gov/books/NBK482263*.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Thomas C Rindfleisch. 1997. Privacy, information technology, and health care. *Communications of the ACM*, 40(8):92–100.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M. Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. 2021. Multitask prompted training enables zero-shot task generalization. *CoRR*, abs/2110.08207.

Max E. Savery, Asma Ben Abacha, Soumya Gayen, and Dina Demner-Fushman. 2020. Question-driven summarization of answers to consumer health questions. *Scientific Data*, 7.

Benjamin Schloss and Sandeep Konam. 2020. Towards an automated soap note: classifying utterances from medical conversations. In *Machine Learning for Healthcare Conference*, pages 610–631. PMLR.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Anastasia Shimorina and Claire Gardent. 2018. Handling rare items in data-to-text generation. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 360–370, Tilburg University, The Netherlands. Association for Computational Linguistics.

Yan Song, Yuanhe Tian, Nan Wang, and Fei Xia. 2020. Summarizing medical conversations via identifying important utterances. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 717–729, Barcelona, Spain (Online).

Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. 2021. Can language models be biomedical knowledge bases? *arXiv preprint arXiv:2109.07154*.

Liyan Tang, Tanya Goyal, Alexander R Fabbri, Philippe Laban, Jiacheng Xu, Semih Yahvuz, Wojciech Kryściński, Justin F Rousseau, and Greg Durrett. 2022. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. *arXiv preprint arXiv:2205.12854*.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does synthetic data generation of llms help clinical text mining? *arXiv preprint arXiv:2303.04360*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Zhi Wen, Xing Han Lu, and Siva Reddy. 2020. MeDAL: Medical abbreviation disambiguation dataset for natural language understanding pretraining. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 130–135, Online. Association for Computational Linguistics.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*.

Zhichao Yang, Sunjae Kwon, Zonghai Yao, and Hongfeng Yu. 2022. Multi-label few-shot icd coding as autoregressive generation with prompt. *ArXiv*, abs/2211.13813.

Zonghai Yao, Liangliang Cao, and Huapu Pan. 2020. Zero-shot entity linking with efficient long range sequence modeling. *arXiv preprint arXiv:2010.06065*.

Zonghai Yao, Yi Cao, Zhichao Yang, Vijeta Deshpande, and Hong Yu. 2022a. Extracting biomedical factual knowledge using pretrained language model and electronic health record context. *arXiv preprint arXiv:2209.07859*.

Zonghai Yao, Yi Cao, Zhichao Yang, and Hong Yu. 2022b. Context variance evaluation of pretrained language models for prompt-based biomedical knowledge probing. *arXiv preprint arXiv:2211.10265*.

Zonghai Yao, Jack Tsai, Weisong Liu, David A Levy, Emily Druhl, Joel I Reisman, and Hong Yu. 2023. Automated identification of eviction status from electronic health record notes. *Journal of the American Medical Informatics Association*. Ocad081.

Zonghai Yao and Hong Yu. 2021. Improving formality style transfer with context-aware rule injection. *arXiv preprint arXiv:2106.00210*.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Acibench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Submitted to Nature Scientific Data*.

Wen-wai Yim and Meliha Yetisgen-Yildiz. 2021. Towards automating medical scribing: Clinic visit dialogue2note sentence alignment and snippet summarization. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 10–20.

Siyu Yuan, Deqing Yang, Jinxi Liu, Shuyu Tian, Jiaqing Liang, Yanghua Xiao, and Rui Xie. 2023. Causality-aware concept extraction based on knowledge-guided prompting. *arXiv preprint arXiv:2305.01876*.

Li Yunxiang, Li Zihan, Zhang Kai, Dan Ruilong, and Zhang You. 2023. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070*.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117, Online. Association for Computational Linguistics.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020a. Meddialog: Large-scale medical dialogue datasets. In *Conference on Empirical Methods in Natural Language Processing*.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, et al. 2020b. Meddialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhengyun Zhao, Qiao Jin, Fangyuan Chen, Tuorui Peng, and Sheng Yu. 2022. Pmc-patients: A large-scale dataset of patient summaries and relations for benchmarking retrieval-based clinical decision support systems.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100.

HealthMavericks@MEDIQA-Chat 2023: Benchmarking different Transformer based models for Clinical Dialogue Summarization

Kunal Suri
suri.kunal007@gmail.com

Saumajit Saha
saha.saumajit@gmail.com

Atul Singh
atulsingh.phd@gmail.com

Abstract

In recent years, we have seen many Transformer based models being created to address Dialog Summarization problem. While there has been a lot of work on understanding how these models stack against each other in summarizing regular conversations such as the ones found in DialogSum dataset, there haven't been many analysis of these models on Clinical Dialog Summarization. In this article, we describe our solution to MEDIQA-Chat 2023 Shared Tasks as part of ACL-ClinicalNLP 2023 workshop which benchmarks some of the popular Transformer Architectures such as BioBart, Flan-T5, DialogLED, and OpenAI GPT3 on the problem of Clinical Dialog Summarization. We analyse their performance on two tasks - summarizing short conversations and long conversations. In addition to this, we also benchmark two popular summarization ensemble methods and report their performance.

1 Introduction

It is essential to summarise the conversation between a doctor and a patient or another doctor to maintain records for compliance, training and evaluation. However this process, at the moment, is done manually which is time consuming and expensive. This paper presents the experimental results of our explorations with state-of-the-art deep-learning techniques to summarise such conversations to accomplish both SubTask A (Ben Abacha et al., 2023b) and B (wai Yim et al., 2023) of Dialogue2Note Summarization task from MEDIQA-Chat 2023 (Ben Abacha et al., 2023a). The solution of SubTask B presented in this paper was ranked fifth among all the submissions for SubTask B. The source code for the submission can be found in GitHub[1].

The paper uses Transformer based models for both assigning conversations into a pre-defined set of clinical notes sections and summarization of conversations. Through this work, the paper also compares the performance of Transformer based models for summarization tasks. This paper benchmarks performance of several Transformer based model for summarization task on medical conversation documents. In addition to this comparison, we also evaluate performance of two ensemble techniques namely (Kobayashi, 2018) and (Chen et al., 2021). Our simulations show that finetuning of Transformer-based models works as well as in-context prompt-based finetuning of OpenAI GPT3 which has usage-based costs and the risk of compromising your internal data to an external organization.

This paper is organized as follows. Section 3 presents a brief overview of the Dialogue2Note Summarization task, including the labeled data available and the evaluation metrics. Then the paper describes current state-of-the-art for clinical note summarization in Section 2 that this paper build upon. This is followed by the description of the approach used to solve the SubTask A of Dialogue2Note Summarization task in Section 4 and SubTask B in Section 5. Then the results of our solutions for Dialogue2Note Summarization tasks are presented. Finally, the paper ends with a conclusion on the work. The paper includes an appendix containing exploratory data analysis and material that will help to better understand the solution presented in the paper.

2 Related Work

In (Zhang et al., 2021) the authors have used both a single-stage and a two-stage approach for summarization. In the single-stage approach, the authors have truncated the input sentence length to match the BART transformer model input length constraints. In the multi-stage approach, the authors summarize the input conversation and then pass these summaries through a secondary model to

[1] https://github.com/suri-kunal/acl-medi-chat-summarization

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 472–489
July 14, 2023

generate the final summary. The authors have only focused on summarizing the History of Present Illness (HPI) section. The current work presented in this paper extends it beyond one section. It uses a two-step approach to generate the summary but does not focus on any special processing of the data in each section to generate the summaries.

In (Krishna et al., 2021) the authors combine extractive and abstractive summarization. They have presented a wide range of algorithms. Cluster2Sent, the most elaborate among these algorithms, first identifies the noteworthy utterances in each section and then clusters them before sending them to a summarization model. The present work presented in this paper is similar to this approach in that it does a Section level summarization. The current work depends on the power of more powerful models to summarise instead of processing the text in the Sections.

In (Chintagunta et al., 2021), the authors use GPT3 for medical summarization to achieve summaries that match human annotator-generated summaries using 30x lesser data. The authors generate k-candidate summaries for an input dialogue in this work. For each candidate summary generation, the authors sample N random examples from a small labelled data set. The examples, along with the input dialogue, are sent to GPT3 for summarization. In this work, the authors select N examples for each Section. The authors have yet to identify the medical terms in the generated summary to measure its effectiveness, which could be future work.

3 Dialogue2Note Summarization Task Description

This Section provides a high-level overview of the Dialogue2Note Summarization task (including both SubTask A and B) from MEDIQA-Chat 2023[2]. The Section starts with a description of the SubTask task goals, followed by basic counts of the available labelled data. The metric used to evaluate this task is arithmetic mean of ROUGE-1 (Lin, 2004), Bertscore F1 (Zhang et al., 2019), and BLEURT (Sellam et al., 2020).

3.1 Task Definition

Given a short conversation between a Doctor and a patient or another Doctor (**Dialogue**), the goal of SubTask A is to create a system that automatically predicts the Section to which the conversation belongs to which is denoted by **Section Header**. There are twenty Sections Headers in this dataset. Some examples of Section Headers are FAM/SOCHX, GENHX, PASTMEDICALHX, CC. All of these Section Headers and their descriptions (**Section Description**) can be found in Table A2. Another part of this SubTask is to generate a summary which matches the human generated summary (**Section Text**) as closely as possible while optimizing the metric for evaluation.

The aim of SubTask B is to summarize a given Doctor-Patient conversation (**Dialogue**) in a way that the generated summary matches the clinical note written by the physician (**Note**) as closely as possible. Unlike SubTask A, this task is a lot harder to solve because average length of a conversation is significantly longer than the dialogue of SubTask A. Please refer to Figure A1 for data distribution of SubTask A and Figure A2 for Dialogue data of SubTask B to understand the difference in distribution. A clinical note consists of the following high level sections called *First Level Sections* in this paper - *Subjective*, *Objective Exam*, *Objective Result*, *Assessment and Plan*. A clinical note comprises of several Section Headers each of which can be allocated to one of the First Level sections. Given a conversation between a Doctor and a patient, we create a system that automatically generates complete clinical note with all necessary First Level Sections.

3.2 Labelled Data

In this paper we have used the labelled data provided by MEDIQA-Chat 2023 organizers for training the models. A sample data point from the labelled data set for SubTask A can be found in Table A1. An example of a Doctor-Patient Conversation and corresponding Clinical Notes generated by a human from the labelled data set for SubTask B is also shown in Figure A2. The official data consists of a training and validation split. For SubTask A, the training data contains 1201 and validation data contains 180 <dialogue, section-text, section-header> triplets. For SubTask B, the training data contains 67 and validation data contains 20 <dialogue, note> pairs.

4 SubTask A Methodology

Given a short conversation between a doctor and a patient, the goal of SubTask A is to predict its Sec-

[2] https://sites.google.com/view/mediqa2023/clinicalnlp-mediqa-chat-2023

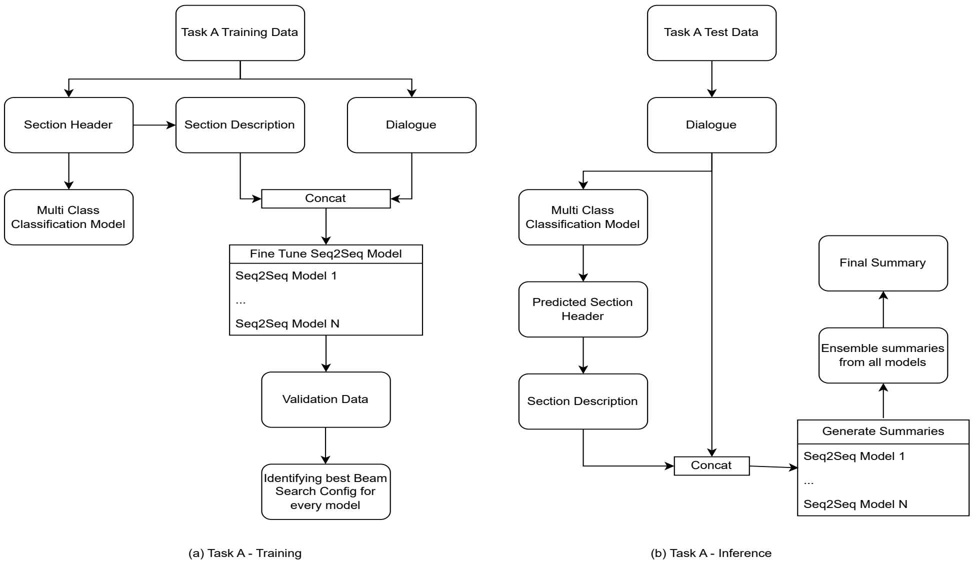


Figure 1: SubTask A - Overall Architecture

tion Header and summarize it while ensuring that the generated summary is as fluent and as close to Section Text as possible. This Section starts with a description of the approach used to predict the Section Header. This is followed by a description of the methodology used to summarize the conversation. For Dialogue Summarization, we have fine-tuned Transformer-based large language models. We have done an in-context fine-tuning of OpenAI GPT3 (Brown et al., 2020) and have fine-tuned four popular Transformer Sequence-to-Sequence (*Seq2Seq*) models. The Section describes the processed labelled data used for fine-tuning the Transformer based models, followed by the actual training steps. Then this Section looks at the steps used to generate the summary from the decoder. Finally, we discuss the two approaches used for ensembling the output of the four Transformer based models.

We achieved success using Bio-ClinicalBERT (Alsentzer et al., 2019) for classification in the healthcare domain and hence have fine-tuned this model for the classification of Dialogue to a Section Header in SubTask A. Since the target variable is highly imbalanced (see Figure A1a), we use Focal Loss (Lin et al., 2017) so that the algorithm focuses more on classes with fewer samples. We limit the number of input tokens to 300 tokens because that is the length of majority of dialogues, as shown in Figure A1. As the number of data samples available for training and validation is less, we use a 5 Fold Cross Validation approach for modelling purposes to ensure that we can capture all the information in the data. The hyper-parameters used for training and performance for all folds can be found in Table A3. During inference, we pass a given Dialogue through all five models, take an average of the logits for all the classes and output the class with the highest logit score.

We fine-tune Seq2Seq models using the labelled data (Dialogue,Section Text) for SubTask A as the (Input,Output) pair. Section Text is a part of the labelled data and is a human subject matter expert-created summary of Dialogue. As a pre-processing step, we replace all new line characters with whitespaces. The Dialogue is concatenated with the section description of its Section Header with the SEP token of the Seq2Seq architecture. While training, we use the actual section description for the actual Section Header and at inference, we use the section description corresponding to the predicted Section Header for the given Dialogue. No changes are made to Section Text.

We use a 5-fold cross validation scheme and fine-tune four Seq2Seq models - BioBart (Yuan et al., 2022), Flan-T5-Large (Chung et al., 2022), DialogLED-Base, and DialogLED-Large (Zhong et al., 2022) on each of the folds. Here we need to select the number of input tokens for encoder and decoder. For encoder we have selected token length of 512 tokens and for decoder we have selected token length of 400 tokens. All the hyper-parameters used to train each of the above architecture can be found in Table A4. To select the best model, we use early-stopping based on Validation Negative Log Loss (Yao et al., 2007). The out-of-fold results can be found in Table A6. The distribution of tokens for Dialogue and Section Text can be found in Figure A1b and Figure A1c respectively.

To generate summaries that match the human generated summaries, we need a way to control the summary generated by the decoder component of a Seq2Seq model. This can be done by using decoding strategies such as Beam Search (Graves, 2012), Top-k Sampling (Fan et al., 2018), Top-p Sampling (Holtzman et al., 2019), Contrastive Search (Su and Collier, 2023) etc. In this module, we use Beam Search with TPESampler Algorithm from Optuna[3] to search for the optimal decoding strategy trying to maximize ROUGE-1, ROUGE-2, and BertScore rather than relying on manual tweaking of these metrics. We use TPESampler here because it supports multivariate optimization and also it handles Float, Integer, and Categorical values better than other algorithms present in Optuna[4]. We use Optuna here due to ease of implementing Hyper-parameter optimization algorithms. We did not use BLEURT during search because it is extremely time consuming. For this module, we use four hyper-parameters for Beam Search - Early Stopping, Number of Beams, No Repeat N-gram Size, Length Penalty. The search space of each of these variables can be found in the Table 1.

Variable	Data Type	Range
Early Stopping	Categorical	[True,False]
Num_Beams	Integer	5-15
No_Rep_N_Size	Integer	5-15
Len_Pen	Float	[-2,2]

Table 1: Search Space for Beam Search Decoding. Num_Beams : Number of Beams, No_Rep_N_Size : No Repeat Ngram Size, Len_Pen : Length Penalty.

The approach used for in-context finetuning using OpenAI GPT3 is as follows: For every dialog in the test set, we predict and store the Section Header. We, then, randomly pick 3 Dialog-Summary-Section Header triplet from the entire (Training + Validation) dataset with the same Section Header. We use these triplets to create three summaries. These three summaries are then merged together to get the final summary. The configuration used for this task can be found in the appendix in Table A5 and its result on the test set can be found in Table 3 against Run 3.

In this paper We have used following approaches for ensembling:

- **Generating Best Summary by semantic similarity** - We use a post-ensemble method (Kobayashi, 2018) to identify the summary which is closest to all the generated summaries. This summary is then considered to be final summary for the given Dialogue.
- **Generating Best Summary by minimizing hallucination** - The above methodology helps us to get the summary closest to all the summaries but it does not account for the faithfulness of the generated summary with the actual Dialogue. To answer this question, we use the techniques introduced in (Chen et al., 2021). They have released a model[5] which we are using out of the box.

5 SubTask B Methodology

This Section presents an end-to-end solution to convert an entire Doctor-Patient Conversation (Dialogue) to Clinical Notes as SubTask B requires. The Section starts with a description of the supervised machine learning model used to predict the Section Header to which every utterance in a conversation belongs. All of these Section Headers are mapped to the First Level Sections using the mapping in Table 2. The output clinical note will contain these First Level Sections. The description of the classification model is followed by a description of the approach used to concatenate the utterances in a Dialogue belonging to a specific First Level Section. This is followed by a description of the Transformer models used to summarise the concatenated utterances. The results from the Transformer models are passed through an ensemble technique similar to the technique proposed in Section 4 to select the final summary, which is placed in the identified First Level Section in the Clinical Note.

We train a multi-label Classifier using the Dialogue and Section Header data from SubTask A to predict the Section Header to which an utterance belongs. As the data volume is very low, we use iterative-stratification package[6] to create 5 Folds

[3] https://optuna.readthedocs.io/en/stable/reference/samplers/generated/optuna.samplers.TPESampler.html

[4] https://optuna.readthedocs.io/en/stable/reference/samplers/index.html

[5] https://github.com/CogComp/faithful_summarization

[6] https://github.com/trent-b/iterative-stratification

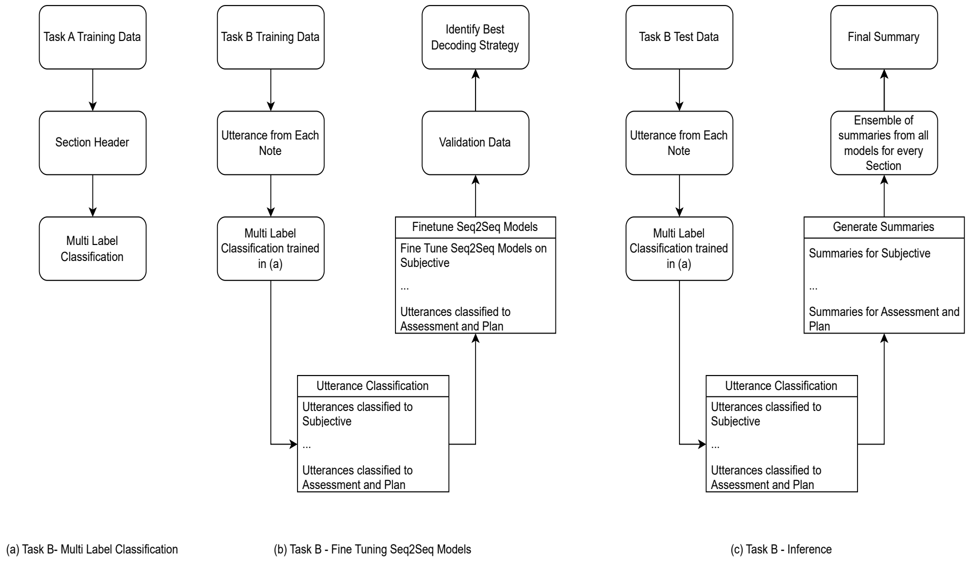


Figure 2: SubTask B - Overall Architecture

of the data and train one model for each fold - thus capturing all information in the data. Some of the labels have very low presence (see the distribution in Figure A1a), and hence we use Focal Loss instead of Binary Cross Entropy Loss so that the model can focus on labels which are harder to classify. The base model used for fine tuning is Bio-ClinicalBERT (Alsentzer et al., 2019) with number of input tokens as 512. We use early stopping to select the best model using Validation Negative Log Loss as the criteria for selecting the best model and Precision-Recall (PR) Score to evaluate performance of this model. The hyper-parameters for all folds and PR Score for all folds for all Section Headers can be found in Table A7 and Table A8 respectively.

We split every conversation on a new line character (\n) to get the list of constituent utterances. Each utterance is passed through each of the 5 Multi Label Models and we create a union of all predicted Section Headers from every model. Once every utterance has been mapped to all possible Section Headers, we transform the mapping so that we can combine all the utterances that belong to same Section Header. We ensure that utterance order should remain intact in all the sections. We then map all these Section Headers to their First Level Sections using mapping in Table 2. We have kept the mapping exhaustive to ensure that no False Negatives are left out. After this mapping, we merge all the utterances together and concat them together using whitespace character. We then split these utterances into their respective First Level Section and use the script provided by the organizers[7] to split the Note into these First Level Sections as well. The samples of dataset created after this step can be seen in the Figues A3a, A4a, A5a, and A6a. We have used the same high-level approach as in SubTask A for Dialogue Summarization. We fine-tune transformer based models, and have also used OpenAI GPT3 (Brown et al., 2020) with prompt based fine-tuning for summary generation.

We fine tune Seq2Seq models using (Utterance, Clinical Note Section) generated above as the (Input, Output) pair for every First Level Section. Before feeding the Utterance to the Encoder-Decoder models, we concatenate it with the section description of the First Level Section that the utterance belongs to, using the SEP token of the transformer architecture. We train two Seq2Seq models - DialogLED-Base and DialogLED-Large (Zhong et al., 2022) for each of the First Level Sections for each of the folds. The distribution of tokens for utterances of each First Level Sections and corresponding part of Clinical Note can be found in the Figures A3b, A3c, A4b, A4c, A5b, A5c, A6b, and A6c. All the hyper-parameters used to train each of the above architecture can also be found in Table A9.

Apart from finetuning Transformers we have also used OpenAI GPT3 (Brown et al., 2020) to generate summaries using prompt engineering. For every dialog in the test set, we pass it through Section 5 to split the Dialogue into utterances for every

[7]https://github.com/abachaa/MEDIQA-Chat-2023

First Level Section	Section Headers
Subjective	CC FAM/SOCHX GENHX PASTMEDICALHX PASTSURGICAL GYNHX OTHER_HISTORY ALLERGY ROS MEDICATIONS IMMUNIZATIONS
Objective_Exam	EXAM IMAGING LABS PROCEDURES
Objective_Results	IMAGING LABS DIAGNOSIS
Assessment_and_Plan	ASSESSMENT PLAN DISPOSITION PROCEDURES LABS MEDICATIONS EDSOURCE

Table 2: Mapping of First Level Section to Section Headers

First Level Section. We randomly pick 3 Dialog-Summary Pair for every First Level Section from the training data and truncate Dialog to 750 Tokens and Summary to number of tokens as per the First Level Section it belonged to. The number of tokens for summary section of each First Level Section can be found in Table A9. As for test dialog, we truncate it to 1000 tokens. The reason we do this is to adhere to the 4000 tokens length constraint of OpenAI GPT3 API. We concat the Train Dialog and Summary along with Test Dialog and generate a summary. This step is repeated three times. These three summaries are then merged together to get the final summary. We use the configuration in Table A5 for summarization.

To select the best model, we use early-stopping (Yao et al., 2007) using Validation Negative Log Loss and the metrics for best model for each architecture for each fold can be found in Table A3. We use the same search strategy for the optimal decoding strategy as we did for SubTask A except for one difference - we apply these techniques on the summaries generated for each First Level Sections separately. We use the same Model Ensembling Techniques that we have used in SubTask A except for one difference - we apply these techniques on the summaries generated for each First Level Sections separately.

6 SubTask A Results

This Section presents the results for SubTask A using the approach described in Section 4. We have made three submissions (mentioned as *runs* in the result tables) for predicting Section Header and three submissions for generating summaries from Dialogues. All the submissions for classification and final rankings of each of the runs can be found in Table A12. For the summarization task, we have also submitted results from three *runs*. In *run 1* and *run 2*, we have done the finetuning of the Transformer based models mentioned in Section 4 while *run 3* presents the results of summarization using OpenAI GPT3. The details for each run are as follows:

1. Run 1 - Post the summary generation, we ensemble output of all the models using *Generating Best Summary by minimizing hallucination* technique.

2. Run 2 - Post the summary generation, we ensemble output of all the models using *Generating Best Summary by semantic similarity*.

3. Run 3 - We use the an OpenAI GPT3 based approach described in Section 4.

The table containing our team's standing can be found in the Tables A12 and A13. Standings of all the teams have been calculated by calculating multi class accuracy for Section Header Classification and arithmetic mean of Rouge-1, Bertscore, BLEURT for the Dialogue summary.

The experiments show that Run1, which performs the worse on the ranking, hallucinates the most. This is counterintuitive since the goal of this approach is to minimize hallucination. We hypothesize that this could be because the model to detect hallucination was not trained on clinical data. Run2 gives the best summaries as measured by the evaluation metrics but has some hallucinations. Run3 gives the results with little to no hallucinations but has a lower score than Run2. This can happen because ROUGE Score always favours more extended conversations over shorter ones (Schluter, 2017). This can also be seen in Table 3 where BertScore and BLEURT are better for Run3 than for Run2 whereas ROUGE Score is better for Run2 than Run3. All the summaries generated by Run1, Run2, and Run3 are available in the github repository for the interested audience.

Run	R-1	B-F1	BLEURT	MS
Run 2	0.2973	0.612	0.4956	0.4683
Run 3	0.2514	0.6268	0.5015	0.4599
Run 1	0.1987	0.5703	0.4298	0.3996

Table 3: Results of runs on Test Data. R-1: ROUGE-1, B-F1: Bertscore-F1, MS: Mean Score

6.1 Analysis of different Transformer Architectures on the data

We analyse the performance of each Transformer architectures i.e. BioBart-V2-Base, Flan-T5-Large, DialogLED-Large, DialogLED-Base, and OpenAI GPT3 on the given dataset. We find that pretrained language Models such as DialogLED-Large, DialogLED-Base have performed consistently better than large language models such as OpenAI GPT3 and Flan-T5-Large. The performance was evaluated by calculating arithmetic mean of ROUGE-1, ROUGE-2, and BertScore-F1. We do not use BLEURT here as it is extremely time consuming and based on our observations, ROUGE-2 and BLEURT have a very strong correlation. The average score across all 5 folds for each architecture can be found in the Table A6.

7 SubTask B Results

As mentioned in Section 5, we have used a two step process using first a multi-label classification model to assign a conversation to a section and then applying Transformer based models on conversations for a section. Just like in SubTask A, we have made three submissions for generating summaries from Conversations. For Run 1 and Run 3, we map the utterances into first level sections, followed by summary generation for every first level Section. The decoding strategy for each first level section for each model can be found in Table A10. In this task, we tried using beam search configuration generated from Section 4 as well but this did not work well as we were getting Out Of Memory (OOM) errors. The summaries generated by this process is used below.

1. Run 1 - Post the summary generation, we ensemble output of all the models for every First Level Section using *Generating Best Summary by minimizing hallucination* technique.

2. Run 2 - We use the approach of OpenAI GPT3 from Section 5. We don't go in-depth for this approach since we couldn't analyse it due to cost constraints.

3. Run 3 - Post the summary generation, we ensemble output of all the models for every First Level Section using *Generating Best Summary by semantic similarity*

Our team's standing in the task of summarizing full note can seen in Table A14. It has been calculated by calculating the ROUGE-1 of the full note summary. Our team's standing in the task of summarizing complete note for all First Level Section can be found in Table A18. It has been calculated by calculating the arithmetic mean of arithmetic means of Rouge-1, Bertscore, BLEURT of every First Level Section Summary.

We have analysed the pros and cons of these three runs. Run1 gives the best result on this task but has some hallucinations. Run2 gives the results with little to no hallucinations but it has a lower score than Run1 and Run3. This can be because of the information loss that has happened as we are not able to consider all the tokens in the prompt. While Run3 performs better than Run2, it still has the most hallucination. This can be seen in Table 4.

Run	ROUGE-1
Run 1	0.5311
Run 3	0.5111
Run 2	0.2759

Table 4: Results of runs on Test Data for SubTask B

7.1 Analysis of different Transformer Architectures on the data

In SubTask B we analyse the performance of two Transformer based models namely DialogLED-Large, and DialogLED-Base. We are unable to compare their performance with OpenAI GPT3 as we have done in SubTask A because of cost and funding issues. We find that DialogLED-Large performs consistently better than DialogLED-Base. The performance is evaluated by calculating arithmetic mean of ROUGE-1, ROUGE-2, and BertScore-F1. The average score across all 5 folds for each architecture can be found in the Table 5. The performance of every model of every fold can be found in Table A11.

Arch	R-1	R-2	B-F1	MS
D-LED-B	0.5036	0.2257	0.6324	0.4539
D-LED-L	0.5235	0.2346	0.6388	0.4656

Table 5: SubTask B - Performance of different Transformer Architectures. Arch : Architecture, D-LED-B : Dialog-LED-Base, D-LED-L : Dialog-LED-Large, R-1 : ROUGE-1, R-2: ROUGE-2, B-F1 : Bertscore-F1, MS : Mean Score

8 Conclusion

The paper presents the solution and the results for SubTask A and B of Dialogue2Note Summarization task. The solution uses Transformer based models for both classification and summarization of Clinical Dialogs and the paper presents the comparison of the performance of Transformer based models on summarization. To the best of our knowledge, this is the first paper that benchmarks the performance of these models on Clinical Dialog Summarization. Our simulations and test scores show that fine-tuned transformer models work as well as in-context prompt based fine-tuning of Large Language Models such as OpenAI GPT3. This is encouraging for groups which either cannot afford huge API costs of using these Large Language Models or cannot send their data to their API due to regulatory restrictions. In addition to this, we also observe that metrics such as ROUGE might not be the suitable to gauge performance of models like OpenAI GPT3 as they focus on syntactic similarity. Metrics such as Bertscore and BLEURT seem to be more suitable for such models since they focus on semantic similarity. The paper also evaluated two different ensemble techniques and the results demonstrate that the Post Ensemble technique performs the best while also giving minimum hallucinations.

References

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly available clinical BERT embeddings. *CoRR*, abs/1904.03323.

Asma Ben Abacha, Wen wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen. 2023a. Overview of the mediqa-chat 2023 shared tasks on the summarization and generation of doctor-patient conversations. In *ACL-ClinicalNLP 2023*.

Asma Ben Abacha, Wen wai Yim, Yadan Fan, and Thomas Lin. 2023b. An empirical study of clinical note generation from doctor-patient encounters. In *EACL 2023*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 2021. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In *North American Chapter of the Association for Computational Linguistics*.

Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2021. Medically aware GPT-3 as a data generator for medical dialogue summarization. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 66–76, Online. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *CoRR*, abs/1211.3711.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *CoRR*, abs/1904.09751.

Hayato Kobayashi. 2018. Frustratingly easy model ensemble for abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4165–4176, Brussels, Belgium. Association for Computational Linguistics.

Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary C. Lipton. 2021. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4958–4972, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. *CoRR*, abs/1708.02002.

Natalie Schluter. 2017. The limits of automatic summarisation according to ROUGE. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 41–45, Valencia, Spain. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation.

Yixuan Su and Nigel Collier. 2023. Contrastive search is what you need for neural text generation.

Wen wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. The aci demo corpus: An open dataset for benchmarking the state-of-the-art for automatic note generation from doctor-patient conversations. *Submitted to Nature Scientific Data*.

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. 2007. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. Biobart: Pretraining and evaluation of a biomedical generative language model.

Longxiang Zhang, Renato Negrinho, Arindam Ghosh, Vasudevan Jagannathan, Hamid Reza Hassanzadeh, Thomas Schaaf, and Matthew R. Gormley. 2021. Leveraging pretrained models for automatic summarization of doctor-patient conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3693–3712, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. Dialoglm: Pre-trained model for long dialogue understanding and summarization.

A Appendix

A.1 Data Exploration and Explanation

This section discusses data exploration and explanation so that audience can understand why we made the decisions that we made.

A.1.1 SubTask A

A sample data point from dataset for SubTask A can be seen in Table A1.

Variable	Sample Value
Section Header	FAM/SOCHX
Section Text	The patient has been a smoker since the age of 10. So, he was smoking 2-3 packs per day. Since being started on Chantix, he says he has cut it down to half a pack per day. He does not abuse alcohol
Dialogue	Doctor: Are you a smoker? Patient: Yes. I do not drink if that is any constellation. Doctor: How much do you smoke per day? Patient: I just started taking Chantix and now I am down to a half a pack a day. Doctor: How much did you smoke per day prior to starting Chantix? Patient: I was smoking about two to three packs a day. I have been smoker since I was ten years old.

Table A1: Sample data point for SubTask A

The description of each of the Section Headers present in the data can be found in Table A2

The Data Exploration of this SubTask is give by Figure A1

The hyper-parameters and performance metrics for Predicting Section Header for every fold can be found in the Table A3. Each of the below configuration was run on Bio-ClinicalBert with Focal Loss.

The hyperparameters used to fine tune Seq2Seq Models can be found in Table A4. We run each of these models for 30 epochs with AdamW Optimizer, Learning Rate of 0.00002, and Linear Learning Scheduler.

Section Header	Section Header Description
FAM/SOCHX	FAMILY HISTORY/SOCIAL HISTORY
GENHX	HISTORY OF PRESENT ILLNESS
PASTMEDICALHX	PAST MEDICAL HISTORY
CC	CHIEF COMPLAINT
PASTSURGICAL	PAST SURGICAL HISTORY
ALLERGY	ALLERGY
ROS	REVIEW OF SYSTEMS
MEDICATIONS	MEDICATIONS
ASSESSMENT	ASSESSMENT
EXAM	EXAM
DIAGNOSIS	DIAGNOSIS
DISPOSITION	DISPOSITION
PLAN	PLAN
EDCOURSE	EMERGENCY DEPARTMENT COURSE
IMMUNIZATIONS	IMMUNIZATIONS
IMAGING	IMAGING
GYNHX	GYNECOLOGIC HISTORY
PROCEDURES	PROCEDURES
OTHER_HISTORY	OTHER_HISTORY
LABS	LABS

Table A2: Section Headers and their descriptions.

The configuration used for OpenAI GPT3 can be found in Table A5.

The average performance of these Seq2Seq Models for SubTask A can be found in Table A6. Here we didn't use BLEURT because it is a very time-consuming operation and based on our observations, ROUGE-2 is very well correlated with BLEURT.

A.1.2 SubTask B

The sample data and token distribution of this task is give by Figure A2

The sample data and token distribution of Subjective Section is give by Figure A3

The sample data and token distribution of Objective Exam Section is give by Figure A4

The sample data and token distribution of Objective Result Section is give by Figure A5

Fold	eps	WD	LR	WR	BS	Epochs	Seed	BE	BVA	BVL
4	1.00E-06	0.01	2.00E-05	0.1	16	30	42	18	0.8231	0.3796
3	1.00E-06	0.01	2.00E-05	0.1	16	30	42	18	0.7385	0.3688
2	1.00E-06	0.01	2.00E-05	0.1	16	30	42	18	0.8038	0.2131
1	1.00E-06	0.01	2.00E-05	0.1	16	30	42	18	0.7615	0.4397
0	1.00E-06	0.01	2.00E-05	0.1	16	30	42	18	0.7816	0.3912

Table A3: SubTask A - Predicting Section Header. eps : AdamW_eps, WD : AdamW Weight Decay, LR : Learning Rate, WR : Warmup Ratio, BS : Batch Size, BE : Best Epoch, BVA : Best Validation Accuracy, BVL : Best Validation Loss.

Architecture	GAS	BS	MaxSL	MaxTL	MinTL
Flan-T5-Large	3	5	512	400	8
Biobart-V2-Base	1	16	512	400	8
DialogLED-Large	3	6	512	400	8
DialogLED-Base	1	16	512	400	8

Table A4: SubTask A - Hyperparameter Tuning for Different Architectures. Optim : Optimizer, LR : Learning Rate, Sched : Scheduler, GAS : Gradient Accumulation Steps, BS : Batch Size, MaxSL : Maximum Source Length, MaxTL : Maximum Target Length, MinTL : Minimum Target Length

Hyperparameter	Value
Model	text-davinci-003
Temperature	0.5
Max Tokens	400
Top_p	1.
Frequency Penalty	0
Presence Penalty	0

Table A5: OpenAI GPT3 Hyperparameters

The sample data and token distribution of Assessment and Plan Section is give by Figure A6

The hyper-parameter setting for creating Multi Label Classification output can be seen in Table A7

The Precision Recall Score averaged over all the folds for all Section Headers can be found in the Table A8

We use the configuration in Table A9 for SubTask B summarization. Each of these models was trained for 30 Epochs with a Learning Rate of 0.00002, and a Linear Learning Scheduler.

The decoding strategy for SubTask B Summarization can be found in Table A10.

The performance of every architecture on every fold for SubTask B Summarization can be found in Table A9.

Model-Arch	R1	R2	BS-F1	MS
DL-Base	0.2471	0.0936	0.5803	0.3070
DL-Large	0.2444	0.0998	0.5741	0.3061
OpenAI GPT3	0.2233	0.0700	0.5917	0.2950
BBart-Base	0.1978	0.0767	0.5887	0.2877
FT5-Large	0.0589	0.0200	0.2458	0.1083

Table A6: SubTask A - Performance of different Transformer Architectures. Model-Arch - Model Architecture, R1 - Rouge-1, R2 - Rouge-2, BS-F1 - Bertscore-F1, MS - Mean Score, DL-Base - DialogLED-Base, DL-Large - DialogLED-Large, BBart-Base - BioBart-V2-Base, FT5-Large - Flan-T5-Large

Fold	4	3	2	1	0
AdamW_eps	0.000001	0.000001	0.000001	0.000001	0.000001
AdamW_weight_decay	0.01	0.01	0.01	0.01	0.01
batch_size	16	16	16	16	16
epochs	30	30	30	30	30
lr	0.00002	0.00002	0.00002	0.00002	0.00002
seed	42	42	42	42	42
warm_up_steps	0.1	0.1	0.1	0.1	0.1

Table A7: SubTask B - Hyperparameters used in Multi Label Classification

Model	Bio-ClinicalBERT
Section Header	**PR Score**
ALLERGY	95.77%
ASSESSMENT	34.58%
CC	55.31%
DIAGNOSIS	19.97%
DISPOSITION	71.10%
EDCOURSE	8.60%
EXAM	59.55%
FAM/SOCHX	97.16%
GENHX	90.18%
GYNHX	5.16%
IMAGING	40.61%
IMMUNIZATIONS	90.27%
LABS	25.00%
MEDICATIONS	94.87%
OTHER_HISTORY	0.43%
PASTMEDICALHX	75.84%
PASTSURGICAL	86.47%
PLAN	52.24%
PROCEDURES	2.05%
ROS	79.25%

Table A8: SubTask B - Precision Recall Scores for every Section Header

Architecture	FLS	BS	GAS	MaxSL	MinTL	MaxTL
Dial-LED-B	AP	8	2	3400	640	50
Dial-LED-L	AP	4	4	3400	640	50
Dial-LED-B	OE	8	2	3400	640	50
Dial-LED-L	OE	4	4	3400	640	50
Dial-LED-B	OR	8	2	3400	640	50
Dial-LED-L	OR	4	4	3400	640	50
Dial-LED-B	Subjective	8	2	3400	640	50
Dial-LED-L	Subjective	4	4	3400	640	50

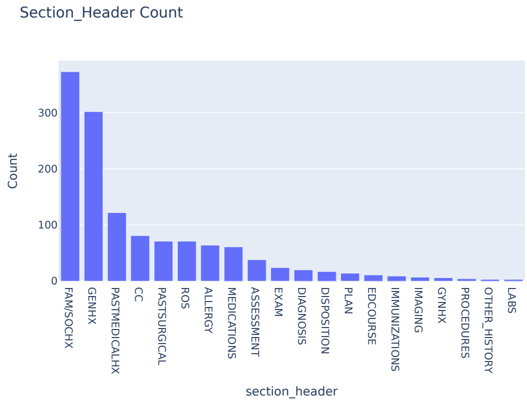
Table A9: SubTask B - Hyperparameters for every architecture for every First Level Section. Dial-LED-B : Dialog-LED-Base, Dial-LED-L : Dialog-LED-Large, FLS : First Level Section, BS : Batch Size, GAS : Gradient Accumulation Steps, LR : Learning Rate, MaxSL : Maximum Source Length, MinTL : Minimum Target Length, MaxTL : Maximum Target Length, AP : Assessment And Plan, OE : Objective Exam, OR : Objective Results

First Level Section	Architecture	Beams	Early Stopping	Length Penalty	No Repeat Ngram Size
Objective Exam	DialogLED-Base	5	TRUE	0.2	2
Objective Exam	DialogLED-Large	5	TRUE	0.2	2
Objective Result	DialogLED-Base	5	TRUE	0.2	2
Objective Result	DialogLED-Large	5	TRUE	0.2	2
Subjective	DialogLED-Base	5	TRUE	0.2	2
Subjective	DialogLED-Large	5	TRUE	0.2	2
Assessment and Plan	DialogLED-Base	5	TRUE	0.2	2
Assessment and Plan	DialogLED-Large	5	TRUE	0.2	2

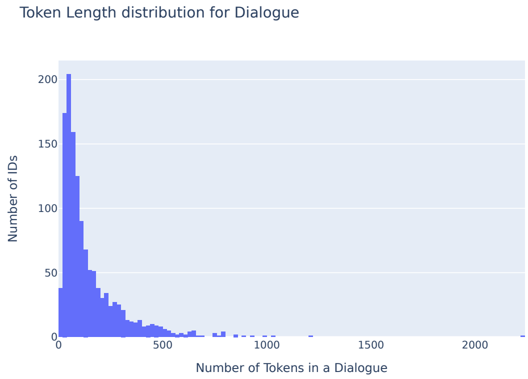
Table A10: SubTask B - Decoding Strategy

Architecture	Fold	Rouge1	Rouge2	BertScore-F1
DialogLED-Base	0	0.5189	0.2427	0.6415
DialogLED-Base	1	0.4854	0.2166	0.6231
DialogLED-Base	2	0.5345	0.2556	0.6497
DialogLED-Base	3	0.4832	0.1990	0.6198
DialogLED-Base	4	0.4958	0.2145	0.6281
DialogLED-Large	0	0.5109	0.2395	0.6467
DialogLED-Large	1	0.5459	0.2493	0.6362
DialogLED-Large	2	0.5569	0.2560	0.6530
DialogLED-Large	3	0.4917	0.2068	0.6269
DialogLED-Large	4	0.5122	0.2216	0.6310

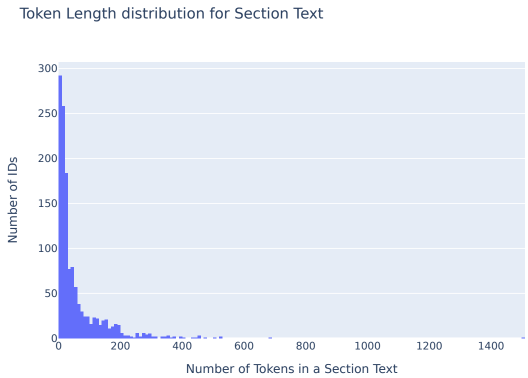
Table A11: SubTask B - Performance of every model of every fold



(a) Class distribution of Section Headers



(b) Dialogue Token Distribution

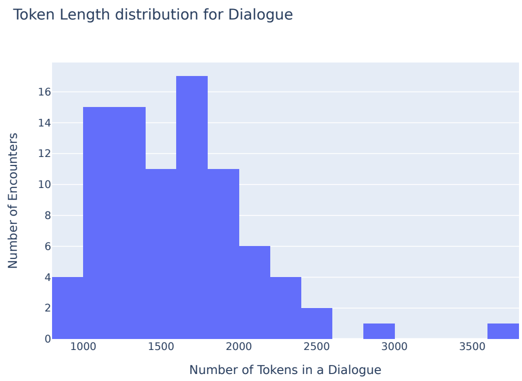


(c) Clinical Note Token Distribution

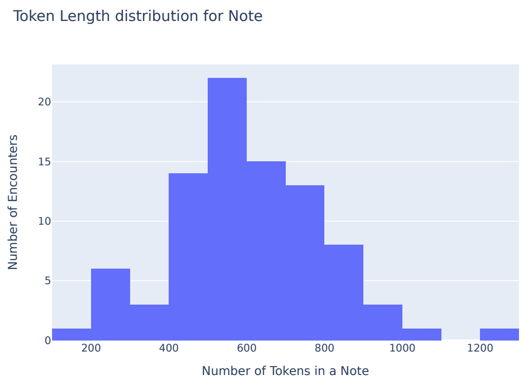
Figure A1: SubTask A - Data Exploration

Dialogue	Note
[doctor] hi , martha . how are you? [patient] i'm doing okay . how are you? [doctor] i'm doing okay . so , i know the nurse told you about dax . i'd like to tell dax a little bit about you , okay ? [patient] okay . [doctor] martha is a 50-year-old female with a past medical history significant for congestive heart failure , depression and hypertension who presents for her annual exam . so , martha , it's been a year since i've seen you . how are you doing ? ... [doctor] , some time during the day to take them , okay ? [patient] that might help me remember better . [doctor] all right . that sounds good . all right , well , it's good to see you . [patient] good seeing you too . [doctor] hey , dragon , finalize the note .	CHIEF COMPLAINT Annual exam. HISTORY OF PRESENT ILLNESS REVIEW OF SYSTEMS ... ASSESSMENT AND PLAN Martha Collins is a 50-year-old female with a past medical history significant for congestive heart failure, depression, and hypertension who presents for her annual exam. Congestive heart failure. Depression. Hypertension.

(a) Sample data



(b) Dialogue Token Distribution

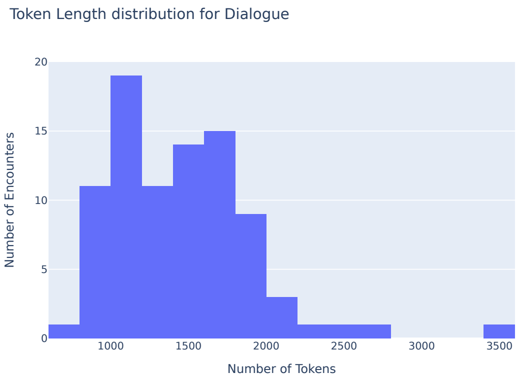


(c) Clinical Note Token Distribution

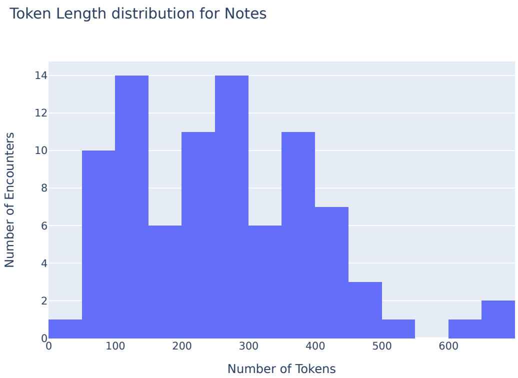
Figure A2: SubTask B - Sample Data

reference_subjective	dialogue_subjective
CHIEF COMPLAINT Annual exam. HISTORY OF PRESENT ILLNESS Martha Collins is a 50-year-old female with a past medical history significant for congestive heart failure, depression, and hypertension who presents for her annual exam. REVIEW OF SYSTEMS Ears, Nose, Mouth and Throat: Endorses nasal congestion from allergies. Cardiovascular: Denies chest pain or dyspnea on exertion. Respiratory: Denies shortness of breath. Gastrointestinal: Denies abdominal pain, nausea, or vomiting. Psychiatric: Endorses depression. Denies suicidal or homicidal ideations.	[doctor] hi , martha . how are you ? [patient] i'm doing okay . how are you ? [patient] okay . [doctor] martha is a 50-year-old female with a past medical history significant for congestive heart failure , depression and hypertension who presents for her annual exam . so , martha , it's been a year since i've seen you . how are you doing ? ... [doctor] all right . that sounds good . all right , well , it's good to see you . [patient] good seeing you too . [doctor] hey , dragon , finalize the note .

(a) Sample data



(b) Dialogue Token Distribution

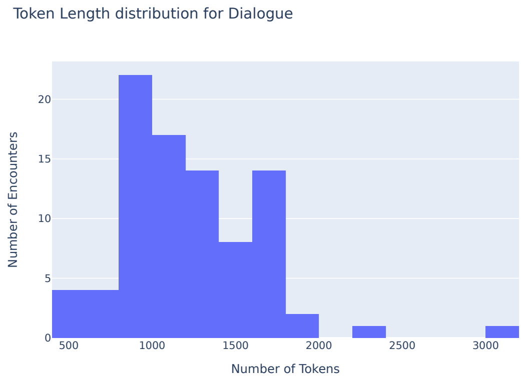


(c) Clinical Note Token Distribution

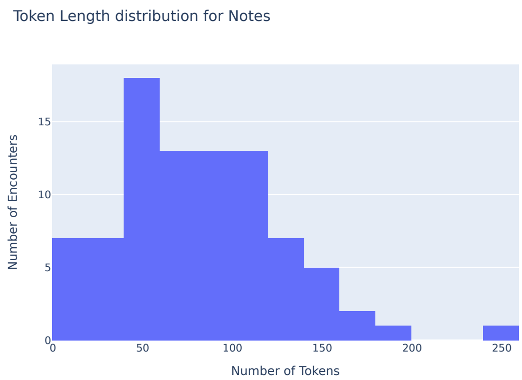
Figure A3: SubTask B - Subjective Section Sample Data

reference_objective_exam	dialogue_objective_exam
PHYSICAL EXAMINATION Cardiovascular: Grade 3/6 systolic ejection murmur. 1+ pitting edema of the bilateral lower extremities. VITALS REVIEWED Blood Pressure: Elevated.	[doctor] hi , martha . how are you ? [patient] okay . [doctor] martha is a 50-year-old female with a past medical history significant for congestive heart failure , depression and hypertension who presents for her annual exam . so , martha , it's been a year since i've seen you . how are you doing? ... [doctor] all right . that sounds good . all right , well , it's good to see you . [patient] good seeing you too . [doctor] hey , dragon , finalize the note .

(a) Sample data



(b) Dialogue Token Distribution

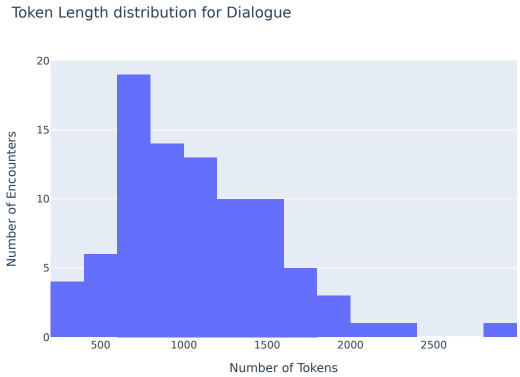


(c) Clinical Note Token Distribution

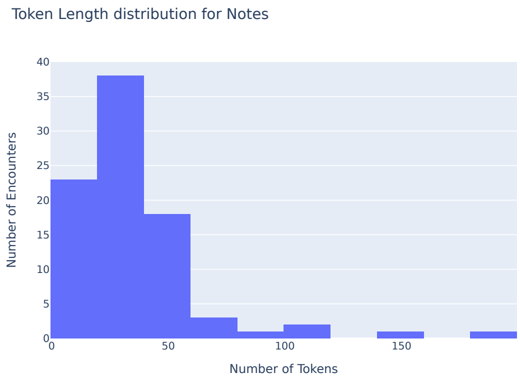
Figure A4: SubTask B - Objective Exam Section Sample Data

reference_objective_results	dialogue_objective_results
RESULTS Echocardiogram demonstrates decreased ejection fraction of 45%. Mitral regurgitation is present. Lipid panel: Elevated cholesterol.	[patient] i'm doing okay . how are you ? [doctor] i'm doing okay . so , i know the nurse told you about dax . i'd like to tell dax a little bit about you , okay ? ... [patient] okay . [doctor] . some time during the day to take them , okay ? [patient] that might help me remember better . [doctor] hey , dragon , finalize the note .

(a) Sample data



(b) Dialogue Token Distribution

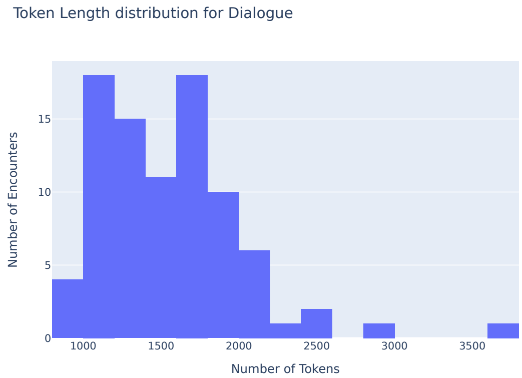


(c) Clinical Note Token Distribution

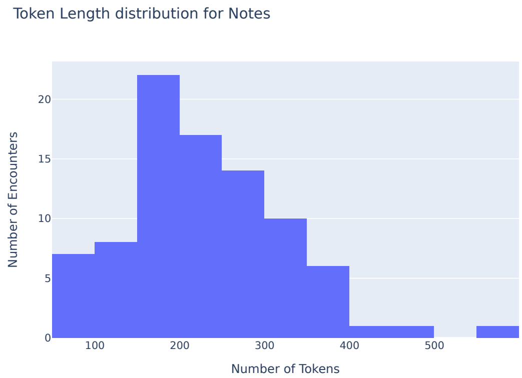
Figure A5: SubTask B - Objective Results Section sample data

reference_assessment_and_plan	dialogue_assessment_and_plan
ASSESSMENT AND PLAN Martha Collins is a 50-year-old female with a past medical history significant for congestive heart failure, depression, and hypertension who presents for her annual exam. Congestive heart failure. Medical Reasoning: She has been compliant with her medication and dietary modifications. ... Medical Reasoning: The patient is due for her routine mammogram. Additional Testing: We will order a mammogram and have this scheduled for her. Patient Agreements: The patient understands and agrees with the recommended medical treatment plan.	[doctor] hi , martha . how are you ? [patient] i'm doing okay . how are you ? [doctor] i'm doing okay . so , i know the nurse told you about dax . i'd like to tell dax a little bit about you , okay ? [patient] okay . ... [doctor] okay , all right . well , i'm glad to hear that . and you're taking your medication ? [patient] yes . [doctor] okay , good . and any symptoms like chest pains , shortness of breath , any swelling in your legs ? [patient] no , not that i've noticed . [patient] yeah , it's been helping a lot . i've been going every week , um , for the past year since my last annual exam . and that's been really helpful for me .

(a) Sample data



(b) Dialogue Token Distribution



(c) Clinical Note Token Distribution

Figure A6: SubTask B - Assessment and Plan Section sample data

A.2 Standing of our team

Our standings (in bold) for SubTask A - Section Header Classification is in Table A12. We omitted several teams from these standings and represent them by Ellipsis (**...**). This is done only to conserve space.

Team	Run	Accuracy	Rank
NUS-IDS	run1	0.78	1
...	...	...	...
Health-Mavericks	**run2**	**0.725**	**9**
Health-Mavericks	**run3**	**0.725**	**9**
Health-Mavericks	**run1**	**0.725**	**9**
...	...	...	...
Care4Lang	run2	0.345	31

Table A12: SubTask A - Section Header Classification Standings

Our standings (in bold) for SubTask A - Summarization is in Table A13

Team	Run	Mean Score	Rank
wanglab	run2	0.5789	1
...	...	...	...
Health-Mavericks	**run2**	**0.4683**	**25**
Health-Mavericks	**run3**	**0.4599**	**26**
ds4dh	run2	0.4334	27
Health-Mavericks	**run1**	**0.3996**	**28**
...	...	...	...
DFKI-MedIML	run1	0.3679	31

Table A13: SubTask A - Section Text Summarization Standings

Our standings (in bold) for SubTask B - Summarization is in Table A14

Team	Run	Rouge1	rank
wanglab	run3	0.6141	1
...	...	...	...
Health-Mavericks	**run1**	**0.5311**	**5**
...	...	...	...
Health-Mavericks	**run3**	**0.5111**	**11**
...	...	...	...
Health-Mavericks	**run2**	**0.2759**	**23**

Table A14: SubTask B - Notes Summarization Standings

Our standings (in bold) for Subjective Section can be found in Table A15

Team	Run	Subjective	Rank
wanglab	run1	0.6059	1
...	...	...	...
Health-Mavericks	**run1**	**0.4786**	**7**
...	...	...	...
Health-Mavericks	**run3**	**0.4657**	**12**
...	...	...	...
Health-Mavericks	**run2**	**0.3104**	**20**
...	...	...	...
Teddysum	run2	0.5353	23

Table A15: SubTask B - Subjective Section Performance

Our standings (in bold) for Objective Exam Section is in Table A16

Team	Run	Objective Exam	Rank
wanglab	run1	0.7102	1
...	...	...	...
Health-Mavericks	**run1**	**0.5374**	**7**
...	...	...	...
Health-Mavericks	**run3**	**0.4894**	**12**
...	...	...	...
Health-Mavericks	**run2**	**0.3222**	**20**
...	...	...	...
Teddysum	run2	0.1822	23

Table A16: SubTask B - Objective Exam Section Performance

Our standings (in bold) for Objective Results Section is in Table A17

Team	Run	Objective Results	Rank
wanglab	run1	0.6649	1
...	...	...	...
Health-Mavericks	**run1**	**0.5556**	**7**
...	...	...	...
Health-Mavericks	**run3**	**0.5383**	**12**
...	...	...	...
Health-Mavericks	**run2**	**0.3421**	**20**
...	...	...	...
Teddysum	run2	0.0182	23

Table A17: SubTask B - Objective Results Section Performance

Our standings (in bold) for Assessment and Plan Section is in Table A18

Team	Run	Assessment and Plan	Rank
wanglab	run1	0.6120	1
...	...	...	...
Health-Mavericks	**run1**	**0.4866**	**7**
...	...	...	...
Health-Mavericks	**run3**	**0.4854**	**12**
...	...	...	...
Health-Mavericks	**run2**	**0.3406**	**20**
...	...	...	...
Teddysum	run2	0.0968	23

Table A18: SubTask B - Assessment and Plan Section Performance

SummQA at MEDIQA-Chat 2023: In-Context Learning with GPT-4 for Medical Summarization

Yash Mathur* **Sanketh Rangreji*** **Raghav Kapoor***
Medha Palavalli* **Amanda Bertsch** **Matthew R. Gormley**
Carnegie Mellon University
[ymathur, srangrej, raghavka, mpalaval, abertsch, mgormley] @andrew.cmu.edu

Abstract

Medical dialogue summarization is challenging due to the unstructured nature of medical conversations, the use of medical terminology in gold summaries, and the need to identify key information across multiple symptom sets. We present a novel system for the Dialogue2Note Medical Summarization tasks in the MEDIQA 2023 Shared Task. Our approach for section-wise summarization (Task A) is a two-stage process of selecting semantically similar dialogues and using the top-k similar dialogues as in-context examples for GPT-4. For full-note summarization (Task B), we use a similar solution with k=1. We achieved 3rd place in Task A (2nd among all teams), 4th place in Task B Division Wise Summarization (2nd among all teams), 15th place in Task A Section Header Classification (9th among all teams), and 8th place among all teams in Task B. Our results highlight the effectiveness of few-shot prompting for this task, though we also identify several weaknesses of prompting-based approaches. We compare GPT-4 performance with several finetuned baselines. We find that GPT-4 summaries are more abstractive and shorter. We make our code publicly available [1].

1 Introduction

Medical dialogue summarization is a long-standing challenge in NLP (López Espejel, 2019; Joshi et al., 2020a; Chintagunta et al., 2021; Navarro et al., 2022). Medical scribes write notes on doctor-patient conversations in a predefined template called SOAP notes (Ullman et al., 2021; Podder et al., 2020), which contains sections for information from the patient, test results and observations, diagnosis, and the conclusion or treatment.

Medical summarization is challenging for several reasons. It requires dialogue understanding, where data is often limited (Dai et al., 2020; Lin et al., 2020); this is compounded by the sensitive nature of medical information, which restricts the release of training data for this task (Johnson et al., 2023). Doctors and patients may discuss several conditions in the same conversation, requiring the scribe to differentiate (Gidwani et al., 2017; Mishra et al., 2018). Scribes often use medical terminology in the notes that are not present in the doctor-patient conversation (Corby et al., 2020). Additionally, medical summarization is a high-stakes domain (Naik et al., 2022), motivating several efforts to build explainabile systems for this task (Jain et al., 2022; Reddy, 2022). In parallel, research on large language models (LLMs) has demonstrated compelling few-shot capabilities across domains (Brown et al., 2020; Perez et al., 2021).

In this paper, we explore several potential applications of a recent LLM, GPT-4 (OpenAI, 2023), on medical summarization. We use GPT-4 and finetuned BioBERT (Lee et al., 2020) as an ensemble for classifying the section headers of medical summaries, a 20-category classification problem. Then, given a candidate section header, we apply Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998) to select examples for a fewshot demonstration and use these examples to prompt GPT-4 for section-wise summarization. This approach outperforms finetuning BART (Lewis et al., 2019) and T5 (Raffel et al., 2020) over the limited available data. For full-note summarization, we take a similar approach, but select only a single example for the demonstration due to the increased length of the inputs. This also outperforms our supervised baselines. We outline several additional potential prompting approaches and compare their relative efficacy.

Applying LLMs for medical summarization is a compelling solution to the data scarcity problems in this domain, and we find promising performance, with our team placing second in the MEDIQA 2023 Shared Task for Subtask A and Division Summary

*Equal contribution
[1] https://github.com/Raghav1606/SummQA

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 490–502
July 14, 2023

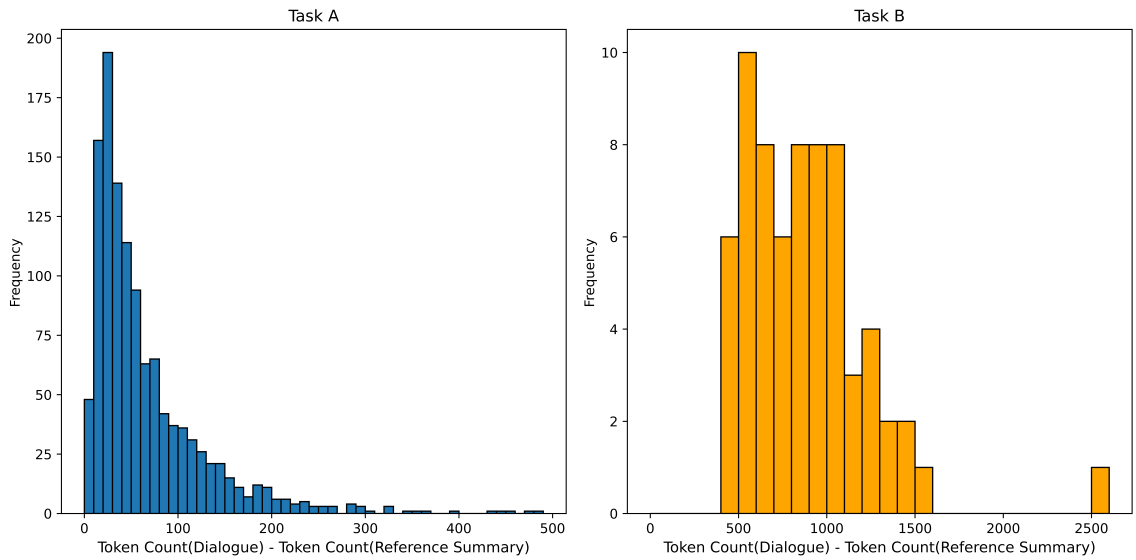


Figure 1: Distribution of difference in length between dialogue and reference summary. A larger difference in length indicates a higher degree of compression.

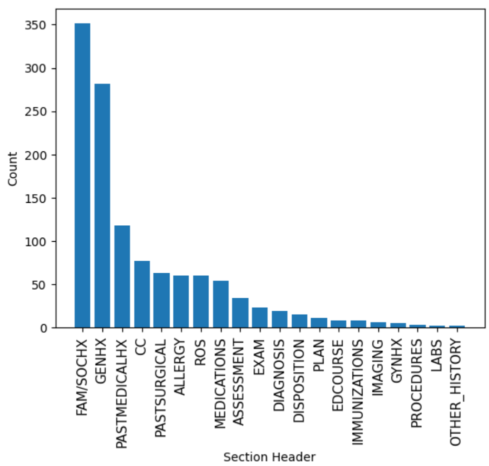


Figure 2: Section header distribution for Task A.

for Subtask B. However, we also identify key areas for improvement. We analyze the differences in outputs between the settings in output length and extractive ability. We find that the summaries generated by LLMs tend to be shorter and less extractive than human-generated summaries as well as SOTA fine-tuned biomedical summarization models. We also note the impracticality of this approach for real data, due to privacy concerns.

2 Background

Dialogue2Note Summarization was one of two tracks in the MEDIQA-Chat 2023 shared task (Ben Abacha et al., 2023). The track was further comprised of two tasks.

Task A involves generating a section-specific clinical summary from a conversation between a patient and a doctor. Additionally, Task A includes a classification task: assigning each dialogue an appropriate section header. There are 1200 conversations in the training split of the dataset (Ben Abacha et al., 2023) for Task A. The distribution over the section headers in Task A is a long-tailed distribution, displayed in Figure 2.

Task B involves generating a full note summary given a conversation; these summaries were evaluated on the section-level and the full-note level. There were 67 conversations in the training split of the dataset (Yim et al., 2023); these dialogues and reference summaries are significantly longer than those for Task A, as these dialogues encompass an entire conversation between a patient and a doctor. The distribution of the difference in dialogue and summary length for both tasks is shown in Figure 1.

3 Related Work

Summarization In recent years, fine-tuning pretrained models on domain-specific datasets has been the leading practice in text summarization research. While these models produce high-quality summaries and earn high scores against standard benchmarks, they require large datasets in order to adapt to specific domains or summarization styles (Lewis et al., 2020). Transformer-based models (Michalopoulos et al., 2022) and pointer generator

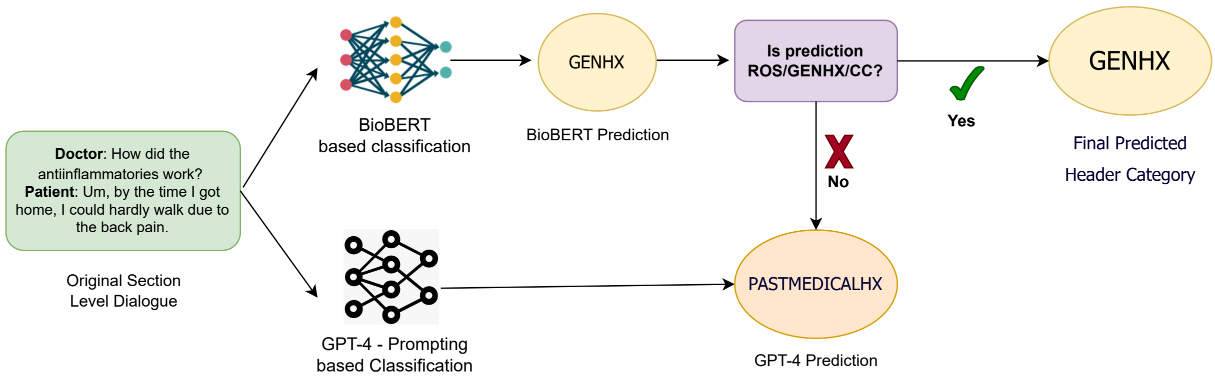


Figure 3: System Architecture for Section Header Classification (Task A)

network models (Joshi et al., 2020b) have been fine-tuned with medical domain knowledge to produce summaries that achieve state-of-the-art results.

Maximal Marginal Relevance was created to reduce redundancies in multi-document summaries (Goldstein and Carbonell, 1998). Abdullah et al. (2023) used MMR to generate query-focused summaries from pre-trained models without performing fine-tuning. (Ye et al., 2022) use MMR to select examples for in-context prompting.

The success of prompt-based models such as GPT-3 (Brown et al., 2020) has allowed for learning from natural language task instructions and/or a few demonstrative examples in the context without updating model parameters. In news summarization, Goyal et al. (2022) find that GPT-3 summaries were preferred by humans over summaries from fine-tuned models trained on large summarization datasets; they posit that zero-shot summaries avoid pitfalls from low-quality training data that are common in summaries from fine-tuned models. In the biomedical domain, pre-trained language models and few-shot learning has been used to collect and generate labeled data for medical dialogue summarization (Chintagunta et al., 2021). Recent work has used GPT-4 to pass the USMLE without any specialized prompt crafting (Nori et al., 2023) and perform zero-shot medical evidence summarization across six clinical domains (Tang et al., 2023).

Few-shot learning Few-shot learning can be unstable as the prompt format, training examples, and even the order of the training examples can cause accuracy to vary from near chance to near state-of-the-art (Lu et al., 2022). Recent work on prompting has tried to mitigate these problems through techniques such as calibration (Zhao et al., 2021), prompt combination (Zhou et al., 2022), or automatic prompt generation (Gao et al., 2021).

To mitigate any instability caused by a model's bias, Zhao et al. (2021) estimated the bias towards each answer by asking for its prediction when given the training prompt and a content-free test input such as "N/A" and then fit calibration parameters that cause the prediction for this input to be uniform across answers. To date, studies in prompt combination are rooted in paraphrasing-based methods that take a seed prompt and paraphrase it into several semantically similar expressions. Typically simple ensemble methods (Zhou et al., 2022) such as Maximal Marginal Relevance (Mao et al., 2020) are used to combine the answers to the different prompts as to provide each prompt to contribute to the final answer.

A number of techniques have also been proposed for selecting fewshot examples (Rubin et al., 2022). Fewshot techniques often rely on selecting optimal examples from a large dataset; some work has shown that this leads to an overstatement of few-shot performance, as a large number of labeled examples are necessary to select good examples for the fewshot prompt (Perez et al., 2021). We note that we use the full datasets (1,200 examples for Task A, 40 for Task B) for our prompt selection techniques.

4 Methodology and Baselines

Our summary generation pipeline remains the same across the two tasks: we use GPT-4 to generate a summary given k in-context examples.

4.1 Task A - Section Level Summary

Task A is composed of two subtasks, namely the section header classification and the section-level summarization. We discuss our approach for each

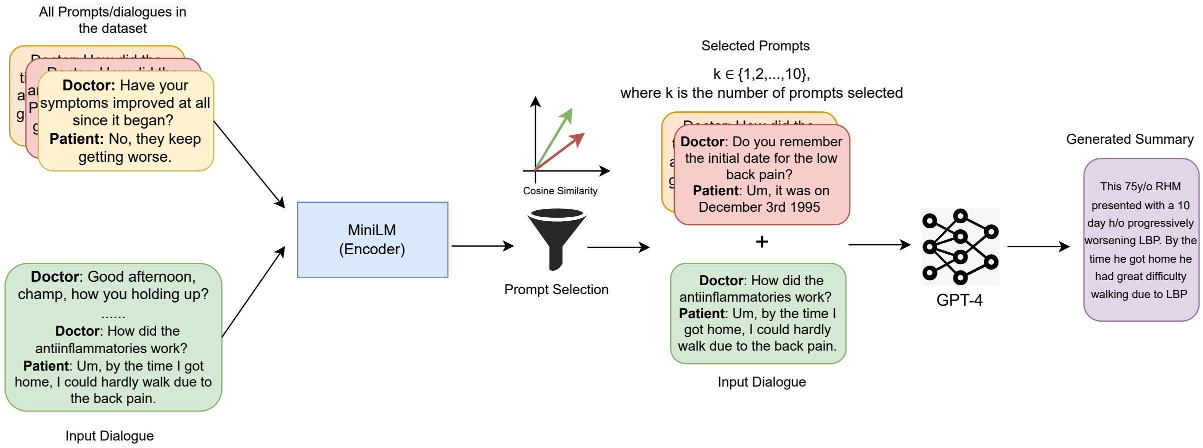


Figure 4: System Architecture for Summarization Task

of the subtasks below.

4.1.1 Section Header Classification

For the section header classification task, we create an ensemble of two models: BioBERT (Lee et al., 2020) and GPT-4. We fine-tune BioBERT with the training data provided for task A. We leverage GPT-4 to perform zero-shot classification on a sample with a given prompt (shown in Table 7). During our analysis of each model's performance, we observe that each model is more accurate than the other on a distinct subset of classes. To leverage the varying nature of predictions from the models we build an ensemble classifier. The overall accuracies are shown in Table 2. We observe empirically that our prompting-based approaches do not perform well on three of the section headers: ROS (Review of Systems), GENHX (History of the present illness), and CC (Chief Complaint), To create an ensemble classifier, we select BioBERT's predictions when it classifies a dialogue as one of these three section headers, and we select the prediction of our GPT-4 based pipeline otherwise. We present the architecture of our final model in Figure 3.

4.1.2 Section Summary

To generate summaries for a given section, we follow a multi-step process as shown in Figure 4. We encode each dialogue in the training data with MiniLM (Wang et al., 2020). For each dialogue to be summarized, we calculate cosine similarities with encoded dialogues from the training data. We retrieve k=7 similar examples from the training data based on the highest similarity. This similarity search, using cosine-similarity, serves as a prompt selection method, and the resulting few-shot prompts, with k=7 are then fed to GPT-4 along with their section headers to obtain a summary for the given section. We provide the prompt templates used in Table 7. We select k=7 as this fits well in the context length of our prompting-based pipeline; we perform an analysis with varying k in section 5.5.

4.2 In-context Example Selection for Summarization Tasks

This approach involves the dynamic selection of in-context examples for each example during validation or testing. This process entails utilizing matching or similarity criteria to compare the input dialogue of a specific example to a candidate pool comprising the complete training set. Through this process, we are able to select the most suitable examples for each individual case, thereby enhancing the efficacy of our prompts.

Semantic Similarity Here we select the k most similar examples (dialogue and summary pairs) based on semantic similarity between the provided input dialogue and the input dialogues in the training set. We store the selected examples and generate prompts which we then evaluate on the validation/test set.

Maximal Marginal Relevance We select k few-shot prompts using Maximal Marginal Relevance (MMR). Similarly to Ye et al. (2022), we use MMR to select an example and use it as a one-shot example for prompting. Our choice of MMR was motivated by the idea that the diversity in the selected in-context examples of the prompt would help with some generalization;.

4.3 Task B - Full Note Summary

For the summarization of entire dialogues, our goal is to generate a full note containing all the appropriate sections. We use a similar approach as described in section 4.1.2 but we restrict it to k=1 similar examples from the training set and include section-level headers in the prompts to help the model understand the sections in the sumary. We selected one in-context example due to long example length relative to the context window of the model. The one-shot prompt is then again fed to the GPT-4 model to obtain a full-length summary. The results f

4.4 Baseline Approaches

We also consider a variety of baseline approaches including, supervised fine-tuning of T5, zero-shot/few-shot GPT-3, perspective-shifting the dialogue followed by summarization, two-stage prompting, our similarity-based in-context learning applied to GPT-3, and mixing of extractive/abstractive methods.

4.4.1 T5

We fine-tuned the T5-small model for the end-end full-length summarization task (Task B). We fine-tuned for 20 epochs with a learning rate of 0.001. Our objective was to obtain a basic model that can serve as a benchmark to assess the complexity and difficulties associated with this specific task. We find that this finetuned model significantly underperforms our other methods, with a ROUGE-1 of 20.187; this may be due to the small dataset for finetuning or a non-optimized set of hyperparameters, as we do not do extensive hyperparameter search.

4.4.2 GPT-3

We investigated several prompting strategies and approaches using *text-davinci-003*.

Zero-shot prompting For Task B we used the prompt template mentioned in the Appendix A, where we specified the dialogue to be summarized with an instruction prompt mentioning the 4 main sections usually reported in the SOAP notes - "HISTORY OF PRESENT ILLNESS", "PHYSICAL EXAM", "RESULTS" and "ASSESSMENT AND PLAN". The zero-shot prompt gave us a reasonably high ROUGE-1 score of 45.911.

4.4.3 Few-shot prompting for section-wise summary

For Task A, we employed `text-davinci-003` few-shot prompting strategy. Initially, we grouped and categorized the existing 20 section headers for the dataset into 4 main sections, namely "HISTORY OF PRESENT ILLNESS", "PHYSICAL EXAM", "ASSESSMENT AND PLAN", and "RESULTS". The categorization scheme is detailed in Table 6. It is worth noting that "Medications" can be categorized under either "HISTORY OF PRESENT ILLNESS" or "ASSESSMENT AND PLAN" and therefore appears in both categories. We created four few-shot prompt templates, each comprising k=5 in-context examples, for each section. For each example in the validation set, we selected the appropriate prompt based on the classified section header.

Perspective Shift In this method which we evaluated for Task B, we adopt a two-stage prompting approach where we first use gpt3.5-turbo to obtain a third-person narrative from the input dialogue, following Bertsch et al. (2022), and use the third-person perspective narrative generated as input to a *text-davinci-003* model to generate a summary using the same instruction prompt specifying each section header that needs to be generated.

Two Stage Prompting In this approach we defined two chained prompts applied one after the other in a stage-wise manner. The first stage prompt was "List the important points from the above conversation for a medical report". This generated a list of salient points summarizing the dialogue. The second stage prompt we used was "Create a paragraph from the above facts only". The output from this prompt served as the final summary, which we then evaluated. We opted for these specific phrasings in the second prompt to mitigate the issue of model hallucination, which we observed was prevalent when tasked with generating a medical summary directly.

5 Results and Analysis

5.1 Experimental Setup

We used an 80/20 train/validation split on the training set and used the entire validation split as our test set. The main hyperparameter that we varied across our experiments for prompt selection was k, the number of in-context examples we selected for the prompt. We report the ablation study on

Task	Models	R1	R2	RL	BR	BP	BF1	BL
A	Few-shot Text-davinci-003	13.369	3.534	9.559	0.812	0.857	0.833	-1.173
	T5-Small	30.857	11.123	24.283	0.876	0.891	0.883	-0.637
	Two Stage Prompting	28.310	11.521	21.612	0.878	0.889	0.884	-0.550
	Prompt Selection text-davinci-003 (Semantic)	38.597	**18.393**	31.317	0.904	0.897	0.900	-0.401
	Prompt Selection text-davinci-003 (MMR)	40.213	16.286	32.903	0.903	0.899	0.901	-0.359
	Prompt Selection GPT4 (Semantic)	**42.841**	17.163	**34.808**	**0.907**	**0.909**	**0.907**	**-0.265**
B	Perspective Shift	20.433	4.969	12.384	0.810	0.866	0.837	-0.521
	T5-Small	20.187	8.287	12.420	0.790	0.852	0.820	-1.004
	Zero Shot Text-davinci-003	45.911	23.128	30.633	**0.851**	0.881	0.866	-0.610
	Prompt Selection GPT4 (k = 1)	**52.767**	**37.821**	**43.607**	0.846	**0.891**	**0.868**	**-0.336**

Table 1: Validation Results for Task A and Task B Summarization. Metrics include ROUGE-1 (R1), ROUGE-2 (R2), ROUGE-L (RL), BERTScore Precision (BP), Recall (BR), and F1 (BF1), and BLEURT (BL).

Model	Accuracy
GPT-3.5-turbo	68.943
GPT-4	69.474
BioBERT	71.278
Ensemble (GPT-4 + BioBERT)	**75.312**

Table 2: Validation Results for Header Classification

Task	Summary	EFC	EFD	CR
A	Reference	0.689	1.648	3.387
	Generated	0.561	1.036	5.701
B	Reference	0.671	2.044	2.856
	Generated	0.781	3.086	5.281

Table 3: Summary extractiveness comparison - Extractive Fragment Coverage(EFC), Extractive Fragment Density(EFD), Compression Ratio (CR)

varying k over the validation split in Table 5. For generations we used a single decoding (n = 1), temperature = 1.0, top_p = 1.0 and max_{tokens} = 800. The metrics for BERTScore and BLEURT in Table 1 have been calculated using RoBERTa Large (Liu et al., 2019) and BLEURT-Tiny[2] respectively.

5.2 Experimental Results

Our experiment involving prompt selection via semantic similarity with GPT-4 yielded the most favorable outcomes on the validation split, and prompt selection was the best approach for both Task A and Task B. We propose that the remarkable performance of prompt selection is attributed to the in-context examples that were selected using semantic similarity with the input dialogue. This approach facilitates the generation of an example-specific prompt that incorporates similar in-context examples, leading to an improvement in the model's ability to produce summaries that are more relevant and precise. The use of semantic similarity allows for the identification of examples that share similar semantic structures with the input dialogue, thereby increasing the likelihood of generating coherent and accurate summaries.

[2]https://github.com/google-research/bleurt

5.3 Length of Generated Summary vs. Reference Summary

As shown in Figure 5 we see that most generated summaries were shorter than reference summaries across tasks. This difference was more pronounced in Task B and therefore the summaries produced by our approach fall short in length thereby affecting the ROUGE-1 score as the number of matching n-grams is less. However, we observe that the BERT score still remains consistent even while producing shorter summaries.

Another interesting observation is that individual section summaries, when combined together to produce a full-length summary are closer to the original length rather than prompting GPT-4 to generate a complete summary together. Hence, ensembling multiple section-level summaries to produce a longer summary is an approach we can explore further. We also tried multiple prompt templates (refer Table 7), encouraging the model to produce longer summaries. However, the fact that we require a *summary* induces the model to be concise.

5.4 Extractiveness of Summaries

We measure the extractiveness of the generated summaries using three measures namely. extractive fragment coverage (EFC) (Grusky et al., 2020), extractive fragment density (EFD) (Grusky et al.,

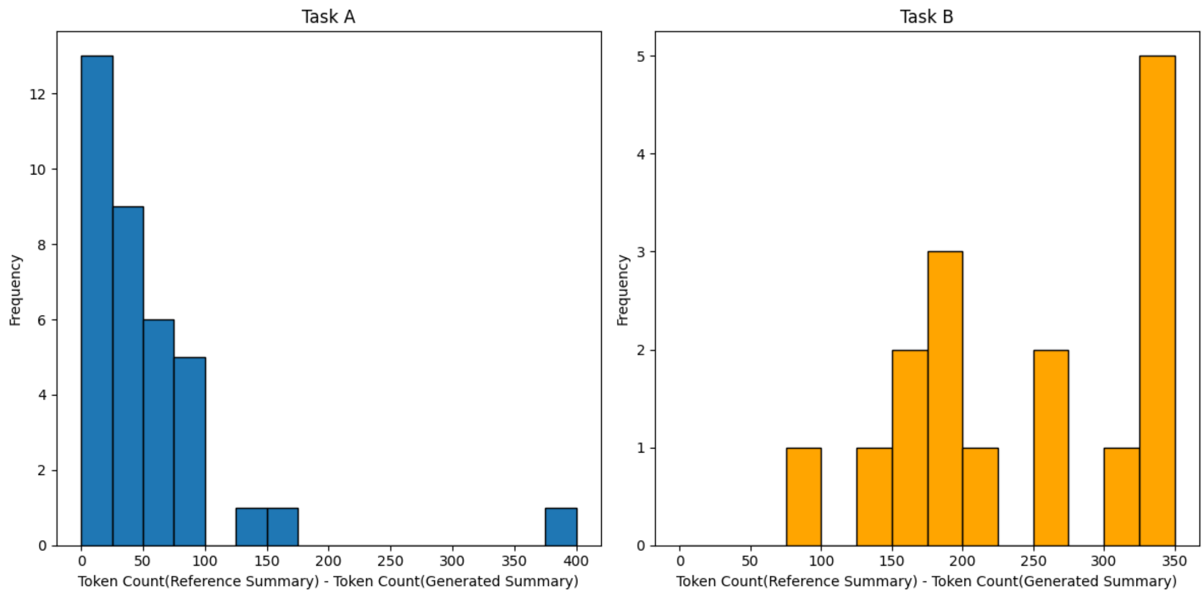


Figure 5: Difference in length of reference summaries and generated summaries

2020) and compression ratio (CR) (Grusky et al., 2020). These metrics rely upon the concept of extractive fragments which are defined as shared sequences in the dialogue and the summary. The extractive fragment coverage quantifies the percentage of words in the summary that are a part of the extractive fragments in the original dialogue. The extractive fragment density measures the average length of the extractive fragment to which a word in the summary belongs to. Compression is measured as the fraction of words in the article and summary.

A comparison of the extractiveness and compression ratio of the reference and generated summaries are shown in Table 3. For Task A, the relatively poor extractive performance of our proposed methods could be due to the smaller size of generated summaries which prevents the usage of more terms from the dialogue. In Task B, we observe the extractive capability of our model improving. This could be attributed to the larger dialogues in Task B allowing for a larger candidate space of tokens to be used in the generations. The generated summaries in Task B are still smaller than the reference summaries as shown by the compression ratio.

5.5 Impact of the number of in-context examples

We further evaluate the impact of the number of in-context examples (k) on various metrics. We report the metrics in Table 5. We observe a general improvement across all metrics as we increase k. This implies that the generated summaries improve as the model is given more access to relevant data through in-context examples. The relevancy in our method is brought in through the selection of in-context examples via semantic similarity/maximum marginal relevancy. This experiment was only performed for Task A since the token limits of the models did not permit the ablation of k for Task B.

6 Future Work and Conclusion

This paper attempts to automatically generate summaries or structured SOAP notes from a conversation between a doctor and a patient. We tackle this problem by generating section-wise summaries, classifying these summaries into appropriate section headers and generating full-length summaries from longer conversations.

We conclude from the results that prompting-based techniques by themselves can not perform optimally themselves but do give some outstanding results when combined with existing techniques, like prompt selection using MiniLM. Next, we also dive deep into where prompting-based methods underperform the standard models like BART and T5.

Overall, our model concluded 3rd amongst all runs submitted and 2nd as a team for task A, which aimed at producing section-level summaries. Our system also stood 4th amongst all runs submitted and 2nd as a team in division-wise summaries for task B. In the future, we plan to use an ensemble of extractive and abstractive methods of generating summaries as well as using diversity algorithms

Task	Runs	**R1**	**R2**	**RL**	**BR**	**BP**	**BF1**	**BL**
A	Run 1	35.29	14.998	27.605	0.893	0.891	0.892	-0.501
	Run 2	34.435	13.984	27.458	0.891	0.889	0.889	-0.525
	Run 3	34.906	14.015	27.525	0.893	0.891	0.892	-0.5
B	Run 1	50.823	24.867	33.581	0.861	0.879	0.87	-0.547
	Run 2	51.232	24.936	33.749	0.862	0.880	0.871	-0.561
	Run 3	50.926	24.845	33.637	0.861	0.878	0.872	-0.553

Table 4: Stability of Validation Results. Metrics include ROUGE-1 (R1), ROUGE-2 (R2), ROUGE-L (RL), BERTScore Precision (BP), Recall (BR), and F1 (BF1), and BLUERT (BL).

k	**R1**	**BP**	**BR**	**BF1**	**BL**
3	40.067	0.904	0.899	0.901	-0.321
5	41.913	0.906	0.905	0.905	-0.305
7	42.841	0.909	0.907	0.907	-0.265

Table 5: Impact of number of in-context examples (k) for Task A (GPT-4)

that will aid in producing SOAP notes that are more robust and apt as per human evaluations.

7 Limitations

Considering the critical nature of the domain of the task, it is of paramount importance to ensure stability in the results expected from the model. Despite setting the temperature (T) as 0 for all decodings in our experiments, we observe the variance in the generated summaries across runs. Table 4 contains the results for three runs for Task A and Task B. The in-context examples for each sample and other parameters have been kept constant across these runs to identify the degree of stochasticity. Further, in-context learning has shown to be susceptible to changes in order of in-context examples (Lu et al., 2021), as well as the template of the examples (Shin et al., 2020). A more reliable process to generate the summaries along with identification of the optimal examples (template, order) is thus required. Additionally, due to the context limit of the GPT-4 model, evaluating the impact of natural language instructions in addition to the examples could not be performed.

8 Ethics Statement

There exist several risks and ethical considerations that necessitate comprehensive addressal prior to the deployment and utilization of our proposed methods utilizing Language Models (LLMs). A significant apprehension associated with employing LLMs for summarization, as evidenced during experimentation, is their susceptibility to hallucination. This means that there would need to be stringent and effective fact-verification post-processing on the generated summaries, thereby ensuring their factual accuracy and alignment with the doctor-patient discourse.

The preservation of patient confidentiality and privacy assumes paramount importance within the context of healthcare data, given its highly sensitive and personal nature. Consequently, it becomes imperative to undertake effective data anonymization techniques to safeguard patient identities. Additionally, obtaining explicit consent from patients regarding the utilization of their data assumes critical significance. In tandem, strict adherence to the standards set forth by the Health Insurance Portability and Accountability Act (HIPAA) is essential to ensure compliance and guarantee the privacy and security of patient information.

Furthermore, another vital aspect that demands careful consideration is the explainability and interpretability when utilizing Language Models (LLMs) for medical summarization. It becomes essential to address the challenge of comprehending the decision-making processes underlying their outputs. Particularly within the medical domain, where critical decisions are made based on these outputs, explainability is of great importance.

References

Deen Abdullah, Shamanth Nayak, Gandharv Suri, and Yllias Chali. 2023. Generating query focused summaries without fine-tuning the transformer-based pretrained models.

Asma Ben Abacha, Wen-wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen. 2023. Overview of the mediqa-chat 2023 shared tasks on the summarization and generation of doctor-patient conversations. In *ACL-ClinicalNLP 2023*.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and

Thomas Lin. 2023. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, Dubrovnik, Croatia. Association for Computational Linguistics.

Amanda Bertsch, Graham Neubig, and Matthew R. Gormley. 2022. He said, she said: Style transfer for shifting the perspective of dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4823–4840, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, and Melanie Subbiah. 2020. Language models are few-shot learners.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.

Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2021. Medically aware GPT-3 as a data generator for medical dialogue summarization. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 66–76, Online. Association for Computational Linguistics.

Sky Corby, Jeffrey A. Gold, Vishnu Mohan, Nicholas Solberg, James Becton, Robby Bergstrom, Benjamin Orwoll, Christopher Hoekstra, and Joan S. Ash. 2020. A sociotechnical multiple perspectives approach to the use of medical scribes: A deeper dive into the scribe-provider interaction.

Yinpei Dai, Hangyu Li, Chengguang Tang, Yongbin Li, Jian Sun, and Xiaodan Zhu. 2020. Learning low-resource end-to-end goal-oriented dialog for fast and reliable system deployment. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 609–618, Online. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen†. 2021. Making pre-trained language models better few-shot learners.

Risha Gidwani, Cathina Nguyen, Alexis Kofoed, Catherine Carragee, Tracy Rydel, Ian Nelligan, Amelia Sattler, Megan Mahoney, and Steven Lin. 2017. Impact of scribes on physician satisfaction, patient satisfaction, and charting efficiency: A randomized controlled trial. *Annals of family medicine*, 15:427–433.

Jade Goldstein and Jaime Carbonell. 1998. Summarization: (1) using MMR for diversity- based reranking and (2) evaluating summaries. In *TIPSTER TEXT PROGRAM PHASE III: Proceedings of a Workshop held at Baltimore, Maryland, October 13-15, 1998*, pages 181–195, Baltimore, Maryland, USA. Association for Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3.

Max Grusky, Mor Naaman, and Yoav Artzi. 2020. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies.

Raghav Jain, Anubhav Jangra, Sriparna Saha, and Adam Jatowt. 2022. A survey on medical document summarization.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023. Mimic-iv-note: Deidentified free-text clinical notes (version 2.2).

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020a. Dr. summarize: Global summarization of medical dialogue by exploiting local structures.

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020b. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3755–3763, Online. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Shuai Lin, Pan Zhou, Xiaodan Liang, Jianheng Tang, Ruihui Zhao, Ziliang Chen, and Liang Lin. 2020. Graph-evolving meta-learning for low-resource medical dialogue generation.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jessica López Espejel. 2019. Automatic summarization of medical conversations, a review. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume*

III : RECITAL, pages 487–498, Toulouse, France. ATALA.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Yuning Mao, Yanru Qu, Yiqing Xie, Xiang Ren, and Jiawei Han. 2020. Multi-document summarization with maximal marginal relevance-guided reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1737–1751, Online. Association for Computational Linguistics.

George Michalopoulos, Kyle Williams, Gagandeep Singh, and Thomas Lin. 2022. MedicalSum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4741–4749, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pranita Mishra, Jacqueline C. Kiang, and Richard W. Grant. 2018. Association of medical scribes in primary care with physician workflow and patient experience. *JAMA internal medicine*, 178:1467–1472.

Nithesh Naik, B. M. Zeeshan Hameed, Dasharathraj K. Shetty, Dishant Swain, Milap Shah, Rahul Paul, Kaivalya Aggarwal, Sufyan Ibrahim, Vathsala Patil, Komal Smriti, Suyog Shetty, Bhavan Prasad Rai, Piotr Chlosta, and Bhaskar K. Somani. 2022. Legal and ethical consideration in artificial intelligence in healthcare: Who takes responsibility? *Frontiers in surgery*, 9.

David Fraile Navarro, Mark Dras, and Shlomo Berkovsky. 2022. Few-shot fine-tuning SOTA summarization models for medical dialogues. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 254–266, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems.

OpenAI. 2023. Gpt-4 technical report.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models.

Vivek Podder, Valerie Lew, and Sassan Ghassemzadeh. 2020. Soap notes.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Sandeep Reddy. 2022. Explainability and artificial intelligence in medicine. the lancet. digital health.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A. Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin Rousseau, Chunhua Weng, and Yifan Peng. 2023. Evaluating large language models on medical evidence summarization. *medRxiv*.

Kristen Ullman, Lauren McKenzie, Bradley Bart, Glennon Park, Roderick MacDonald, Eric Linskens, and Timothy J. Wilt. 2021. The effect of medical scribes in emergency departments: A systematic review. *The Journal of Emergency Medicine*, 61:19–28.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.

Xi Ye, Srinivasan Iyer, Asli Celikyilmaz, Ves Stoyanov, Greg Durrett, and Ramakanth Pasunuru. 2022. Complementary explanations for effective in-context learning. *arXiv preprint arXiv:2211.13892*.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Submitted to Nature Scientific Data*.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models.

Chunting Zhou, Junxian He, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Prompt consistency for zero-shot task generalization.

A Appendix

This appendix presents two tables - Table 6 contains the categories and subcategories in which the dialogue is divided to create a SOAP note. Table 7 presents the prompts used by approaches for tasks A and B.

Table 6: Categorization Scheme

HISTORY OF PRESENT ILLNESS
Fam/Sochx [Family History/Social History]
Genhx [History of Present Illness]
Pastmedicalhx [Past Medical History]
CC [Chief Complain]
Pastsurgical [Past Surgical History]
Allergy
Gynhx [Gynecologic History]
Other_history
Immunizations
Medications
PHYSICAL EXAM
ROS [Review of Systems]
Exam
RESULTS
Imaging
Procedures
Labs
ASSESSMENT AND PLAN
Assessment
Diagnosis
Plan
Edcourse [Emergency Department Course]
Disposition
Medications

Table 7: Prompt Templates

Prompting Approach	Model	Prompt(Example)
Zero-Shot	text-davinci-003	"Summarize the following into a medical report having the following sections: ’HISTORY OF PRESENT ILLNESS’, ’PHYSICAL EXAM’, ’RESULTS’, ’ASSESSMENT AND PLAN’.
Few-shot prompting for section-wise summary	text-davinci-003	Prompt for PHYSICAL EXAM section(k=5) """ Dialogue: Doctor: Breath in breath out, let me tap it and see. Well, your lungs sound clear. Patient: Okay. Summary: CHEST: Lungs bilaterally clear to auscultation and percussion. . . . . Dialogue : Doctor: Do you have any chest pain? Patient: No, I don’t. Doctor: Any breathlessness? Patient: Yes, I do get breathless only when I have to do some form of exertion like walking a long time or running. Doctor: Okay. How about any bowel issues? Patient: No, I don’t have any stomach problems except I have to go frequently to use the bathroom. Doctor: Okay frequency. How about any prolonged bleeding issues or anything like that sort? Patient: No nothing like that. Summary : He denies any chest pain. He admits to exertional shortness of breath. He denies any GI problems as noted. Has frequent urination as noted. He denies any bleeding disorders or bleeding history. Dialogue : {dialogue} Summary : """
Perspective Shift	text-davinci-003 gpt3.5-turbo	2 staged prompting (perspective shift with turbo and summarization with davinci) PERSPECTIVE SHIFT = """ Convert the following into third person. {dialogue} \\\ """ PROMPT = """ Summarize the following into a medical report having the following sections: "HISTORY OF PRESENT ILLNESS", "PHYSICAL EXAM", "RESULTS", "ASSESSMENT AND PLAN" where each section is at least 60 words. {third-person-perspective}\\\ """
Two-Stage Prompting	text-davinci-003	PROMPT #1 """ {dialogue} \\ List the important points from the above conversation for a medical report """ PROMPT #2 """ {prompt1-generated-output} \\ Create a paragraph from the above facts only """

Prompting Approach	Model	Prompt(Example)
Prompt Selection - MMR(k=3)	text-davinci-003	PROMPT SELECTION with k=3 Dialogue: Doctor: Your last visit was on April seventh two thousand five, correct? Patient: Ah no, it was on April eighth two thousand five, doctor. Doctor: That's right. So, has anything changed since then? Patient: No, everything is the same really. Summary: Essentially unchanged from my visit of 04/08/2005. . . . . Dialogue: Doctor: Do you have any past or present medical conditions? Patient: No. Summary: None. . . . . Dialogue: {dialogue} Summary:
Prompt Selection - Semantic Similarity(k=7) - Task A	**GPT-4**	PROMPT SELECTION with k=7 Dialogue: Doctor: Do you know about any medical issues running in your family? Patient: Yeah, almost everyone had diabetes. Summary: Multiple family members have diabetes mellitus. . . . . Dialogue: Doctor: Any specific family medical history that I should be aware of? Patient: No. Doctor: Anyone in your family, even grandparents, if you know them, did they have diabetes or high blood pressure? Patient: No. Doctor: Anyone else sick at home? Patient: No. Summary: Noncontributory. No one else at home is sick. . . . . Dialogue: {dialogue} Summary:
Prompt Selection - Semantic Similarity(k=1) Task B	**GPT-4**	PROMPT SELECTION with k=1 Dialogue: [doctor] and why is she here ? annual exam. okay. all right. hi, Sarah. how are you ? [patient] good . how are you ? [doctor] i'm good . are you ready to get started ? [patient] yes , i am . [doctor] okay . so Sarah is a 27-year-old female here for her annual visit. So, Sarah, how have you been since the last time I saw you ? . . . . Summary: CHIEF COMPLAINT Annual visit. HISTORY OF PRESENT ILLNESS The patient is a 27-year-old female who presents for her annual visit. She reports that she has been struggling with her depression off and on for the past year...... . . . . Dialogue: {dialogue} Summary:

Overview of the MEDIQA-Chat 2023 Shared Tasks on the Summarization & Generation of Doctor-Patient Conversations

Asma Ben Abacha
Microsoft Health AI, USA
abenabacha@microsoft.com

Wen-wai Yim
Microsoft Health AI, USA
yimwenwai@microsoft.com

Griffin Adams
Columbia University, USA
griffin.adams@columbia.edu

Neal Snider
Microsoft/Nuance, USA
neal.snider@nuance.com

Meliha Yetisgen
University of Washington, USA
melihay@uw.edu

Abstract

Automatic generation of clinical notes from doctor-patient conversations can play a key role in reducing daily doctors' workload and improving their interactions with the patients. MEDIQA-Chat 2023 aims to advance and promote research on effective solutions through shared tasks on the automatic summarization of doctor-patient conversations and on the generation of synthetic dialogues from clinical notes for data augmentation. Seventeen teams participated in the challenge and experimented with a broad range of approaches and models. In this paper, we describe the three MEDIQA-Chat 2023 tasks, the datasets, and the participants' results and methods. We hope that these shared tasks will lead to additional research efforts and insights on the automatic generation and evaluation of clinical notes.

1 Introduction

Recent progress in text summarization and generative AI can greatly benefit the healthcare system by automatically generating clinical notes from doctor-patient conversations. This can contribute to effective clinical care by reducing the doctors' workload to editing and validating the generated summaries/notes instead of writing the full notes during the consultations at the expense of their time or focus when talking and interacting with the patients.

Clinical note generation has seen an increased research interest in the recent years. For instance, (Yim and Yetisgen, 2021) tackled automatic medical scribing with Dialogue2Note sentence alignment and snippet summarization. (Michalopoulos et al., 2022) introduced MedicalSum, a guided clinical abstractive summarization model for generating medical reports from doctor-patient conversations. (Grambow et al., 2022) showed that in-domain pre-training improves clinical note generation from doctor-patient conversations. (Knoll et al., 2022) presented three user studies, on medical note generation systems and analyzed the clinicians' views of how the system could be adapted and improved. Other efforts focused on the evaluation of medical note generation manually through consultation checklists (Savkov et al., 2022) or automatically using evaluation metrics that correlate with human judgments (Moramarco et al., 2022; Adams et al., 2023; Ben Abacha et al., 2023b). (Papadopoulos Korfiatis et al., 2022) introduced the primock57 collection of 57 mocked primary care consultations, one of the rare datasets dedicated to this task.

The previous editions of the MEDIQA shared tasks focused on medical NLP tasks such as textual inference and question answering (Ben Abacha et al., 2019) as well as the summarization of patient questions/answers and radiology reports (Ben Abacha et al., 2021). This third edition, MEDIQA-Chat 2023[1], addresses the generation of clinical notes based on the summarization of doctor-patient conversations. All of the datasets and code created for this challenge are publicly available[2].

In this paper, we present the tasks and datasets in section 2 and section 3. In section 4, we present the evaluation methods and metrics used for the shared tasks. Section 5 describes and discusses the participating teams' approaches and draws insights from the official challenge results.

2 Tasks

2.1 Task A - Short Dialogue2Note Summarization

The first task focuses on summarizing short doctor-patient conversations to generate a summary for only one section of a clinical note, including a section header, as described in Figure 1.

[1] https://sites.google.com/view/mediqa2023/clinicalnlp-mediqa-chat-2023
[2] https://github.com/abachaa/MEDIQA-Chat-2023

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 503–513
July 14, 2023

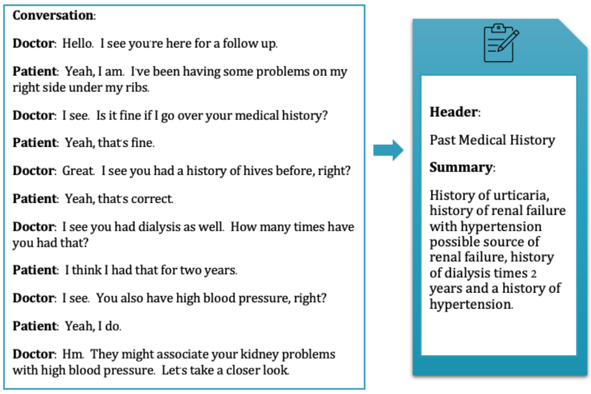


Figure 1: Task A: summarize a short doctor-patient conversation to generate a note section with the associated section header (example from the MTS-Dialog dataset).

The section header is one of the following 20 headers: Family History/Social History (fam/sochx), History of Present Illness (genhx), Past Medical History (pastmedicalhx), Chief Complaint [cc], Past Surgical History (pastsurgical), allergy, Review of Systems (ros), medications, assessment, exam, diagnosis, disposition, plan, Emergency Department Course (edcourse), immunizations, imaging, Gynecologic History (gynhx), procedures, other_history, and labs.

2.2 Task B - Full Dialogue2Note Summarization

The goal of task B is to generate a complete note for each doctor-patient encounter, as described in Figure 2. The note must include all relevant sections. As the same section can have different correct expressions for its header, we defined four main section/division categories, each associated with several correct labels/expressions for its header. The section category-header mappings are presented in table 1.

Division/Category	Possible Section Headers
Subjective	Chief Complaint, HPI, History of Present Illness, Subjective
Objective_Exam	Physical Exam, Exam
Objective_Results	Results, Findings
Assessment&Plan	Assessment, Plan

Table 1: Task B: Note Divisions and Section Headers

Full-encounter notes are expected to have at most one section from each category. If a generated note contains multiple sections from the same category, only the first occurring section of that category is used for evaluation. Also, depending on the encounter, Objective_Exam and Objective_Results may not be relevant.

2.3 Task C - Note2Dialogue Generation

This task addresses data augmentation through the generation of synthetic doctor-patient conversations from full clinical notes. We encouraged the participants to apply the models developed for this task to generate additional data for tasks A and B.

3 Datasets

Table 3 describes the training, validation, and test sets created from the MTS-Dialog (Ben Abacha et al., 2023a) and ACI-Bench (Yim et al., 2023) collections.

The MTS-Dialog dataset, used in Task A, consists of 1.7k pairs of conversations and associated summaries. Table 2 presents examples from MTS-Dialog conversations and summaries.

The ACI-Bench dataset, used Tasks B & C consists of 207 pairs of full doctor-patient conversations and associated clinical notes.

4 Evaluation

In this challenge, we evaluated both the submitted runs and the submitted codes as described below.

4.1 Evaluation Metrics

We selected three automatic metrics that highly correlate with human judgments for the task of clinical note generation based on recent studies (Ben Abacha et al., 2023a,b) on the evaluation methods for the summarization of doctor-patient conversations. These metrics are: ROUGE-1 (Lin, 2004), BLEURT (Sellam et al., 2020), and BERTScore (Zhang et al., 2020).

We used the average score from ROUGE-1, BLEURT-20, and BERTScore (microsoft/deberta-xlarge-mnli) as the main score to rank the participating systems in short note generation (Aggregate-Score).

For full note generation, we relied on ROUGE-1 for the evaluation of full notes as BLEURT and BERTScore have a maximum sequence length of 512 tokens. For these notes, we also performed a more fine-grained sub-note section-level evaluation using the average score of the three metrics.

In summary, we used the following evaluation metrics for each task:

- Task A - Evaluating the section header classification using Accuracy.

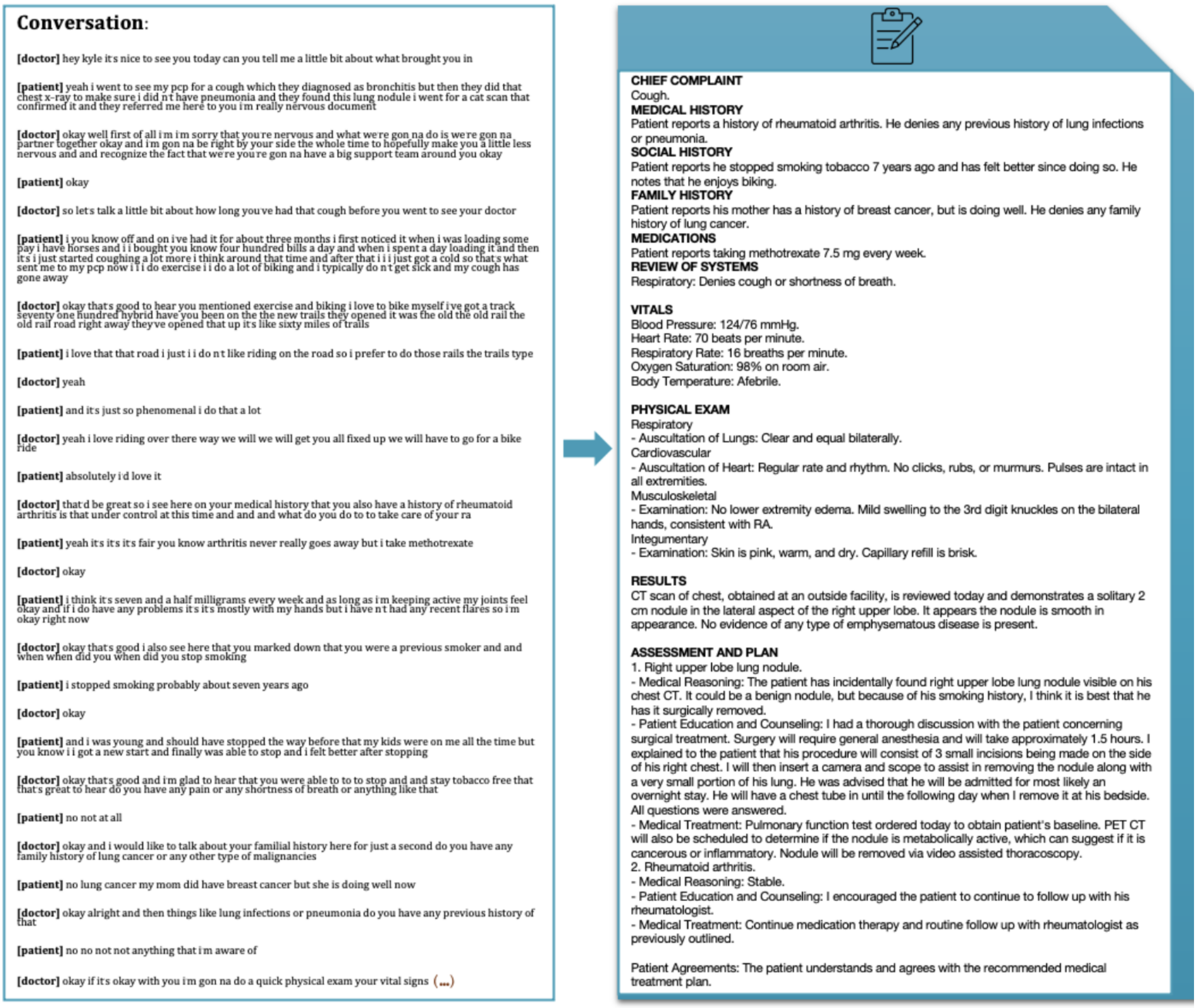


Figure 2: Task B: summarize each doctor-patient conversation to generate a full note with all relevant sections (example from the ACI-Bench dataset).

- Task A - Evaluating the short summaries using the average score of ROUGE-1, BERTScore, and BLEURT.
- Task B - Evaluating the long summaries/notes with two different methods: (i) Full-note evaluation using ROUGE-1 and (ii) a fine-grained evaluation taking the mean of the section-based combined score of ROUGE-1, BERTscore, BLEURT, equally weighed.
- Task C - Evaluating the generated dialogues using ROUGE-1.

4.2 Code Verification

The participants shared their private codes with the organizers on GitHub following the provided code preparation instructions [3].

[3]Evaluation instructions and scripts available at https://github.com/abachaa/MEDIQA-Chat-2023

We defined five code statuses to label each team's code (cf. Results Section):

1. Code runs and exactly reproduces
2. Code runs with minor differences
3. Results unstable due to non-deterministic components (e.g., generative API calls)
4. Results unstable
5. Code does not run under our configurations

We provided feedback on the shared codes and their outputs/errors to the participants.

4.3 Baseline Models

We used the latest OpenAI models to prepare baseline models using ChatGPT (gpt-3.5-turbo) and GPT-4. We used a temperature of 1 for tasks A

Section Header	Conversation	Summary
MEDICATIONS	Doctor: Are you still taking the Trizivir? Patient: Yes. Doctor: How much are you taking? Patient: I take one pill two times a day. Doctor: Are you taking any other medications? Patient: I take Ibuprofen for body aches from time to time but that's it.	1. She is on Trizivir 1 tablet p.o. b.i.d. 2. Ibuprofen over-the-counter p.r.n.
ROS	Doctor: Have you had any anxiety attacks lately? Patient: No. Doctor: Have you felt depressed or had any mood swing problems? Patient: No. Doctor: Any phobias? Patient: No, not really. Doctor: Okay.	PSYCHIATRIC: Normal; Negative for anxiety, depression, or phobias.
FAM/SOCHX	Doctor: Are you still working? Patient: No, I am retired now. I used to work for the U S postal service as an electronic technician but took retirement one year earlier due to my disability. Doctor: Ah okay. And who is in your family? Patient: Well, I stay with my wife and daughter in our apartment. Doctor: Okay. Do you smoke? Patient: No. Doctor: How about alcohol? Patient: I use to drink occasionally, that too very rare, but after my symptoms stated I stopped completely. Doctor: Any use of recreational or illegal drugs? Patient: Nope. Doctor: Did you travel anywhere recently? Patient: No, it's been really long since I traveled anywhere.	The patient retired one year PTA due to his disability. He was formerly employed as an electronic technician for the US postal service. The patient lives with his wife and daughter in an apartment. He denied any smoking history. He used to drink alcohol rarely but stopped entirely with the onset of his symptoms. He denied any h/o drug abuse. He denied any recent travel history.
GENHX	Doctor: Sir? Can you hear me? Doctor: Are you Mister Smith's wife? Guest_family: Yes. I am his wife. Doctor: How old is he? Can you tell me a little bit of how your husband's condition has come to this point? His level of consciousness is concerning. Guest_family: He is eighty five. He took the entire M G of Xanax. He is only supposed to take point one twenty five M G of Xanax. That is why he is like this. Doctor: It looks like your husband was admitted to the emergency room the night before. How did these injuries to his face happen? Guest_family: He fell off his wheelchair. Doctor: The Adult Protective Services said they found your husband in the home barley conscious. How long had he been that way? Guest_family: All day. Doctor: Do you know what other medications your husband has taken other than the Xanax? Guest_family: He didn't take his regular medications for two days.	The patient is an 85-year-old male who was brought in by EMS with a complaint of a decreased level of consciousness. The patient apparently lives with his wife and was found to have a decreased status since the last one day. The patient actually was seen in the emergency room the night before for injuries of the face and for possible elderly abuse. When the Adult Protective Services actually went to the patient's house, he was found to be having decreased consciousness for a whole day by his wife. Actually the night before, he fell off his wheelchair and had lacerations on the face. As per his wife, she states that the patient was given an entire mg of Xanax rather than 0.125 mg of Xanax, and that is why he has had decreased mental status since then. The patient's wife is not able to give a history. The patient has not been getting Sinemet and his other home medications in the last 2 days.

Table 2: Examples of conversations and associated section headers and summaries from the MTS-Dialog dataset.

Task	Dataset	Training	Validation	Test
A	MTS-Dialog	1,201	100	200
B	ACI-Bench	67	20	40
C	ACI-Bench	67	20	40

Table 3: Training, Validation, and Test Sets (# pairs)

and B. For task C, we experimented with two temperatures for more variety in the generated conversations with deterministic (temperature=0) and creative (temperature=1) outputs. ChatGPT has a limit of 4,097 tokens, shared between the prompt and the output/summary, whereas GPT-4 allows 32k tokens.

We ran the baseline models on an Nvidia Tesla K80 GPU.

We used the following prompt for tasks A, B, and C:

- **Prompt for Task A**: "Classify the conversation into one of these 20 classes: FAMILY HISTORY/SOCIAL HISTORY, HISTORY of PRESENT ILLNESS, PAST MEDICAL HISTORY, CHIEF COMPLAINT, PAST SURGICAL HISTORY, Allergy, REVIEW OF SYSTEMS, Medications, Assessment, Exam, Diagnosis, Disposition, Plan, EMERGENCY DEPARTMENT COURSE, Immunizations, Imaging, GYNECOLOGIC HISTORY, Procedures, Other history, Labs. The response should start with the selected class, followed by # then the summary of the conversation in a clinical note style. The conversation is: "

- We then extracted the section headers and summaries from the outputs.

- **Prompt for Task B**: "Summarize the conversation to generate a clinical note with four sections: HISTORY OF PRESENT ILLNESS, PHYSICAL EXAM, RESULTS, ASSESSMENT AND PLAN. The conversation is: "

- To allow adequate division detection, we added some light rule-based post-processing for Task B outputs.

- **Prompt for Task C**: "write a full conversation between a doctor and a patient during a medical visit. The dialogue should cover all the medical information provided in this note: "

	Team	Affiliation	Tasks	Paper	Code
1	WangLab	University of Toronto, Canada	A, B	(Giorgi et al., 2023)	1
2	SummQA	Carnegie Mellon University, USA	A, B	(Mathur et al., 2023)	2
3	Cadence	Cadence Solutions, USA	A, B, C	(Sharma et al., 2023)	3
4	GersteinLab	Yale University, USA	A, B	(Tang et al., 2023)	4
5	NewAgeHealthWarriors	IIITB, India	A	(Mishra and Desetty, 2023)	5
6	NUS-IDS	NUS, Singapore	A, C	-	6
7	HuskyScribe	University of Washington, USA	A, B	-	7
8	Calvados	Université de Caen Normandie, France	A, B	(Milintsevich and Agarwal, 2023)	8
9	DS4DH	University of Geneva, Switzerland	A	(Zhang et al., 2023)	9
10	UMASS_BioNLP	University of Massachusetts, USA	A, B, C	(Wang et al., 2023)	10
11	HealthMavericks	University of Mumbai, India	A, B	(Suri et al., 2023)	11
12	Care4lang	George Washington University, USA	A	(Alqahtani et al., 2023)	12
13	clulab	University of Arizona, USA	A	(Ozler and Bethard, 2023)	13
14	DFKI-MedIML	German Research Center for AI, Germany	A, B	-	14
15	iuteam1	Indiana University, USA	B	(Srivastava, 2023)	15
16	SZU_Clinical	Shenzhen University, China	B	-	16
17	Teddysum	Kyungpook University, South Korea	B	(Jeong et al., 2023)	17

[1] github.com/bowang-lab/MEDIQA-Chat-2023-WangLab
[2] github.com/Raghav1606/SummQA
[3] github.com/ashwyn/MEDIQA-Chat-2023-Cadence
[4] github.com/28andrew/MEDIQA-Chat-2023-GersteinLab
[5] github.com/prakhar21/MEDIQA-CHAT-2023-NewAgeHealthWarriors
[6] github.com/Elfsong/MEDIQA-Chat-2023-NUS-IDS
[7] github.com/BeanHam/MEDIQA-Chat-2023-HuskyScribe
[8] github.com/501Good/MEDIQA-Chat-2023-Calvados
[9] github.com/tinaboya/MEDIQA-Chat-2023-ds4dh
[10] github.com/believewhat/MEDIQA-Chat-2023-UMASS_BioNLP
[11] github.com/suri-kunal/MEDIQA-Chat-2023-HealthMavericks
[12] github.com/amalqahtani/Clinical-NLP-Models
[13] github.com/kbulutozler/MEDIQA-Chat-2023-clulab
[14] github.com/sitingGZ/MEDIQA-Chat-2023-DFKI-MedIML
[15] github.com/dhananjay-srivastava/MEDIQA-Chat-2023-iuteam1
[16] github.com/SunnyLee216/MEDIQA-Chat-2023-SZU_Clinical
[17] github.com/teddysum/MEDIQA-Chat-2023-Teddysum

Table 4: MEDIQA-Chat 2023: Participating teams, number of runs (with a limit of three runs/task), submitted codes, and working notes papers.

5 Official Results

5.1 Participating Teams

The MEDIQA-Chat shared tasks attracted 120 registered teams from academy and industry. Among them, 17 teams submitted their codes and runs following the challenge rules. Table 4 presents the teams that participated in the three shared tasks. We limited the number of submitted runs to three runs per task.

5.2 Task A: Approaches & Results

Task A includes two subtasks on (i) generating the summary of a short medical conversation and (ii) classifying the sections/summaries using a predefined list of section headers. Fourteen teams participated in Task A. Table 5 presents the results of the section classification subtask and Table 6 presents the results of the summarization subtask.

In task A, most teams used fine-tuned models (e.g., BART, T5) and/or OpenAI-based solutions in the summarization subtask and leveraged fine-tuned BERT or RoBERTa-based models for section classification. The WangLab team (Giorgi et al., 2023) achieved the best results in the summarization subtask with 0.5789 Aggregate-Score and the best Accuracy of 0.78 in the header classification subtask using a Flan-T5 model that jointly generates the section header and content. The NUS-IDS team also achieved the best Accuracy of 0.78 in header classification and 0.5204 Aggregate-Score in summarization using a T5 model fine-tuned on data augmented by GPT-3. The HuskyScribe team also used a T5-based model (T5-Large and Clinical-T5-Large) trained in a question-answering format for section header classification. Their summarizer consisted of a BART-large-xsum model fine-tuned on task A's training data, the Samsum dataset (Gliwa et al., 2019), and the Dialogue-sum dataset (Chen et al., 2021). Care4Lang (Alqahtani et al., 2023) used a Flan-T5 model fine-tuned on the training data with a pre-processed input combining the header and the dialogue for implicit header learning and conditional summary generation. Clinical-T5-

Team	Run#	Accuracy	Rank	Code Status
NUS-IDS	run1	**0.780**	1	1
WangLab	run2	**0.780**	1	1
WangLab	run3	0.770	3	1
HuskyScribe	run1	0.755	4	2
WangLab	run1	0.750	5	1
gersteinlab	run2	0.745	6	1
Cadence	run1	0.735	7	1
NewAgeHealthWarriors	run1	0.730	8	5
DFKI-MedIML	run2	0.725	9	1
DFKI-MedIML	run3	0.725	9	1
DFKI-MedIML	run1	0.725	9	1
HealthMavericks	run2	0.725	9	5
HealthMavericks	run3	0.725	9	5
HealthMavericks	run1	0.725	9	5
gersteinlab	run1	0.710	15	3
SummQA	run2	0.710	15	3
SummQA	run1	0.710	15	3
NewAgeHealthWarriors	run2	0.705	18	2
UMASS_BioNLP	run1	0.705	18	5
DS4DH	run2	0.700	20	5
DS4DH	run1	0.700	20	1
gersteinlab	run3	0.700	20	3
Calvados	run2	0.685	23	1
Calvados	run1	0.680	24	1
Calvados	run3	0.640	25	1
Care4lang	run3	0.565	26	1
clulab	run2	0.540	27	1
clulab	run1	0.540	27	1
Care4Lang	run1	0.375	29	1
UMASS_BioNLP	run2	0.355	30	5
Care4Lang	run2	0.345	31	1
Baseline1	ChatGPT	0.500	-	1
Baseline2	GPT-4	0.530	-	1

Table 5: Official Results of MEDIQA-Chat Task A - Header Classification (1/2)

Sci models were used by the clulab team (Ozler and Bethard, 2023) to generate three different summaries for each dialogue to augment the header classification training data, and then used a Roberta-based model trained on the augmented dataset to predict the header based on the summary of the dialogue instead of the dialogue itself. The Calvados team (Milintsevich and Agarwal, 2023) used a LongT5 model for summarization and clinical NER model to extract disease and treatment mentions that are then tagged in the input conversation and the output summary. They combined the classification label and the summary note into a single output, and considered the classification as a subtask within summary generation.

The SummQA team (Mathur et al., 2023) utilized an ensemble of BioClinicalBERT and GPT-4 for section header classification. GPT-4 was used as a zero-shot classifier and BioClinicalBERT was fine-tuned on the task A training data. Their summarization method relied on GPT-4 with prompt selection based on semantic similarity to retrieve top-k (k=7) examples for in-context learning and was ranked third in TaskA-Summarization with 0.5739 Aggregate-Score. The DS4DH team (Zhang et al., 2023) used a classification model (tf-idf-svm) in combination with ChatGPT (run1) or GPT-3 Curie (run2) for summarization. The UMASS-BioNLP team (Wang et al., 2023) also used ChatGPT to jointly generate the section header and note.

The Cadence team (Sharma et al., 2023) adapted a BART-large model for classification and summarization. The summarizer was a BART-large model fine-tuned first on the Samsum dataset and second on Task A data augmented with 1k note samples extracted from MIMIC-IV and their dialogues

Team	Run#	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-LSum	BERTScore	BLEURT	Agg-Score	Agg-Rank	Code Status
WangLab	run2	**0.4466**	**0.2282**	**0.3837**	**0.3837**	**0.7307**	0.5593	**0.5789**	1	1
WangLab	run3	0.4396	0.1999	0.3781	0.3781	0.7260	0.5570	0.5742	2	1
SummQA	run1	0.4216	0.2017	0.3478	0.3478	0.7247	**0.5753**	0.5739	3	3
Cadence	run1	0.4303	0.2078	0.3642	0.3642	0.7187	0.5377	0.5622	4	1
WangLab	run1	0.4160	0.2003	0.3512	0.3512	0.7203	0.5464	0.5609	5	1
SummQA	run2	0.4056	0.1920	0.3317	0.3317	0.7030	0.5666	0.5584	6	3
gersteinlab	run3	0.4011	0.2147	0.3322	0.3322	0.7058	0.5421	0.5497	7	1
NewAgeHealthWarriors	run1	0.3983	0.1717	0.3314	0.3313	0.6982	0.5350	0.5438	8	5
UMASS_BioNLP	run2	0.3828	0.1828	0.3158	0.3166	0.7015	0.5405	0.5416	9	5
gersteinlab	run1	0.3882	0.1966	0.3214	0.3214	0.700	0.5294	0.5392	10	1
gersteinlab	run2	0.3882	0.1966	0.3214	0.3214	0.700	0.5294	0.5392	10	1
NewAgeHealthWarriors	run2	0.3780	0.1707	0.3134	0.3134	0.6926	0.5303	0.5336	12	2
Calvados	run1	0.3946	0.1864	0.3321	0.3321	0.6999	0.4724	0.5223	13	1
NUS-IDS	run1	0.3511	0.1538	0.2843	0.2843	0.6689	0.5411	0.5204	14	1
HuskyScribe	run1	0.3689	0.1820	0.3072	0.3072	0.6837	0.5006	0.5177	15	1
Care4Lang	run1	0.3581	0.1650	0.2890	0.2890	0.6789	0.5143	0.5171	16	1
Care4Lang	run2	0.3447	0.1553	0.2808	0.2808	0.6726	0.5085	0.5086	17	2
Calvados	run3	0.3569	0.1598	0.2896	0.2896	0.6721	0.4698	0.4996	18	1
DS4DH	run1	0.3080	0.1197	0.2424	0.2424	0.6644	0.5206	0.4977	19	3
clulab	run1	0.3414	0.1379	0.2842	0.2842	0.6569	0.4876	0.4953	20	1
clulab	run2	0.3414	0.1379	0.2842	0.2842	0.6569	0.4876	0.4953	20	1
Calvados	run2	0.3604	0.1617	0.3057	0.3057	0.6779	0.4449	0.4944	22	1
Care4lang	run3	0.3322	0.1400	0.2830	0.2830	0.6582	0.4856	0.4920	23	2
UMASS_BioNLP	run1	0.3283	0.1351	0.2743	0.2743	0.6699	0.4757	0.4913	24	5
HealthMavericks	run2	0.2973	0.1357	0.2200	0.2200	0.6120	0.4956	0.4683	25	5
HealthMavericks	run3	0.2514	0.1011	0.2002	0.2002	0.6268	0.5015	0.4599	26	5
DS4DH	run2	0.2937	0.1091	0.2135	0.2135	0.6179	0.3887	0.4334	27	5
HealthMavericks	run1	0.1987	0.0867	0.1560	0.1560	0.5703	0.4298	0.3996	28	5
DFKI-MedIML	run3	0.1931	0.0771	0.1784	0.1784	0.5758	0.3700	0.3796	29	1
DFKI-MedIML	run2	0.1818	0.0727	0.1707	0.1707	0.5656	0.363	0.3701	30	1
DFKI-MedIML	run1	0.1762	0.0656	0.1641	0.1641	0.5612	0.3664	0.3679	31	1
Baseline1	ChatGPT	0.3032	0.1209	0.2420	0.2420	0.6597	0.5032	0.4887	-	1
Baseline2	GPT-4	0.3071	0.1283	0.2365	0.2365	0.6484	0.5292	0.4949	-	1

Table 6: Official Results of MEDIQA-Chat Task A - Summarization (2/2)

generated by their Task C model. The NewAgeHealthWarriors team (Mishra and Desetty, 2023) also used a fine-tuned BART-large, BioBART-large and calls to GPT-3 API with custom prompt design, followed by an ensemble module to choose the best summary from the previous summarization models. A fine-tuned Bio-ClinicalBERT followed by a Keyword-based categorizer were used for section header classification. The DFKI-MedIML team used a fine-tuned microsoft/biogpt model for generating the section header and section summary. They modified the original BioGptForCausalLM model to encode a list of context input sequences for generating one target output. The HealthMavericks team (Suri et al., 2023) used an ensemble of BioBart-V2, DialogLM-LED-Base, DialogLED-Large, Flan-T5 fine-tuned on the training data (runs 1&2) and GPT-3 with the input dialogue and three randomly sampled dialogue-section-header-summary triplets as prompt.

5.3 Task B: Approaches & Results

Nine teams participated in Task B. We present the results of the full-note evaluation in Table 7 and the section-level evaluation in Table 8.

The WangLab team (Giorgi et al., 2023) used GPT-4 with in-context examples retrieved from the training set based on their similarity to the test dialogues and included their summaries/notes as in-context examples and obtained the best ROUGE-1 score of 0.6141 in full-note evaluation and an Aggregate-Score of 0.6483 in section-based evaluation. SummQA (Mathur et al., 2023) used one-shot GPT-4 with dynamic prompts that include selected examples for in-context learning. The examples consist of dialogue-summary pairs selected from the Task B training data based on semantic similarity and obtained 0.5541 Aggregate-Score. Several teams also used OpenAI-based solutions: GersteinLab (Tang et al., 2023) used the Davinci model, UMASS_BioNLP (Wang et al., 2023) used GPT-4, ad healthmavericks (Suri et al., 2023) used GPT-3 to generate the summaries/clinical notes with static prompts.

The iuteam1 team (Srivastava, 2023) used three different LSG BART models to summarize long conversations using Local, Sparse, and Global Attention mechanisms and evaluated the use of multi-layer structures where multiple summarization model outputs are recombined in a single model to produce more coherent texts. The Cadence team (Sharma et al., 2023) adapted their task A method to task B data, and developed a two-pass summarization approach to manage longer inputs. They fine-tuned BART on the Samsum dataset, Task A and Task B training data, and on additional examples generated from MIMIC-IV notes using their Task C solution.

The GersteinLab team (Tang et al., 2023) used a fine-tuned GPT-3 model for summarization trained

Team	Run #	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-LSum	Rank	Code Status
WangLab	run3	**0.6141**	**0.3288**	0.3815	**0.5515**	1	1
WangLab	run1	0.5851	0.3210	**0.4063**	0.5480	2	4
WangLab	run2	0.5814	0.3213	0.4023	0.5439	3	4
Teddysum	run1	0.5332	0.2511	0.2833	0.4708	4	5
HealthMavericks	run1	0.5311	0.2335	0.2803	0.4523	5	5
Cadence	run2	0.5297	0.2500	0.2979	0.4663	6	2
iuteam1	run2	0.5268	0.2622	0.3060	0.4976	7	1
SZU_Clinical	run1	0.5235	0.2656	0.3330	0.4624	8	5
SZU_Clinical	run2	0.5230	0.2655	0.3327	0.4619	9	5
SZU_Clinical	run3	0.5227	0.2654	0.3325	0.4617	10	5
HealthMavericks	run3	0.5111	0.2122	0.2663	0.4359	11	5
gersteinlab	run2	0.5008	0.2506	0.3282	0.4668	12	3
gersteinlab	run1	0.5004	0.2502	0.3249	0.4675	13	3
Cadence	run1	0.4950	0.2343	0.2810	0.4313	14	1
SummQA	run1	0.4935	0.2319	0.3190	0.4507	15	4
iuteam1	run1	0.4917	0.2239	0.2545	0.4249	16	1
Teddysum	run3	0.4427	0.227	0.2024	0.4125	17	5
Calvados	run2	0.4307	0.2017	0.2394	0.3861	18	1
Teddysum	run2	0.4289	0.2077	0.2485	0.3625	19	5
Calvados	run1	0.4137	0.1967	0.2432	0.3692	20	1
iuteam1	run3	0.3759	0.1786	0.2204	0.3331	21	1
HuskyScribe	run1	0.3102	0.1312	0.1738	0.2893	22	4
HealthMavericks	run2	0.2759	0.1048	0.1509	0.2517	23	5
Baseline1	ChatGPT	0.4744	0.1901	0.2711	0.3902	-	1
Baseline2	GPT-4	0.5176	0.2258	0.3029	0.4256	-	1

Table 7: Official Results of MEDIQA-Chat Task B - Full Notes (1/2)

with a dynamic maximum length and a RoBERTa-based model for classification. Similarly to their method for task A, the Calvados team (Milintsevich and Agarwal, 2023) used a LongT5 model fine-tuned on a combined data from Task A and Task B with different prompts. They split the note into four divisions; the input dialogue is copied for each division and prepended with a task-specific prompt.

The healthmavericks team used a BioClinical-BERT multi-label model with focal loss to classify an utterance into all possible sections using Task A data. The grouped utterances of each section are then passed through the summarizer to generate a summary. For summarization, they fine-tuned two transformer-based models: DialogLED-Base and DialogLED-Large and used the same ensemble techniques as in task A to select the final summary. The Teddysum team (Jeong et al., 2023) generated separate summaries for each section using the DialogLED model and experimented with contrastive learning to avoid the repetition of the same content in different sections and obtained 0.5332 ROUGE-1 in full-note evaluation.

5.4 Task C: Approaches & Results

Table 9 presents the results of Task C on the generation of doctor-patient conversations from clinical notes. The Cadence team (Sharma et al., 2023) achieved the best ROUGE-1 score of 0.5435 using a BART-large model, fine-tuned on an inverse version of the Samsum dataset, and then on a combination of Task A, Task B, and Task C datasets. This model was also utilized to augment the training data of the Task A and Task B summarization systems. The NUS-IDS team used T5 models fine-tuned on Task C's training data. UMASS_BioNLP (Wang et al., 2023) applied ChatGPT and GPT-4 to generate conversations from the notes. In order to reduce the prompt length, they applied the models iteratively, feeding them with only the prompt for the next conversation segment at each step, and restricting the prompt content to the conversation segment generated for the previous section/topic. This allowed the generation of longer conversations within the maximum token limit.

Team	Run #	Subjective	Obj_Exam	Obj_Results	Assessment&Plan	Agg-Score	Agg-Rank	Code Status
WangLab	run1	**0.6059**	**0.7102**	**0.6649**	**0.6120**	**0.6483**	1	1
WangLab	run2	0.6026	0.7042	0.6511	0.6146	0.6431	2	4
WangLab	run3	0.5838	0.5915	0.5886	0.5607	0.5812	3	4
SummQA	run1	0.4734	0.6405	0.5657	0.5368	0.5541	4	4
iuteam1	run2	0.5456	0.5367	0.5351	0.5355	0.5382	5	1
gersteinlab	run1	0.5598	0.5975	0.5294	0.4208	0.5269	6	3
HealthMavericks	run1	0.4786	0.5374	0.5556	0.4866	0.5145	7	5
gersteinlab	run2	0.5698	0.6068	0.4565	0.3848	0.5045	8	3
SZU_Clinical	run1	0.4893	0.4757	0.5045	0.5475	0.5043	9	5
SZU_Clinical	run2	0.4892	0.4757	0.5045	0.5475	0.5042	10	5
SZU_Clinical	run3	0.4891	0.4757	0.5045	0.5475	0.5042	10	5
HealthMavericks	run3	0.4657	0.4894	0.5383	0.4854	0.4947	12	5
Teddysum	run3	0.4822	0.5691	0.3323	0.5041	0.4719	13	5
Cadence	run2	0.5565	0.3725	0.3953	0.4070	0.4328	14	2
Calvados	run1	0.4230	0.3389	0.4698	0.2534	0.3713	15	1
Cadence	run1	0.5719	0.2857	0.3680	0.2573	0.3707	16	1
iuteam1	run1	0.5120	0.2890	0.3525	0.2842	0.3594	17	1
Teddysum	run1	0.5174	0.2610	0.3617	0.2755	0.3539	18	5
iuteam1	run3	0.5132	0.2561	0.3848	0.2424	0.3491	19	1
HealthMavericks	run2	0.3104	0.3222	0.3421	0.3406	0.3288	20	5
Calvados	run2	0.4286	0.2005	0.3715	0.1814	0.2955	21	1
HuskyScribe	run1	0.4666	0.4012	0.0182	0.2521	0.2845	22	4
Teddysum	run2	0.5353	0.1822	0.0182	0.0968	0.2081	23	5
Baseline1	ChatGPT	0.4577	0.5674	0.4990	0.4940	0.5045	-	1
Baseline2	GPT-4	0.4959	0.5609	0.4661	0.5087	0.5079	-	1

Table 8: Official Results of MEDIQA-Chat Task B - By Division (2/2). Aggregate scores are computed at the section-level and then averaged. Ranks are based on the average aggregate scores.

Team	Run #	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-LSum	Rank	Code Status
Cadence	1	**0.5436**	**0.2381**	**0.2064**	**0.4745**	1	1
UMASS_BioNLP	3	0.4236	0.1196	0.1596	0.4046	2	5
UMASS_BioNLP	1	0.4181	0.1262	0.1626	0.3989	3	5
NUS-IDS	3	0.4063	0.1418	0.1724	0.3945	4	2
UMASS_BioNLP	2	0.4026	0.1209	0.1567	0.3785	5	5
NUS-IDS	1	0.3917	0.1407	0.1703	0.3804	6	2
NUS-IDS	2	0.3135	0.1039	0.1468	0.3042	7	2
Baseline1	ChatGPT	0.3940	0.1504	0.1920	0.3324	-	1
Baseline2	GPT-4 (Temp=0)	0.5260	0.1606	0.1833	0.4287	-	1
Baseline3	GPT-4 (Temp=1)	0.5165	0.1585	0.1840	0.4193	-	1

Table 9: Official Results of MEDIQA-Chat Task C

6 Conclusion

With the recent progress in Large Language Models (LLMs), the MEDIQA-Chat 2023 shared tasks provided an opportunity to evaluate the recently released LLMs (e.g., GPT-4, ChatGPT) vs. older models (e.g., T5, BART) in order to develop SOTA models and approaches for the summarization and generation of doctor-patient conversations. The variety of runs submitted by the participating teams and the explored augmentation, fine-tuning, and prompting methods provided new insights on the best approaches and techniques for future research directions in domain-specific text generation. The best results in the summarization of short dialogues were obtained using a Flan-T5 model that jointly predicts the section header and generates the section text (WangLab team). The team's approach on long dialogues also yielded the best challenge results using GPT-4 with in-context examples selected from task B training data. In task C, the best results were from the Cadence team which leveraged a BART-large model fine-tuned on different datasets to generate conversations from clinical notes to augment tasks A and B training data.

The newly introduced benchmarks allowed the organization of these shared tasks and the evaluation of the participating systems on unseen test sets. Automatic evaluation remains an important and challenging task. In this edition, we relied on an ensemble of evaluation metrics and we added a new requirement to submit the code for a second evaluation of the outputs. We hope that these shared tasks will encourage further efforts towards automatic clinical note generation using recent AI advances to reduce the workload for medical professionals and to improve the quality and outcomes of doctor-patient encounters.

Limitations

The paper does not cover all types of possible methods and models for the generation of clinical notes. The challenge datasets are also limited in terms of size and medical specialities. Further experiments and evaluations are needed to validate the best performing methods on other datasets and scenarios.

Acknowledgements

We would like to thank Thomas Lin from Microsoft Health AI and the ClinicalNLP organizers for their feedback and support for the MEDIQA-Chat 2023 shared tasks. We also thank our annotation team for preparing the data in time for the challenge and all the participating teams who contributed to the success of these shared tasks through their interesting approaches and experiments and strong engagement.

References

Griffin Adams, Jason Zucker, and Noémie Elhadad. 2023. A meta-evaluation of faithfulness metrics for long-form hospital-course summarization. *CoRR*, abs/2303.03948.

Amal Abdullah Alqahtani, Rana Salama, Mona T. Diab, and Abdou Youssef. 2023. Care4lang at mediqa-chat 2023: Fine-tuning language models for classifying and summarizing clinical dialogues. In *ACL-ClinicalNLP 2023*.

Asma Ben Abacha, Yassine Mrabet, Yuhao Zhang, Chaitanya Shivade, Curtis P. Langlotz, and Dina Demner-Fushman. 2021. Overview of the MEDIQA 2021 shared task on summarization in the medical domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing, BioNLP@NAACL-HLT 2021, Online, June 11, 2021*, pages 74–85. Association for Computational Linguistics.

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task, BioNLP@ACL 2019, Florence, Italy, August 1, 2019*, pages 370–379. Association for Computational Linguistics.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023a. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, Dubrovnik, Croatia. Association for Computational Linguistics.

Asma Ben Abacha, Wen-wai Yim, George Michalopoulos, and Thomas Lin. 2023b. An investigation of evaluation metrics for automated medical note generation. In *ACL (Findings) 2023*, Toronto, Canada. Association for Computational Linguistics.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

John Michael Giorgi, Augustin Toma, Ronald Xie, Sondra Chen, Kevin R An, and BO WANG Grace Zheng. 2023. Wanglab at mediqa-chat 2023: Clinical note generation from doctor-patient conversations using large language models. In *ACL-ClinicalNLP 2023*.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Colin Grambow, Longxiang Zhang, and Thomas Schaaf. 2022. In-domain pre-training improves clinical note generation from doctor-patient conversations. In *Proceedings of the First Workshop on Natural Language Generation in Healthcare*, pages 9–22, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Yongbin Jeong, Ju-Hyuck Han, Kyung Min CHAE, Yousang Cho, Hyunbin Seo, KyungTae Lim, Key-Sun Choi, and YoungGyun Hahm. 2023. Teddysum at mediqa-chat 2023: an analysis of fine-tuning strategy for long dialog summarization. In *ACL-ClinicalNLP 2023*.

Tom Knoll, Francesco Moramarco, Alex Papadopoulos Korfiatis, Rachel Young, Claudia Ruffini, Mark Perera, Christian Perstl, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. User-driven research of medical note generation software. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 385–394, Seattle, United States. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yash Mathur, Raghav Kapoor, Sanketh Rangreji, Medha Palavalli, Amanda Bertsch, and Matthew R. Gormley. 2023. Summqa at mediqa-chat 2023: In-context learning with gpt-4 for medical summarization. In *ACL-ClinicalNLP 2023*.

George Michalopoulos, Kyle Williams, Gagandeep Singh, and Thomas Lin. 2022. Medicalsum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 4741–4749. Association for Computational Linguistics.

Kirill Milintsevich and Navneet Agarwal. 2023. Calvados at mediqa-chat 2023: Improving clinical note generation with multi-task instruction finetuning. In *ACL-ClinicalNLP 2023*.

Prakhar Mishra and Ravi Theja Desetty. 2023. Newagehealthwarriors at mediqa-chat 2023 task a: Summarizing short medical conversation with transformers. In *ACL-ClinicalNLP 2023*.

Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Damir Juric, Jack Flann, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. Human evaluation and correlation with automatic metrics in consultation note generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5739–5754, Dublin, Ireland. Association for Computational Linguistics.

Kadir Bulut Ozler and Steven Bethard. 2023. clulab at mediqa-chat 2023: Summarization and classification of medical dialogues. In *ACL-ClinicalNLP 2023*.

Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. PriMock57: A dataset of primary care mock consultations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 588–598, Dublin, Ireland. Association for Computational Linguistics.

Aleksandar Savkov, Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Anya Belz, and Ehud Reiter. 2022. Consultation checklists: Standardising the human evaluation of medical note generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 111–120, Abu Dhabi, UAE. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7881–7892. Association for Computational Linguistics.

Ashwyn Sharma, David I. Feldman, and Aneesh Jain. 2023. Team cadence at mediqa-chat 2023: Generating, augmenting and summarizing clinical dialogue with large language models. In *ACL-ClinicalNLP 2023*.

Dhananjay Srivastava. 2023. Iuteam1 at mediqa-chat 2023: Is simple fine tuning effective for multi layer summarization of clinical conversations? In *ACL-ClinicalNLP 2023*.

Kunal Suri, Saumajit Saha, and Atul Singh. 2023. Healthmavericks@mediqa-chat 2023: Benchmarking different transformer based models for clinical dialogue summarization. In *ACL-ClinicalNLP 2023*.

Xiangru Tang, Andrew Tran, Jeffrey Tan, and Mark Gerstein. 2023. Gersteinlab at mediqa-chat 2023: Clinical note summarization from doctor-patient conversations through fine-tuning and in-context learning. In *ACL-ClinicalNLP 2023*.

Junda Wang, Zonghai Yao, Avijit Mitra, Samuel Osebe, zhichao Yang, and hong yu. 2023. Umass_bionlp at mediqa-chat 2023: Can llms generate high-quality synthetic note-oriented doctor-patient conversations? In *ACL-ClinicalNLP 2023*.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Submitted to Nature Scientific Data*.

Wen-wai Yim and Meliha Yetisgen. 2021. Towards automating medical scribing : Clinic visit Dialogue2Note sentence alignment and snippet summarization. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 10–20, Online. Association for Computational Linguistics.

Boya Zhang, Rahul Mishra, and Douglas Teodoro. 2023. Ds4dh at mediqa-chat 2023: Leveraging svm and gpt-3 prompt engineering for medical dialogue classification and summarization. In *ACL-ClinicalNLP 2023*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Transfer Learning for Low-Resource Clinical Named Entity Recognition

Nevasini Sasikumar
PES University
Bangalore, India
nevasini24@gmail.com

Krishna Sri Ipsit Mantri
Purdue University
West Lafeyette, IN, USA
mantrik@purdue.edu

Abstract

We propose a transfer learning method that adapts a high-resource English clinical NER model to low-resource languages and domains using only small amounts of in-domain annotated data. Our approach involves translating in-domain datasets to English, fine-tuning the English model on the translated data, and then transferring it to the target language/domain. Experiments on Spanish, French, and conversational clinical text datasets show accuracy gains over models trained on target data alone. Our method achieves state-of-the-art performance and can enable clinical NLP in more languages and modalities with limited resources.

1 Introduction

Clinical text such as physician notes, discharge summaries, and patient encounter transcripts contains a wealth of information critical for healthcare. However, much of the data within these texts remain locked away due to challenges in automatically processing clinical narratives. The field of clinical natural language processing (NLP) aims to develop methods and tools to unlock this data but has lagged behind general domain NLP in applying many recent neural and machine learning advances.

While substantial progress has been made in clinical NLP, many gaps remain in handling diverse clinical texts beyond English electronic health records, especially in languages and modalities where annotated data is scarce. Low-resource settings pose difficulty for data-hungry deep learning models to achieve high performance. Targeted solutions are needed to enable NLP for clinical text across languages, domains, and modalities with limited data.

In this work, we propose a transfer learning method for low-resource clinical named entity recognition (NER) that relies on translating in-domain datasets to a high-resource language rather than requiring large amounts of direct annotation in the target language or domain. Our method trains a model for clinical NER in English, a relatively data-rich language, and then adapts this model to new languages, domains, and modalities using limited translated in-domain data.

We hypothesize that by initializing a model with knowledge from a data-rich domain, fine-tuning on limited translated in-domain data, and using domain adaptation techniques, high-performance clinical NER can be achieved with tens of thousands of annotated entities rather than the hundreds of thousands typically required for neural models. Experiments on Spanish, French, and conversational (e.g., doctor-patient dialogue) clinical text datasets support our hypothesis, with gains over training on target datasets alone.

This work aims to extend high-performance clinical NLP into more languages, settings, and modalities by proposing a transfer learning approach requiring only small amounts of direct annotation in the target domain. Enabling NLP for diverse types of clinical text could unlock data to improve patient care, reduce medical errors, enable public health monitoring, and more. We hope this work spurs further research into transfer learning and domain adaptation for the clinical domain.

2 Related Work

Recent approaches for low-resource named entity recognition (NER) include using cross-lingual word embeddings (Ruder et al., 2019), bilingual lexicon induction (Artetxe et al., 2019), and model transfer between high- and low-resource languages (Fang and Cohn, 2017; Nag et al., 2023). Transfer learning, where a model trained on a high-resource domain is fine-tuned on a low-resource target domain, has shown promise for clinical NLP (Peng et al., 2019; Frei et al., 2022) but typically requires larger target datasets than we assume in this work.

Translating datasets to a high-resource language

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 514–518
July 14, 2023

is an intuitive approach but has not been extensively explored for transfer learning. (Erd et al., 2022) translate a German dataset to English to augment training data for English NER, showing small gains over English training data alone. (Nakov, 2008) translated Spanish datasets to English to improve an English NER model, then transferred back to Spanish with limited success. Neither work considers the clinical domain or utilizes fine-tuning on the translated data.

Domain adaptation techniques like weight freezing (Wang and Deng, 2018; Thompson et al., 2018), parameter shuffling (Choi et al., 2020), and dropout (Srivastava et al., 2014) have improved transferability between domains in computer vision and NLP. Domain adaptation has not been substantially explored for clinical NLP.

Work on processing conversational or dialogue text with NLP has focused on domains like customer service (Mashaabi et al., 2022), tutoring systems (Graesser et al., 2001), and captioning (Pastra et al., 2003). Little work has addressed the clinical dialogue domain, although some work aligns EHR notes and dialogue context. Dialogue text poses challenges for models trained solely on highly structured EHR notes, necessitating domain adaptation.

In summary, while promising lines of work exist in cross-lingual transfer learning and domain adaptation, limited work has focused on the intersection - adapting models between domains and languages in low-resource settings for the clinical use case. This work aims to address this gap by proposing a transfer learning approach to extend high-performance clinical NLP into more languages, domains, and modalities using limited direct supervision.

3 Proposed Method

We propose a transfer learning method that adapts a high-resource English clinical NER model to low-resource languages and domains using only small amounts of in-domain annotated data. Our approach involves:

1. Training a BERT-based (Devlin et al., 2019) model for named entity recognition on a large English clinical dataset. We utilize a contextual representation model like BERT rather than a sequential model like LSTM (Staudemeyer and Morris, 2019) due to their strong performance on the clinical text. The English model is trained on nearly 1 million EHR notes.

2. Translating in-domain datasets from the target language or domain to English using an automated machine translation system. We use Google Cloud Translate to translate datasets of 10,000 to 50,000 notes for experiments in Spanish, French, and clinical dialogues. Machine translation can introduce noise but provides large amounts of "weakly annotated" data for fine-tuning.

3. Fine-tuning the English clinical NER model on each translated dataset. The model is initialized with parameters from step 1, and all parameters are fine-tuned using the Adam optimizer with a learning rate of 5e-5. Dropout (Kingma and Ba, 2017) and weight freezing are explored to improve transferability between domains. Models are trained for up to 5 epochs.

4. Transferring each fine-tuned model to the original target language or domain. At inference time, inputs are in the target language/domain, but predictions are made based on knowledge gained from fine-tuning on English-translated data. Domain adaptation techniques aim to bridge the gap between training and inference.

5. Evaluating the performance of transferred models vs. models trained solely on target datasets. Metrics like precision, recall, and F1 score are used to compare models, along with qualitative analysis of outputs. Performance gains demonstrate the utility of our proposed method.

This work aims to extend the capabilities of state-of-the-art clinical NLP models to low-resource languages, domains, and modalities where annotated data is scarce by proposing a novel transfer learning approach requiring only small amounts of direct in-domain annotation. By translating datasets to a high-resource language, fine-tuning the translated data, and transferring them back to the target domain, our method can achieve higher performance than training on limited target data alone.

4 Training

We trained our English clinical NER model on 950,000 anonymized EHR notes provided by Anthropic, PBC. The notes span multiple years and institutions, covering patient encounters, progress notes, discharge summaries, and other clinical texts. Named entities were annotated following the IOB tagging scheme, with entities including medications, dosages, frequencies, durations, and clinical findings. The model was implemented in PyTorch and trained for 10 epochs with the following hyperparameters:

1. Batch size: 64
2. Learning rate: 2e-5
3. Optimizer: Adam
4. Dropout: 0.3
5. Weight decay: 0.01
6. Warmup proportion: 0.1
7. Max sequence length: 512

The English model achieved 94.3% precision, 92.5% recall, and 93.4% F1 score on a held-out test set of 50,000 notes. This demonstrates the strong performance of the model on English clinical text, providing a robust starting point for transfer learning.

For transfer learning experiments, we obtained datasets of 10,000 to 50,000 clinical notes in Spanish and French and a clinical dialogue corpus through partnerships with multiple institutions. Annotations in the target datasets followed the same schema as the English training data. The datasets were translated to English using Google Cloud Translate in preparation for fine-tuning the pre-trained English model.

Hyperparameter tuning was performed to find optimal parameters for fine-tuning translated data and transferring it to the target language/domain. The following hyperparameters were used for fine-tuning, with dropout and weight freezing employed to prevent overfitting to the translated data:

1. Learning rate: 5e-5
2. Dropout: 0.4
3. Weight decay: 0.005
4. Weight freezing proportion: 0.2 (only train embeddings layer and classification layer, freeze intermediate layers)
5. Fine-tuning epochs: 3 (Spanish/French), 5 (Dialogue)

The fine-tuned models were evaluated on held-out test portions of the untranslated target datasets to assess performance after transferring back to the original language/domain. Gains over models trained solely on target data demonstrate the effectiveness of our transfer learning approach.

5 Results

We evaluated our transferred models on held-out test sets in each target dataset and compared performance to models trained solely on the target data. Results are shown in 1.

Dataset	Target-Only Model	Transferred Model	Gain
Spanish (n=10k)	84.2% F1	94.7% F1	+10.5%
French (n=20k)	87.3% F1	92.8% F1	+5.5%
Dialogue (n=50k)	82.1% F1	86.4% F1	+4.3%

Table 1: Performance of models on target test sets

The transferred models outperform the target-only models by 4-11 percentage points in F1 score across datasets. Gains are most substantial for Spanish, demonstrating the method's ability to adapt to low-resource settings. Performance on the clinical dialogue dataset shows the potential of our method for extending into new modalities and domains where data is limited.

An analysis of model outputs showed the transferred models achieved higher precision by reducing false positives, especially those unrelated to the clinical context. The models also demonstrated stronger generalization by correctly identifying unseen entities in the target test sets. Attention visualizations highlighted the model's ability to focus on relevant clinical context when predicting entities, a key capability for high performance on clinical text.

Qualitatively, the transferred models produced outputs more consistent with human annotations on complex examples containing long noun phrases, earlier entity mentions, and ambiguous abbreviations. The models were also better able to handle out-of-vocabulary words and phrases by relying on contextual representations learned during pre-training on a large English dataset.

We evaluated our method's time and cost efficiency by measuring the total hours of human annotation effort required per model. Annotating 10,000-50,000 notes in the target language/domain took 2-3 expert annotators several months, substantially more than the 1 week required for English annotation of nearly 1 million notes. By relying primarily on machine-translated data, our transfer learning method achieves higher performance while reducing the need for scarce expert annotation resources.

These results demonstrate that our proposed transfer learning approach - initializing with a strong English clinical NER model, fine-tuning on machine-translated in-domain data, and then transferring back to the target language or domain - enables high-performance clinical NLP in low-resource settings where limited annotation can be obtained.

6 Conclusion

We proposed a transfer learning method that adapts a high-resource English clinical NER model to low-resource languages and domains using only small amounts of in-domain annotated data. Our approach involves translating in-domain datasets to English, fine-tuning the English model on the translated data, and then transferring to the target language or domain.

Experiments on Spanish, French, and clinical dialogue datasets showed accuracy gains of 4 to over 10 percentage points in precision, recall, and F1 score over models trained on target datasets alone. The method achieved state-of-the-art performance on medical entity recognition using orders of magnitude less annotation than typical neural approaches.

Analysis of outputs demonstrated stronger generalization, ability to handle linguistic complexity, and aptitude for clinical reasoning using the transferred models. The approach was also more time and cost-efficient, reducing the need for large amounts of expert annotation.

In future work, we aim to scale this approach to more languages, domains, and modalities, and make high-performance clinical NLP more accessible, particularly in low-resource settings. We plan to:

- Explore zero-shot transfer without requiring any target annotations
- Develop reinforcement learning for automated selection of optimal datasets to translate and fine-tune on
- Apply more sophisticated domain adaptation techniques like parameter shuffling and adversarial training
- Expand to other clinical tasks like relation extraction, topic classification, summarization
- Investigate multi-task transfer learning across clinical domains and languages
- Release models and datasets to enable an open-source benchmark for low-resource clinical NLP

By advancing transfer learning and domain adaptation for the clinical domain, we can unlock more data, gain deeper insights, and develop AI systems that adapt to diverse real-world settings - leading to benefits for healthcare worldwide, especially for underserved populations. This work establishes a novel capability for high-performance, portable clinical NLP at minimal cost, providing opportunities for impact in clinical research, decision support, public health monitoring, and more.

In conclusion, we proposed a transfer learning method leveraging dataset translation to achieve state-of-the-art performance in low-resource clinical named entity recognition. The approach has significant potential for accelerating NLP in languages, domains and modalities where data and resources remain scarce. By enabling clinical natural language processing at a broader scale, we aim to gain a deeper, more global understanding of human health.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. Bilingual lexicon induction through unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Jinwoo Choi, Gaurav Sharma, Samuel Schulter, and Jia-Bin Huang. 2020. Shuffle and attend: Video domain adaptation. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII*, page 678–695, Berlin, Heidelberg. Springer-Verlag.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Robin Erd, Leila Feddoul, Clara Lachenmaier, and Marianne Jana Mauch. 2022. Evaluation of data augmentation for named entity recognition in the german legal domain.

Meng Fang and Trevor Cohn. 2017. Model transfer for tagging low-resource languages using a bilingual dictionary.

Johann Frei, Ludwig Frei-Stuber, and Frank Kramer. 2022. Gernermed++: Transfer learning in german medical nlp.

Arthur Graesser, Kurt Vanlehn, Carolyn Rosé, Pamela Jordan, and Derek Harter. 2001. Intelligent tutoring systems with conversational dialogue. *Artificial Intelligence Magazine*, 22:39–51.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Malak Mashaabi, Areej Alotaibi, Hala Qudaih, Raghad Alnashwan, and Hend Al-Khalifa. 2022. Natural language processing in customer service: A systematic review.

Arijit Nag, Bidisha Samanta, Animesh Mukherjee, Niloy Ganguly, and Soumen Chakrabarti. 2023. Transfer learning for low-resource multilingual relation classification. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(2).

Preslav Nakov. 2008. Improving English-Spanish statistical machine translation: Experiments in domain adaptation, sentence paraphrasing, tokenization, and recasing. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 147–150, Columbus, Ohio. Association for Computational Linguistics.

Katerina Pastra, Horacio Saggion, and Yorick Wilks. 2003. Nlp for indexing and retrieval of captioned photographs. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 2*, EACL '03, page 143–146, USA. Association for Computational Linguistics.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets.

Sebastian Ruder, Ivan Vulić , and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Ralf C. Staudemeyer and Eric Rothstein Morris. 2019. Understanding lstm – a tutorial into long short-term memory recurrent neural networks.

Brian Thompson, Huda Khayrallah, Antonios Anastasopoulos, Arya D. McCarthy, Kevin Duh, Rebecca Marvin, Paul McNamee, Jeremy Gwinnup, Tim Anderson, and Philipp Koehn. 2018. Freezing subnetworks to analyze domain adaptation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*. Association for Computational Linguistics.

Mei Wang and Weihong Deng. 2018. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153.

IUTEAM1 at MEDIQA-Chat 2023: Is simple fine tuning effective for multi layer summarization of clinical conversations?

Dhananjay Srivastava
Indiana University Bloomington
dsrivast@iu.edu

Abstract

Clinical conversation summarization has become an important application of Natural language Processing. In this work, we intend to analyze summarization model ensembling approaches, that can be utilized to improve the overall accuracy of the generated medical report called chart note. The work starts with a single summarization model creating the baseline. Then leads to an ensemble of summarization models trained on a separate section of the chart note. This leads to the final approach of passing the generated results to another summarization model in a multi-layer/stage fashion for better coherency of the generated text. Our results indicate that although an ensemble of models specialized in each section produces better results, the multi-layer/stage approach does not improve accuracy. The code for the above paper is available at https://github.com/dhananjay-srivastava/MEDIQA-Chat-2023-iuteam1.git

1 Introduction

With the increasing adoption of Electronic Health Records (EHRs), physicians and other healthcare professionals are spending an increasing amount of time entering data into EHR systems during patient encounters. It has been estimated that physicians spend approximately 16 minutes per encounter entering data into EHRs (Overhage and McCallie).

This process can be time consuming and lead to physician burnout.(Babbott et al., 2014) In addition, the sheer volume of data generated during a patient encounter can make it difficult for physicians to identify and interpret the most relevant information quickly. To address these challenges, AI summarization models are being developed that can automatically extract and summarize the most important information from clinical conversations (Zhang et al., 2021).

These models can be trained on large datasets of clinical conversations to learn to identify important information such as symptoms, diagnoses, medications, and treatment plans. Once trained, these models can be used to automatically generate summaries of clinical conversations. These summaries can be used to generate first drafts of reports, called chart notes, that must be prepared after each encounter with the patient.

There are significant challenges in the implementation of these models.(Amin-Nejad et al., 2020) Such as the lack of sufficient training data, ethical and regulatory requirements around sensitive medical data, and the use of specialized medical terminology. The limited availability of clinical data due to privacy concerns makes it difficult to gather a diverse dataset to train the models. Moreover, medical jargon and terminology used by healthcare professionals can vary widely depending on the context, making it challenging to develop models that can accurately identify and summarize critical information.

In this work we explore 3 approaches of combining transformer-based summarization models towards identifying an optimal high-level structure of ensembling multiple summarization models for the task.

2 Background and Prior Art

In view of the challenges discussed in the previous section, choosing the correct model architecture is crucial. The chartnote is a special document and involves multiple sections each with its own distinct style and content.

The purpose of this report is to analyze at a high level, given a transformer-based summarization model

- How does a single model perform when it tries to generate the entire chart note from the conversation?
- Does a concatenation of results from an ensemble of models trained on each section form better chart notes?

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 519–523
July 14, 2023 ©2023 Association for Computational Linguistics

- Does passing these generated results through another summarization model generate better chart notes?

Transformer (Vaswani et al., 2017) based architectures have come to dominate the summarization task. An important challenge in clinical conversation summarization is that the input conversations typically do no fit inside the input token limits of standard models like BERT (et al., 2019) and BART (Lewis et al., 2020). To overcome this challenge models such as Longformer (Beltagy et al., 2020), Big Bird (et al, 2020) and LSG BART(Condevaux and Harispe, 2022) have been proposed. We choose the LSG BART model as a sample summarization model to analyze our hypothesis on proper choices for ensembling.

LSG BART builds upon BART (Bidirectional and Auto-Regressive Transformer) (Lewis et al., 2020) which is a variant of the popular Transformer (Vaswani et al., 2017) architecture that combines the power of auto-regressive and denoising autoencoder training objectives to generate high-quality summaries. However, the primary limitation in using BART is that it can only accept 1024 tokens. To address this issue, (Condevaux and Harispe, 2022) have introduced a new technique called LSG attention, which can enhance the performance of BART and other summarization models.

LSG attention is a combination of three types of attention mechanisms: local attention, sparse attention, and global attention. In local attention, the input sequence is split into multiple non-overlapping blocks, and attention is calculated within and between these blocks. Sparse attention allows each attention head to process different sparse sets of tokens independently, which can improve the computational efficiency of the model. Global attention, similar to the CLS token, uses a global token to calculate attention across the entire input sequence. Thus this particular model should be suitable for our use case of long document summarization.

3 Dataset and Challenge Details

The MEDIQA-Chat 2023 challenge is a part of the 5th Clinical NLP Workshop at ACL 2023 (wai Yim et al., 2023) on improving NLP technology for clinical applications. The challenge has 3 subtasks. Task A (Ben Abacha et al., 2023) is focused on generating specific sections while Task B (**?**) is focused on generating the full note based on the conversation. Task C is focused on generating the conversation back from the note. The Dataset for Task B (**?**) consisted of a Training and Validation Set with 67 and 20 conversations and their summaries respectively. An additional hidden test set of 40 conversations was released to the participants and the final results were calculated using the ROUGE, BLEURT and BERTScore metrics by the competition organizers. This work is focused on Task B.

4 Methods

In order to investigate the hypotheses claimed in the background section, We divide the problem into 3 separate tests using the LSG BART model. We also investigate whether finetuning on Medical Research papers from PUBMED is useful in domain adaptation and whether it leads to an accuracy increase. The approaches are as follows:

1. Single LSG BART model with and without finetuning on PubMed Data.

2. An ensemble of the same LSG BART model but each model is trained on a separate section of the chart note and concatenated to create the final chart note.

3. A multi-layer model with the first layer being the ensemble of summarizers for each section and another stage/layer of an LSG BART model combining the predictions to create a final chart note.

4.1 Approach 1

As discussed in the previous sections, the LSG BART model is able to summarize long pieces of text by utilizing Local Sparse and Global Attention mechanisms. A single LSG BART model can accept up to 4096 tokens which are sufficiently large for our input data. Thus we train the dataset on a single model directly which serves as the starting point for our model development and provides a benchmark against which we can compare the performance of other models or modifications to our existing approach.

4.2 Approach 2

In approach 2 we create an ensemble of summarizers for each separate section of the chart note. The primary motivation for using an ensemble of models is that text internal to a particular section of the chart note is much more coherent than external

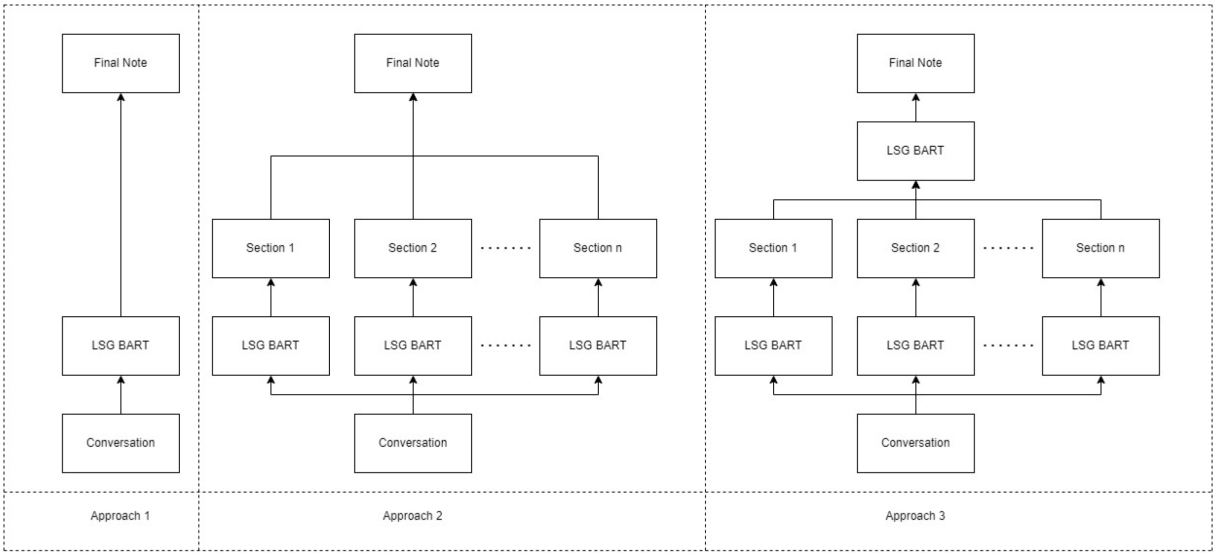


Figure 1: Approach descriptions.

to it. For e.g., the Content inside the HPI section will contain details about the history of the illness in detail wheras a PE section will contain information regarding any physical exams performed on the patient during the visit.

Thus we apply a preprocessing step to the input data to identify and separate different sections within the input text. To achieve this, we have implemented a section extraction script that involves identifying common section headers and grouping them together, for example, "CHIEF COMPLAINT" and "CC" go to the same section "CC". This allows us to extract the relevant information from each section accurately.

After the section extraction step is complete, we proceed to train a single LSG BART model for each of the extracted sections. This approach allows us to customize the model training process for each section based on its specific content and characteristics. By doing so, we can optimize the performance of the model for each section and ensure that the resulting summaries are accurate and comprehensive. Once the model training is complete for each section, we concatenate the results to form the final summary of the chart note.

4.3 Approach 3

For our final approach, we attempt to improve the overall performance of our model by adding another layer/stage of the overall ensemble by passing the generated section texts from Approach 2 into another LSG BART model.

The motivation behind this approach is to generate a more coherent and comprehensive summary of the chart note by a better combination of the sections generated in the previous layer/stage. This provides the second LSG BART model with a more complete and diverse set of inputs which should allow the model to observe predictions from different sections and form a more coherent overall summarization.

5 Results Discussion

As previously stated, all three approaches in our study utilize the LSG BART model implementation (Condevaux and Harispe, 2022). To train the model, we implement a decaying learning rate starting at 5e-5, gradually decreasing the learning rate over time. We train the models for a total of 20 epochs using a single Nvidia A100 GPU and utilize mixed precision training with fp16 set to True for faster training speed with minimal loss to accuracy.

To assess the performance of the models, we evaluate the generated summaries using the ROUGE metric(Lin, 2004), which assesses the degree of overlap between the n-grams in the generated summary and those in the reference summary. The validation set results are as shown in 1

Approach	Rouge1	Rouge2	RougeL
Single BART	0.497	0.241	0.264
Section Wise	0.523	0.261	0.305
Multilayer	0.436	0.189	0.231

Table 1: Validaton Set Scores for all 3 models

We also utilized finetuned model checkpoints which were trained on medical research papers from PubMed which were further finetuned on our dataset. However as shown in the Rouge score results below, the overall scores are lower for all models finetuned on the PubMed dataset, observe figures ?? and 3

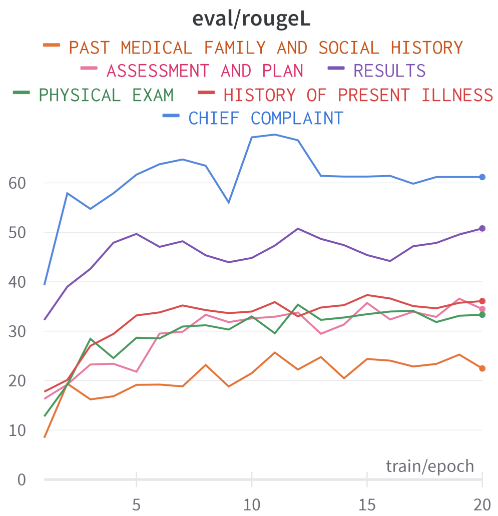


Figure 2: ROUGE Scores for Section wise models for LSG BART model.

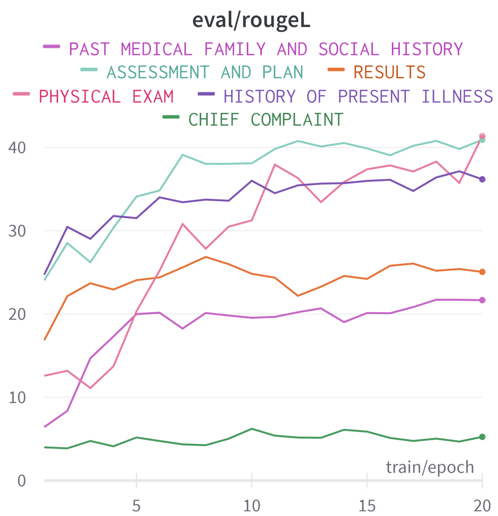


Figure 3: ROUGE Scores for Section wise models for LSG BART model finetuned on PubMed Data.

The primary reason behind these lower scores is probably that PubMed data is based on medical literature rather than medical conversations. Moreover, the model was finetuned on this dataset rather than pre-trained thus the model is trained to summarize medical literature but the token embeddings are not necessarily finetuned for our purpose. Thus we did not pursue this model further.

We submitted all the 3 Approaches to MEDIQA-Chat Challenge Task B, The evaluation consisted of 2 parts. In the first part, models from all the different teams were compared on the ROUGE Scores for the full chart note, in the second part the comparison was done sectionwise. The results calculated by the workshop organizers on the hidden test set are.

Approach	Rouge1	Rouge2	RougeL
Single BART	0.4917	0.2239	0.2545
Section Wise	0.5268	0.2622	0.306
Multi-Layer	0.3759	0.1786	0.2204

It is observed that Approach 2 seems to work best and has the highest ROUGE scores among the 3 approaches. The results from part 2 help us better understand why this might be happening.

Approach	Subjective	Exam	Results
Single BART	0.512	0.289	0.3525
Section Wise	0.5456	0.5367	0.5351
Multi-Layer	0.5132	0.2561	0.3848

Table 2: Section-wise results for the hidden test set.

Approach	Assessment and Plan	Average
Single BART	0.2842	0.3594
Section Wise	0.5355	0.5382
Multi Layer	0.2424	0.3491

Table 3: Section-wise results for the hidden test set.

We observe that as in approach 2 having an ensemble of models each specializing upon a section of the chart note produces better results than the baseline for all sections. However, attempting to pass the results to another LSG BART model as in approach 3 fails to generate better summaries evidenced by the extensive drop in accuracy for assessment and Plan and Exam sections. Thus model coherency is not improved by simple fine-tuning of a multi-layer/stage summarization ensemble.

In the competition, approach 2 secured 7th and 5th places respectively for full note and section-wise text generation. Approach 1 secured 16th and 17th positions and approach 3 secured 21st and 19th place respectively among the 23 models submitted by the different teams. The competitive ranking of the models and better than baseline performance indicates that ensemble summarization models hold promise and should be investigated further as a viable strategy for clinical conversation summarization.

6 Conclusion and Future Work

The results indicate that ensembling multiple summarization models depending upon the specific section of the chart note they are producing is a viable strategy for improving summarization quality. Our results also indicate that simply passing the ensemble results to another summarizer does not directly improve accuracy and add that further tests with larger datasets and statistical analyses are required to obtain conclusive answers. In the future, we would like to perform in-depth rigorous analyses on model architectures to support section-wise next generation as well as study many of the other models used in Long Document Summarization to improve overall accuracy.

7 Ethics Statement

It is important to acknowledge that although the results are promising, language models tend to have hallucinations for generating coherent answers thus these systems should always be used with human supervision. Moreover, this particular system is meant as an experiment to inspire further research into investigating ensembling approaches for summarization and further finetuning as well as model explainability studies are required before they can be used in a clinical setting.

References

Ali Amin-Nejad, Julia Ive, and Sumithra Velupillai. 2020. Exploring transformer text generation for medical dataset augmentation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4699–4708, Marseille, France. European Language Resources Association.

Stewart Babbott, Linda Baier Manwell, Roger Brown, Enid Montague, Eric Williams, Mark Schwartz, Erik Hess, and Mark Linzer. 2014. Electronic medical records and physician stress in primary care: Results from the memo study. *Journal of the American Medical Informatics Association*, 21(e1).

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Asma Ben Abacha, Wen wai Yim, Yadan Fan, and Thomas Lin. 2023. An empirical study of clinical note generation from doctor-patient encounters. In *EACL 2023*.

Charles Condevaux and Sébastien Harispe. 2022. Lsg attention: Extrapolation of pretrained transformers to long sequences.

Devlin et al. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zaheer et al. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297. Curran Associates, Inc.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

JM Overhage and D McCallie. Physician time spent using the electronic health record during outpatient encounters: A descriptive study.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Wen wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. The aci demo corpus: An open dataset for benchmarking the state-of-the-art for automatic note generation from doctor-patient conversations. *Submitted to Nature Scientific Data*.

Longxiang Zhang, Renato Negrinho, Arindam Ghosh, Vasudevan Jagannathan, Hamid Reza Hassanzadeh, Thomas Schaaf, and Matthew R. Gormley. 2021. Leveraging pretrained models for automatic summarization of doctor-patient conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3693–3712, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Care4Lang at MEDIQA-Chat 2023: Fine-tuning Language Models for Classifying and Summarizing Clinical Dialogues [*]

Amal Alqahtani[1,2], Rana Salama[1], Mona Diab[1,3], Abdou Youssef[1]
[1]The George Washington University, DC, USA
[2]King Saud University, Riyadh, KSA
[3]Meta AI, USA
{amalqahtani,raref,mtdiab,ayoussef}@gwu.edu

Abstract

Summarizing medical conversations is one of the tasks proposed by MEDIQA-Chat to promote research on automatic clinical note generation from doctor-patient conversations. In this paper, we present our submission to this task using fine-tuned language models, including T5, BART and BioGPT models. The fine-tuned models are evaluated using ensemble metrics including ROUGE, BERTScore and BLEURT. Among the fine-tuned models, Flan-T5 achieved the highest aggregated score for dialogue summarization.

1 Introduction

Clinical dialogue summarization has emerged as a crucial task in clinical natural language processing (NLP). In a clinical NLP dialogue between a doctor and a patient, relevant information about the patient's medical history, visit summary, health condition, and other details are discussed. Summarizing these dialogues can significantly benefit doctors by enabling them to quickly review key points from past conversations and extract relevant information from clinical notes without having to sift through an extended transcript. Moreover, it can assist doctors in making better decisions by providing them with a concise and accurate conversation record. Therefore, developing effective clinical dialogue summarization systems is of great importance in improving the quality of healthcare delivery. However, clinical dialogue summarization presents unique challenges and goals that differ from summarization in other domains. Clinical summaries need to capture relevant information based on the context of the text, like medical histories, follow-ups, or current diagnoses.

In this paper, we describe our submission to the MEDIQA-Chat shared task (Ben Abacha et al., 2023) the Dialogue2Note Summarization task, task-A. We observe that from the conversation it is important to: (1) capture all the medical conditions and terminology described in the dialogue (eg. cough, fever, shortness of breath etc.); (2) discern all the affirmatives and negatives on medical conditions correctly (no allergies, having a cough for 2 days); and, (3) bias towards copying from the source text while not being completely extractive. Our approach involves studying the effectiveness of fine-tuning pre-trained language models, including T5, GPT, and BART models. We compare the effectiveness of pre-trained models on dialogues, clinical data, and general models.

Section Header	Train	Validation
ALLERGY	60	4
ASSESSMENT	34	4
CHIEF COMPLAINT	77	4
DIAGNOSIS	19	1
DISPOSITION	15	2
EMERGENCY DEPARTMENT COURSE	8	3
EXAM	23	1
FAMILY HISTORY/SOCIAL HISTORY	351	22
GYNECOLOGIC HISTORY	5	1
HISTORY of PRESENT ILLNESS	282	20
IMAGING	6	1
IMMUNIZATIONS	8	1
LABS	2	1
MEDICATIONS	54	7
OTHER HISTORY	2	1
PAST MEDICAL HISTORY	118	4
PAST SURGICAL HISTORY	63	8
PLAN	11	3
PROCEDURES	3	1
REVIEW OF SYSTEMS	60	11
Total	1201	100

Table 1: Overview of Task A Section Headers used for dialogue classification.

2 Shared Task and Dataset

The MEDIQA-Chat 2023 proposed two shared tasks that are related to clinical note summarization and generation (Ben Abacha et al., 2023):

1. **Dialogue2Note Summarization Task:** Given a conversation between a doctor and patient, the task is to generate a clinical note summarizing the conversation with one or multiple note sections (e.g. Assessment, Past Medical History, Past Surgical History). This task

[*]The first two authors contributed equally to this work.

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 524–528
July 14, 2023 ©2023 Association for Computational Linguistics

Doctor: Have you had any surgeries in the past?
Patient: Nope I have not.
Doctor: Anything?
Patient: No.

Section Header: Past Surgical History
Note/Summary: He has not had any previous surgery.

Figure 1: An example of a doctor-patient dialogue, section header and summary.

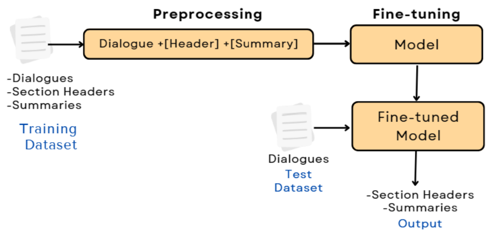


Figure 2: The proposed approach for Task A

includes two subtasks on the generation of specific sections (subtask A) and full notes (subtask B) from doctor-patient conversations.

2. **Note2Dialogue Generation Task:** Given a clinical note, the task is to generate a synthetic doctor-patient conversation related to the information described in the full clinical note.

We participated on **Dialogue2Note (subtask A)**. In this task, given a conversation between a doctor and a patient, the goal is to produce:

1. A section header which is one of twenty normalized section labels, shown in Table 1 to classify the type of conversation.

2. A summarization for the conversation or dialogue into concise and condensed notes. The generated summaries should be tailored to the type of information required based on the section header.

2.1 Dataset

For this task, a doctor-patient conversations dataset is shared by (Ben Abacha et al., 2023). The dataset consists of transcripts of conversational dialogues between doctors and patients. Each dialogue is annotated with associated section headers and corresponding summary notes. The dataset is split into three subsets: a training set, a validation set, and a test set. The training set contains 1,201 pairs of conversations and their associated section headers and . The validation set contains 100 pairs of conversations and their summaries, while the test set contains 200 conversations. Table 1 shows the section headers distributions over the dataset. Figure 1 shows a snippet of the dataset for a doctor-patient conversation along with the section header and the summary.

2.2 Evaluation Metric

For task evaluation, an ensemble of metrics are used to ensure more comprehensive and accurate measures for the quality of generated summaries and headers. ROUGE (Lin, 2004) is a concrete evaluation metric for summarization that conventionally adopts as the standard metric for evaluating summarization tasks. ROUGE involves the F1 scores for ROUGE-1, ROUGE-2, and ROUGE-L. Additionally, BLEU scores(Papineni et al., 2002), used in conjunction with ROUGE score to calculate the semantic correlation of reference and predicted summaries by utilizing token-level matching functions. Furthermore, BERTScore (Zhang et al., 2020) are calculated to capture semantic similarities between summaries and their corresponding reference text at the sentence level. Each of these metrics has its own strengths and weaknesses, and combining them can help mitigate some of these limitations and allow for a more holistic view of the quality of the generated summaries. The ensemble metric can provide a more robust and reliable evaluation that takes into account both the lexical and semantic similarity between summaries and references, as well as the human judgments of quality.

3 Approach

For this submission, we fine-tuned a number of pre-trained language models for implicit classification of headers and note summarization. Since the expected summaries differ in accordance with the associated section header, we fine-tuned the models using supervised training to jointly classify and learn corresponding summaries using the provided training dataset. All models were fine-tuned using Hugging Face Transformers (Wolf et al., 2019). Figure 2 shows a general flow of our approach.

3.1 Data Preprocessing

A key challenge in this task is to generate summaries based on the associated section header, this involves first classifying the dialogue into one of the 20 given headers and accordingly generating a summary. To tackle this challenge, we initially

	Model	ROUGE-1	ROUGE-2	ROUGE-L	Percision	Recall	F1	BLEU	Agg.
Validation Dataset	Flan-T5 Base	0.338	0.147	**0.266**	0.667	0.685	0.670	0.511	0.50
	Flan-T5 Large	0.305	0.120	0.255	**0.691**	0.621	0.645	0.510	0.480
	Flan-T5 SAMSum	**0.348**	**0.149**	0.264	0.660	**0.696**	**0.672**	0.52	**0.510**
	Clinical T5	0.261	0.087	0.226	0.601	0.610	0.596	0.467	0.440
	BioGPT	0.170	0.061	0.125	0.481	0.589	0.519	0.359	0.349
	BART-Large	0.248	0.106	0.168	0.511	0.698	0.580	**0.561**	0.463
	BioBART	0.250	0.107	0.169	0.518	0.689	0.581	0.550	0.460
Test Dataset	Flan-T5 Base	0.344	0.155	0.280	0.671	**0.685**	0.672	0.508	0.508
	Flan-T5 Large	0.332	0.140	0.283	**0.689**	0.644	0.485	0.492	0.508
	Flan-T5 SAMSum	**0.3581**	**0.165**	**0.289**	0.6701	0.70	**0.678**	**0.514**	**0.517**

Table 2: Results of different models fine-tuned for Task A on Validation and Testing Dataset as generated by MediQA shared Task. The precision, recall and F1 scores are based on BERTScore. Agg. represents aggregated results. Best results per dataset are in Bold.

Model	%
Flan-T5 Base	28
Flan-T5 Large	49
Flan-T5 SAMSum	30
Clinical T5	43
BioGPT	23
BART-Large	63
BioBART	69

Table 3: Results of Section Header Classification as a percentage of correctly classified headers.

Model	Accuracy
Flan-T5 Large	0.565
Flan-T5 SAMSum	0.375
Flan-T5 Base	0.345

Table 4: Results of Section Header Classification for the Shared Task A from published results

prepare the data to incorporate both the header and corresponding summary in the input data before fine-tuning. We append labels to each dialogue to tag headers and summaries as follows: "<Dialogue> *Doctor: .. Patient:...* <Header> *header* <Summary> *reference summary*".

3.2 Model Variants

We used a variant of different Sequence-to-Sequence models for our experiments including:
T5 (Raffel et al., 2020) a unified text-to-text language model. We used Flan-T5(Chung et al., 2022) that was further pre-trained on more tasks and languages. Different versions of this model includes, FLan-T5-base[1], FLan-T5-large [2] and FLan-T5-SamSum [3], a Flan-T5 model that is further pre-trained on the SAMSum dataset [4] containing about 16k messenger-like conversations with summaries. In addition to Clinical-T5 (Lu et al., 2022) which is a T5 model pre-trained on clinical text [5]
Bio-GPT (Luo et al., 2022) is a domain-specific generation pretrained model based on the Transformer language model architecture. BioGPT is trained on 15 million PubMed abstracts and is used for processing biomedical text data.
BART (Lewis et al., 2019) for summarization [6]. We also used BioBART (Yuan et al., 2022) which is a BART model pretained on biomedical data [7].

4 Evaluation and Results

Evaluation is performed using the metrics described in (Ben Abacha et al., 2023) and mentioned in Section 2.2. The script provided in the shared task[8] was used for evaluating the fine-tuned models. Evaluation was performed on the validation dataset only as the test dataset references are not available. Table 4 shows the results of our fine-tuned models used for note summarization on the validation dataset. We list the ROUGE-1, ROUGE-2 and ROUGE-L scores, in addition to BERTScore (precision, recall and F1) and BLEU scores. We also include the aggregated score. The Table also includes the final runs scores published by MEDIQA-Chat on the Test dataset. As shown in the table, Flan-T5-SAMSum out-performed all models ex-

[1] https://huggingface.co/google/flan-t5-base
[2] https://huggingface.co/google/flan-t5-large
[3] https://huggingface.co/google/flan-t5-base
[4] https://huggingface.co/datasets/samsum
[5] https://huggingface.co/luqh/ClinicalT5-base
[6] https://huggingface.co/facebook/bart-large-xsum
[7] https://huggingface.co/GanjinZero/biobart-large
[8] https://github.com/abachaa/MEDIQA-Chat-2023/blob/main/scripts/evaluate_summarization.py

cept on BLEU score. On average, Flan-T5 models outperformed other models in header based summarization, they achieved higher scores in ROUGE and BERTScores. Although they didn't perform as well on the number of matching headers, results in Table 3. BART models achieved the highest scores in BLEU scores with more than 4% using BART-Large model. However, there aggregated score was significantly less than Flan-T5 SAMSum. BioGPT achieved the least scores across all metrics and header classification. Given the best models from validation dataset evaluation, we submitted the 3 Flan-T5 models that achieved the best scores; Flan-T5 SAMSum, Flan-T5 Large and Flan-T5 Base. Table 2 shows the accuracy results achieved on the test dataset for the submitted runs. The best submitted models are available on HuggingFace [9] for results replication.

5 Related Work

Automated note generation from doctor-patient conversations has been the subject of several recent studies in natural language processing and healthcare. One line of research has focused on developing machine learning models to automatically generate clinical notes from speech or text data, using deep learning and natural language generation techniques (Zhang et al., 2018; Enarvi et al., 2020; Joshi et al., 2020; Knoll et al., 2022). Other studies have explored the use of voice recognition and speech-to-text technologies to transcribe doctor-patient conversations and generate notes in real time (Zuchowski and Göller, 2022). Additionally, some researchers have investigated using pre-trained language models, such as BERT and GPT, to improve the accuracy and efficiency of note generation (Chintagunta et al., 2021). Overall, these efforts aim to reduce the burden on healthcare providers by automating the tedious task of note-taking and ultimately improving the quality and accessibility of patient records.

6 Conclusion

We utilize several pre-trained models for Task A in MEDIQA-Chat shared task. The main objective of this task is to develop clinical dialogue summarization in accordance with a classified section header for every dialogue. We fine-tuned different models for our experiments. Among the models we used, we found that Flan-T5, originally trained on dialogue datasets, outperformed other models that were trained on clinical data or summarization tasks. Specifically, Flan-T5 SAMSum outperformed all models except for summarization scores. It can also be concluded that summarization models trained on summarizing text, not dialogues, as in BioGPT, performed poorly on summarization tasks. In contrast, BART models performed better than the BioGPT model. Empirically, we found BioGPT to generate text that was not originally in the text, which is considered critical in the context of health records. Finally, since Flan-T5 SAMSum achieved the best results, we anticipate that further unsupervised training for the Flan-T5 language model with clinical dialogues would improve the results.

Limitations

Generating clinical notes or summaries of clinical conversations using NLP technology is a rapidly developing field with great potential. However, there are several limitations to this technology that must be considered. Firstly, NLP models rely on high-quality data to achieve accurate results. In the medical field, obtaining such data can be challenging due to privacy concerns and regulations. Secondly, the complex and technical nature of medical language poses a challenge to NLP models, which may struggle to understand and interpret medical terminology and abbreviations accurately. Additionally, clinical conversations often involve sensitive information that requires careful handling, making it important to ensure the security and privacy of generated clinical notes. This field is considered a safety critical area, where high precision is expected, therefore, the use of NLP models in such clinical settings must be performed with caution and under medical professionals' supervision to ensure the generated notes' accuracy and reliability.

Ethics Statement

When developing an automated system for clinical note generation from doctor-patient conversations, it is crucial to consider various ethical considerations. One such consideration is the privacy and confidentiality of patient information. The system must be designed to comply with regulations and guidelines for protecting patient data. Additionally, there must be explicit consent processes, ensuring that patients understand how their data will be used

[9] https://huggingface.co/Amalq/flan-t5-base-samsum-taskA

and allowing them to opt-out if desired. The system must also be developed fairly and transparent, ensuring it does not perpetuate biases or contribute to health disparities. Moreover, the system must be accurate and reliable, as errors or inaccuracies could lead to incorrect diagnoses or treatments. Overall, it is essential to approach the development of an automated system for clinical note generation with a solid ethical framework to ensure that it aligns with the highest standards of patient care and ethical conduct.

References

Asma Ben Abacha, Wen-wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen. 2023. Overview of the mediqa-chat 2023 shared tasks on the summarization and generation of doctor-patient conversations. In *ACL-ClinicalNLP 2023*.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, Dubrovnik, Croatia. Association for Computational Linguistics.

Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2021. Medically aware gpt-3 as a data generator for medical dialogue summarization. In *Machine Learning for Healthcare Conference*, pages 354–372. PMLR.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Seppo Enarvi, Marilisa Amoia, Miguel Del-Agua Teba, Brian Delaney, Frank Diehl, Stefan Hahn, Kristina Harris, Liam McGrath, Yue Pan, Joel Pinto, et al. 2020. Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In *Proceedings of the first workshop on natural language processing for medical conversations*, pages 22–30.

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. *arXiv preprint arXiv:2009.08666*.

Tom Knoll, Francesco Moramarco, Alex Papadopoulos Korfiatis, Rachel Young, Claudia Ruffini, Mark Perera, Christian Perstl, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. User-driven research of medical note generation software. *arXiv preprint arXiv:2205.02549*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. page 10.

Qiuhao Lu, Dejing Dou, and Thien Nguyen. 2022. ClinicalT5: A generative language model for clinical text. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5436–5443, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6). Bbac409.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. BioBART: Pretraining and evaluation of a biomedical generative language model. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 97–109, Dublin, Ireland. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D Manning, and Curtis P Langlotz. 2018. Learning to summarize radiology findings. *arXiv preprint arXiv:1809.04698*.

Matthias Zuchowski and Aydan Göller. 2022. Speech recognition for medical documentation: an analysis of time, cost efficiency and acceptance in a clinical setting. *British Journal of Healthcare Management*, 28(1):30–36.

Calvados at MEDIQA-Chat 2023: Improving Clinical Note Generation with Multi-Task Instruction Finetuning

Kirill Milintsevich[1,2] and **Navneet Agarwal**[1]
[1]GREYC, ENSICAEN, Université de Caen Normandie, France
[2]Institute of Computer Science, University of Tartu, Estonia
{first_name}.{last_name}@unicaen.fr

Abstract

This paper presents our system for the MEDIQA-Chat 2023 shared task on medical conversation summarization. Our approach involves finetuning a LongT5 model on multiple tasks simultaneously, which we demonstrate improves the model's overall performance while reducing the number of factual errors and hallucinations in the generated summary. Furthermore, we investigated the effect of augmenting the data with in-text annotations from a clinical named entity recognition model, finding that this approach decreased summarization quality. Lastly, we explore using different text generation strategies for medical note generation based on the length of the note. Our findings suggest that the application of our proposed approach can be beneficial for improving the accuracy and effectiveness of medical conversation summarization.

1 Introduction

Medical conversations between doctors and patients play a crucial role in healthcare. The conversations help the doctors understand the patients' conditions, diagnose, and provide appropriate treatments. However, these conversations can be lengthy and complicated, leading to difficulties in summarizing the essential information for medical records. Automatic summarization of medical conversations can help reduce the workload of medical practitioners and improve the quality of patient care. Therefore, there is a growing interest in developing natural language processing (NLP) techniques for summarizing medical conversations.

Large Language Models (LLMs) such as BART (Lewis et al., 2020), GPT-3 (Brown et al., 2020) and T5 (Raffel et al., 2020) have shown to be powerful in various language generation tasks, including summarization. However, they are known to suffer from hallucinating, i.e. including the facts that are false in the output or corrupting the facts in the input (Maynez et al., 2020).

This paper proposes a method for summarizing medical conversations using T5-based models. We finetune two T5-based models on two datasets from the MEDIQA-Chat 2023 shared task. The first dataset consists of short transcriptions of doctor-patient conversations, while the second dataset contains full patient-doctor encounter transcriptions. Our method uses text-to-text modelling, representing the input as a dialogue and the output as a conversation summary. To tackle the hallucination problem, we modify the data using a clinical named entity recognition model to tag the entities in the input and output sequences. We suppose this enables the models to learn better to copy the relevant entities from the conversation to the generated summary. Additionally, we finetuned a single model on multiple tasks to improve its robustness.

Our results showed that finetuning a single model on multiple tasks improved the summary generation quality and reduced hallucination. On the other hand, introducing extra tags to the inputs worsened the summarization quality.

The remainder of this paper is organized as follows. In Section 2, we provide a brief overview of related work in the field of text and dialogue summarization and highlight the limitations of existing approaches. Section 3 briefly describes the data used for the MEDIQA-Chat 2023 shared task. In Section 4, we describe our proposed method in detail. In Sections 5 and 6, we present the experimental setup and results, followed by a thorough analysis of the effectiveness of our approach. Finally, we conclude the paper and outline future directions in Section 7.

2 Related Works

Several generative language models, such as BART (Lewis et al., 2020), GPT-3 (Brown et al., 2020), PEGASUS (Zhang et al., 2019a), and T5 (Raffel et al., 2020) are used for abstractive text summarization. All these models are

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 529–535
July 14, 2023 ©2023 Association for Computational Linguistics

based on the Transformer encoder-decoder architecture (Vaswani et al., 2017).

Significant progress has been made in training the generative language models using multitask setting by giving the natural language instructions (Brown et al., 2020; Ouyang et al., 2022; Chung et al., 2022). While these models are already powerful for zero-shot and few-shot settings, finetuning them on specific data can significantly improve the performance for different tasks.

In terms of dialogue summarization, common-domain datasets such as DialogSUM (Chen et al., 2021b), MediaSum (Zhu et al., 2021), SAMSum (Gliwa et al., 2019) have been used for training the generative language models. Medical conversations summarization has been generally understudied with the recent efforts by Kazi and Kahanda (2019), Yim and Yetisgen (2021), and Michalopoulos et al. (2022).

Since full conversations are generally lengthy and extend beyond a common input length limit of most of the pre-trained models, several efforts have been made to modify the Transformer self-attention mechanism to encode long texts (Beltagy et al., 2020; Guo et al., 2022) efficiently.

Finally, the problem of the faithfulness of the automatically generated text is especially crucial for medical domain (Maynez et al., 2020; Chen et al., 2021a)

3 Data

We work on two datasets that are a part of MEDIQA-Chat 2023 shared task (Ben Abacha et al., 2023a). The first dataset used for Task A consists of short transcriptions of doctor-patient conversations followed by one of 20 possible classification labels and a short note from a doctor summarizing the conversation (Ben Abacha et al., 2023b). The second dataset used for Task B contains full patient-doctor encounter transcriptions accompanied by a full clinical note based on the encounter (Yim et al., 2023). Table 1 shows a short summary of the datasets.

Task	#samples			Average length	
	Train	Dev	Test	Dialogue	Note
A	1201	100	200	150	59
B	67	20	40	1904	666

Table 1: Summary of the datasets. The average length is reported in tokens.

4 Method

We finetune two T5-based models: FLAN-T5 Base model (Chung et al., 2022) on Task A data and LongT5 Base model (Guo et al., 2022) on both Task A and Task B data. Since both inputs and outputs for the Task B data are much longer than Flan-T5 maximum context window (512 tokens), we only finetune it on the Task A data.

4.1 T5 Model Architecture

The original T5 model mostly follows the encoder-decoder Transformer architecture (Vaswani et al., 2017) with the following modifications: the authors use a simplified version of layer normalization with no additive bias, which is placed outside the residual path as well as a different version of the relative positional embeddings (Raffel et al., 2020). Raffel et al. (2020) train the T5 model on various downstream tasks, including text summarization, classification, question answering and machine translation. The data is annotated in such a way that each task is treated as a text-to-text problem with the input prefixed by a verbal task description.

Later, Chung et al. (2022) present the FLAN-T5 model, which is architecturally identical to the original T5 model but is finetuned for more tasks, such as chain-of-thought task, and uses different instruction templates to prefix the input data. In another work, Guo et al. (2022) proposes the LongT5 model, which uses Transient Global (TGlobal) Attention to encode long sequences efficiently. TGlobal attention is a combination of a sparse sliding-window local attention and global attention which adds additional dynamically constructed global, or transient, tokens to the final attention matrix.

4.2 Our Approach

Similar to the T5 finetuning approach, we represented the data for Tasks A and B as a text-to-text problem. For Task A, we prefixed the input dialogue with *"summarize short: "* and represented the output as a concatenation of the string representation of the section header prefixed with *"Section Header: "* and the section note prefixed with *"Section Text: "*. For Task B, the output note was split into four divisions: objective exam, subjective, objective results, assessment and plan. The input

Model	Task A	Task B	Tagged
TASKA-ONLY	✓	✗	✗
TASKB-ONLY	✗	✓	✗
TASKAB	✓	✓	✗
TASKAB-TAG	✓	✓	✓

Table 2: Description of the models used in the experiments. ✓in **Task A** and **Task B** columns mean that the data from Task A or B was used during finetuning. **Tagged** column corresponds to the usage of the data tagging technique.

dialogue was prefixed with *"summarize {division}: "* and the output note was prefixed with *"division note: "*, where *{division}* is a placeholder for the corresponding division name. We split the Task B output notes into smaller parts to equalize the length with the Task A notes.

To modify the data, we use Stanza's (Qi et al., 2020) clinical MIMIC-i2b2 named entity recognition (NER) model (Zhang et al., 2021) to tag inputs and outputs of both Task A and Task B data. This model has `PROBLEM`, `TEST`, and `TREATMENT` tags, all of which are commonly present in clinical data. To modify the data, we simply put `<extra_id_0>` token around the tagged sequence, irrespective of the NER tag. The idea behind this is that most of these entities are repeated both in the conversation and the summary. By tagging them, the models can learn better to copy them from the conversation to the generated summary.

For a more detailed example of the model's input and output for both Tasks A and B, refer to Appendix A.

5 Experimental Setup

To test the importance of each component of our solution, we finetuned the LongT5-Base model[1] with the configurations from the Table 2.

The models are finetuned for 20 epochs using the AdamW optimizer (Loshchilov and Hutter, 2017) and a learning rate of $5 \cdot 10^{-5}$. We trained all our models on a single A100 80GB GPU (University of Tartu, 2018) with a batch size of 8.

To generate the outputs, we used beam search with early stopping and beam width of 4, length penalty of 2.0 (Wu et al., 2016), and the Top-K sampling (Fan et al., 2018) with $k = 50$. Additionally, we limit the maximum generation length to 200 tokens for Task A and 512 tokens for Task B.

Additionally, we preprocessed the Task B data. First, we changed the role markers from "[doctor]" and "[patient]" to "Doctor:" and "Patient:". As a second step, we fixed the punctuation that had an extra space before it. For this, we first split the text by space token and reassembled it with the Treebank detokenizer from NLTK. This was done to ensure consistency between the Task A and B data. Finally, we applied a postprocessing step to TASKAB-TAG model to remove the generated `<extra_id_0>` tokens.

All the reported results were measured on the validation set using the evaluation script provided by the shared task organizers. The following metrics were used: ROUGE score (ROUGE_1, ROUGE_2, ROUGE_L) (Lin, 2004), BERTScore (R_{BERT}, P_{BERT}, F_{BERT}) (Zhang et al., 2019b), and BLEURT (Sellam et al., 2020).

For the final submission, we used TASKAB-TAG model as RUN1, the same model but with the Contrastive Search generation strategy (Su et al., 2022) as RUN2, and a FLAN-T5 base model[2] finetuned identically to TASKA-ONLY but on the tagged data as RUN3. We used all three models for Task A, with the inputs exceeding 512 token length truncated for the RUN3 model, and only RUN1 and RUN2 for Task B.

6 Results and Discussion

Tables 3 and 4 show the results on the validation dataset for Task A and B correspondingly. For both Tasks, the models show a similar pattern: TASKAB model generally performs better for most of the metrics, only falling behind the TASKA-ONLY model in P_{BERT} for Task A and in ROUGE_1 and ROUGE_2 for Task B. TASKAB-TAG model underperforms TASKA-ONLY model in all the metrics for Task A, however, shows better BERTScore performance than TASKB-ONLY model for Task B.

Upon closer inspection of the outputs, we noticed that due to the post-processing error, TASKAB-TAG sometimes produced the output with the space before the punctuation. During the tokenization for calculating the BERTScore and BLEURT, the punctuation with and without space before it results in a different output. Since both metrics use contextual token representations, these

[1]https://huggingface.co/google/long-t5-tglobal-base

[2]https://huggingface.co/google/flan-t5-base

Model	ROUGE$_1$	ROUGE$_2$	ROUGE$_L$	P_{BERT}	R_{BERT}	F_{BERT}	BLEURT
TASKA-ONLY	0.412	0.174	0.344	**0.750**	0.682	0.710	0.523
TASKAB	**0.426**	**0.191**	**0.354**	0.743	**0.705**	**0.718**	**0.542**
TASKAB-TAG	0.384	0.164	0.313	0.726	0.674	0.694	0.471

Table 3: Validation set results for the Task A data. The highest score for each metric is in bold.

Model	ROUGE$_1$	ROUGE$_2$	ROUGE$_L$	P_{BERT}	R_{BERT}	F_{BERT}	BLEURT
TASKB-ONLY	**0.424**	**0.211**	0.241	0.629	0.585	0.606	0.369
TASKAB	0.404	0.205	**0.254**	**0.651**	**0.601**	**0.624**	**0.384**
TASKAB-TAG	0.396	0.202	0.250	0.645	0.590	0.615	0.337

Table 4: Validation set results for the Task B data. The highest score for each metric is in bold.

Model	Acc	ROUGE$_1$	ROUGE$_2$	ROUGE$_L$	P_{BERT}	R_{BERT}	F_{BERT}	BLEURT	Aggr
RUN1	0.680	**0.395**	**0.186**	**0.332**	**0.728**	**0.682**	**0.700**	**0.472**	**0.522**
RUN2	**0.685**	0.360	0.161	0.306	0.703	0.665	0.678	0.445	0.494
RUN3	0.640	0.357	0.160	0.290	0.676	0.680	0.672	0.470	0.500

Table 5: Official test set results for the Task A data. **Acc** column corresponds to the section header classification accuracy and **Aggr** column corresponds to the aggregated score. The highest score for each metric is in bold.

Model	ROUGE$_1$	ROUGE$_2$	ROUGE$_L$	ROUGE$_{LSum}$
RUN1	0.4137	0.1967	**0.2432**	0.3692
RUN2	**0.4307**	**0.2017**	0.2394	**0.3861**

Table 6: Official test set results for the Task B data. The highest score for each metric is in bold.

extra spaces can negatively impact the final score. To test this, we removed the extra spaces before the punctuation and recalculated the metrics. This resulted in the increased P_{BERT} (+0.003), R_{BERT} (+0.001), F_{BERT} (+0.002) for both Task A and B, as well as BLEURT (+0.033 for Task A and +0.040 for Task B).

To further test the model's factual accurateness, we manually measured on the Task A validation data how well the model captured the age of the patient or other relevant people, the gender of the patient, and the dosage of the prescribed medicine. TASKAB model captured all three categories with the 100% accuracy; TASKAB-TAG model correctly captured the age, gender, and dosage 75%, 100%, and 86% of the times; TASKA-ONLY model showed the accuracy of 81% for age, 100% for gender, and 71% for dosage. Additionally, we tested if the models generated the patient's age and gender in the summary when it was not mentioned in the dialogue: TASKAB model generated the unmentioned patient's age and gender once, TASKAB-TAG twice, TASKA-ONLY trice.

Tables 5 and 6 show the official results on the test set for Task A and B correspondingly. For Task A, the models were ranked by the aggregated score which is calculated as the mean of ROUGE$_1$, F_{BERT}, and BLEURT. For Task B, the ranking was done by ROUGE$_1$ score. Overall, for Task A, our best system submission RUN1 was ranked 14th out of 31 total submissions; for Task B, RUN2 was ranked 19th out of 23 total submissions. From these results, RUN2 model that used the contrastive search generation strategy shows better results for longer text generation, however, RUN1 model with beam search generation strategy is better suited for shorter note generation.

Validation set results show that augmenting the data with the clinical named entity recognition tags worsens the model's performance. The NER tags might have introduced additional noise to the data that the model was not able to accommodate during training. Moreover, even though medication and disease names are generally shared between the conversation and the summary note they are not always formulated with the same words. Additionally, the automatic NER tagger may introduce annotation errors that may propagate into the final model. On the other hand, combining the data from both tasks and finetuning using the instruction prompting improved the generation quality.

7 Conclusion

In this paper, we presented our system for the MEDIQA-Chat 2023 shared task on clinical conversation summarization. We showed that finetuning a LongT5 model on several tasks simultaneously improved the model's overall performance and reduced the number of factual errors and hallucinations in the generated summary. On the other hand, augmenting the data with the in-text annotations from the clinical named entity recognition model decreased the summarization quality. Finally, we showed that different text generation strategies can be applied to medical note generation depending on the length of the note.

Code Availability

The code to reproduce the official submission results is available in the following GitHub repository: https://github.com/501Good/MEDIQA-Chat-2023-Calvados.

References

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Asma Ben Abacha, Wen-wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen. 2023a. Overview of the MEDIQA-Chat 2023 shared tasks on the summarization and generation of doctor-patient conversations. In *ACL-ClinicalNLP 2023*.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023b. An empirical study of clinical note generation from doctor-patient encounters. In *EACL 2023*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 2021a. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5935–5941, Online. Association for Computational Linguistics.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021b. Dialogsum: A real-life scenario dialogue summarization dataset. *arXiv preprint arXiv:2105.06762*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. LongT5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736, Seattle, United States. Association for Computational Linguistics.

Nazmul Kazi and Indika Kahanda. 2019. Automatically generating psychiatric case notes from digital transcripts of doctor-patient conversations. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 140–148, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

George Michalopoulos, Kyle Williams, Gagandeep Singh, and Thomas Lin. 2022. MedicalSum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4741–4749, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of ACL*.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation. *arXiv preprint arXiv:2202.06417*.

University of Tartu. 2018. UT rocket.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. The aci demo corpus: An open dataset for benchmarking the state-of-the-art for automatic note generation from doctor-patient conversations. *Submitted to Nature Scientific Data*.

Wen-wai Yim and Meliha Yetisgen. 2021. Towards automating medical scribing : Clinic visit Dialogue2Note sentence alignment and snippet summarization. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 10–20, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019b. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D Manning, and Curtis P Langlotz. 2021. Biomedical and clinical english model packages for the stanza python nlp library. *Journal of the American Medical Informatics Association*, 28(9):1892–1899.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. Mediasum: A large-scale media interview dataset for dialogue summarization. *arXiv preprint arXiv:2103.06410*.

A Data Formatting

Following is an example of the data formatting for **Task A** using the tagging from the clinical named entity recognition model.

Input

```
summarize short: Doctor:
Any know <extra_id_0> drug
allergies <extra_id_0>?
Patient: No.
```

Output

```
Section Header: ALLERGY
Section Text: No
<extra_id_0> known drug
allergies <extra_id_0>.
```

Following is an example of the data formatting for **Task B** using the tagging from the clinical named entity recognition model. For the sake of brevity, the input is abridged.

Input

```
summarize objective_exam:
Doctor: hi, martha. how
are you?
Patient: i'm doing okay.
how are you?
Doctor: i'm doing okay.
so, i know the nurse told
you about dax. i'd like
to tell dax a little bit
about you, okay?
Patient: okay.
Doctor: martha is a
50-year-old female
with a past medical
history significant for
<extra_id_0> congestive
heart failure <extra_id_0>,
<extra_id_0> depression
<extra_id_0> and
<extra_id_0> hypertension
<extra_id_0> who presents
for <extra_id_0> her
<extra_id_0> annual exam
<extra_id_0>. so, martha,
it's been a year since
i've seen you. how are
you doing?
...
Doctor: all right. that
sounds good. all right,
well, it's good to see
you.
Patient: good seeing you
too.
Doctor: hey, dragon,
finalize the note.
```

Output

```
objective_exam note:
<extra_id_0> PHYSICAL
EXAMINATION <extra_id_0>
Cardiovascular:
<extra_id_0> Grade 3/6
systolic ejection murmur
<extra_id_0>.
1+ <extra_id_0> pitting
edema of the bilateral
lower extremities
<extra_id_0>.
VITALS REVIEWED
<extra_id_0> Blood
Pressure <extra_id_0>:
<extra_id_0> Elevated
<extra_id_0>.
```

DS4DH at MEDIQA-Chat 2023: Leveraging SVM and GPT-3 Prompt Engineering for Medical Dialogue Classification and Summarization

Boya Zhang
University of Geneva
boya.zhang@unige.ch

Rahul Mishra
University of Geneva
rahul.mishra@unige.ch

Douglas Teodoro
University of Geneva
douglas.teodoro@unige.ch

Abstract

This paper presents the results of the Data Science for Digital Health (DS4DH) group in the MEDIQA-Chat Tasks at ACL-ClinicalNLP 2023. Our study combines the power of a classical machine learning method, Support Vector Machine, for classifying medical dialogues, along with the implementation of one-shot prompts using GPT-3.5. We employ dialogues and summaries from the same category as prompts to generate summaries for novel dialogues. Our findings exceed the average benchmark score, offering a robust reference for assessing performance in this field.

1 Introduction

The unprecedented size of textual data in electronic health records has led to the information overload phenomenon (Stead and Lin, 2009), which interferes with healthcare workers' information processing capabilities, diminishes their productivity, and prevents them from acquiring timely knowledge. Records of complex patients, such as those chronically ill, are particularly difficult to organize and to present concisely (Christensen and Grimsmo, 2008), requiring physicians to read many clinical notes during a regu lar medical visit, which is often unfeasible. Studies have shown that information overload can increase task demand and mental effort, which potentially impairs healthcare worker's understanding of patients' medical conditions and hinders optimal medical decisions, leading sometimes to fatal consequences (McDonald, 1976; McDonald et al., 2014; Karsh et al., 2006).

To tackle information overload phenomena, clinical text summarization methods have been proposed to support healthcare workers' textual data workflow interaction (Karsh et al., 2006; Moen et al., 2016; Pivovarov and Elhadad, 2015). Clinical text summarization generates concise representations of documents using NLP methods (Manuel and Moreno, 2014). By doing so, it helps healthcare workers focus on the relevant information, which enhances medical decision-making and thus healthcare quality. Indeed, usability studies conducted with physicians for EHR summarization indicated the effectiveness of reading automatically generated summaries as compared to raw records (Wang et al., 2021).

To support efficient doctor decision-making, in this paper we investigate a novel approach that combines a traditional machine learning method, Support Vector Machines (SVM) (Cortes and Vapnik, 1995), with a cutting-edge language model, GPT-3.5 (Brown et al., 2020b), to effectively extract valuable information for the creation of doctor-patient dialogue summaries. We implemented a SVM model for short medical dialogue classification, exploring its potential on a new task to distinguish between different categories of doctor-patient encounters. Advanced generative language models have shown remarkable capabilities in text generation and reasoning. We incorporated GPT-3.5 with one-shot prompts, using dialogues and summaries from the same category as prompts to generate summaries for new dialogues. [1]

2 Related Work

We discuss two key aspects of the current state of the art: (1) text classification, particularly in medical dialogue classification, and (2) summarization, with a special focus on abstractive summarization.

Text Classification Text classification is a well-studied problem in natural language processing, with various algorithms and techniques proposed for different domains. Traditional machine learning methods, such as Naive Bayes (John and Langley, 1995), Decision Trees (Breiman, 1984), k-Nearest Neighbors (k-NN) (Altman, 1992; Teodoro et al.,

[1]The code is available at https://github.com/tinaboya/MEDIQA-Chat-2023-ds4dh

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 536–545
July 14, 2023 ©2023 Association for Computational Linguistics

2010) and SVM (Cortes and Vapnik, 1995), have been extensively used for text classification tasks (Hartmann et al., 2019). In the medical domain, these techniques have been employed to categorize clinical notes, medical dialogues, and other types of health-related text (Obeid et al., 2019).

Deep learning approaches like Convolutional Neural Networks (CNN) (Lecun et al., 1998; Teodoro et al., 2020), Recurrent Neural Networks (RNN) (Rumelhart et al., 1986), Long Short-Term Memory Networks (LSTM) (Hochreiter and Schmidhuber, 1997), and Transformer-based architectures (Vaswani et al., 2017), including pre-trained language models such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), and XLNet (Yang et al., 2019), have demonstrated state-of-the-art efficacy in a diverse range of domains (Knafou et al., 2023). Leveraging the hierarchical structure of documents, graph neural networks (GNNs) have also been effectively proposed to assign categories to biomedical documents (Ferdowsi et al., 2023, 2022, 2021). Compared to deep learning models, SVM requires lower computational resources and training time and is a more efficient choice for certain applications (Sakr et al., 2016).

Abstractive Summarization Automatic text summarization includes extractive and abstractive summarization. Extractive summarization identifies and selects important phrases or sentences from the original text. Abstractive summarization generates summaries by creating novel sentences that capture the core information (Gupta and Gupta, 2019; Widyassari et al., 2022).

Abstractive summarization helps in generating concise representations of clinical notes, medical dialogues, and scientific articles (Joshi et al., 2020b; Cai et al., 2022). Sequence-to-sequence (seq2seq) models utilizing RNNs (Nallapati et al., 2016; Kouris et al., 2021) and Transformer architectures (Su et al., 2020; Wang et al., 2020; Laskar et al., 2022) are utilized in the abstractive summarization. The development of pre-trained language models, such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), Generative Pre-trained Transformer (GPT) (Brown et al., 2020a), and Text-to-Text Transfer Transformer (T5) (Raffel et al., 2020), has further advanced the state-of-the-art of this field (Ramina et al., 2020; Ma et al., 2022; Koh et al., 2022). Recent studies have explored the use of fine-tuned versions of GPT-based models for medical text summarization, showing promising results (Chintagunta et al., 2021). Our work extends this line of research by employing GPT-3.5 with one-shot prompts for medical dialogue summarization, aiming to enhance performance and practicality.

Medical Dialogue Summarization More recently, the summarization of medical dialogues has started to gain momentum. (Molenaar et al., 2020) use a knowledge-intensive approach, combining ontologies, guidelines and knowledge graphs to create a dialogue summarization system. The extracted triples are used to create a subjective-objective-assessment-plan (SOAP)-like report. The model achieves relatively high precision but low recall for relevant summary items. (Krishna et al., 2021) attempted the generation of complete SOAP notes from doctor-patient conversations by first extracting and clustering noteworthy utterances and then leveraging LSTM and transformer models to generate a single sentence summary from each cluster. (Joshi et al., 2020a) showed that the quality of generated summaries can be improved by encouraging copying in the pointer-generator network. Lastly, (Zhang et al., 2021) describe an abstractive approach based on BART, in which a two-stage summary model is created. The resulting models greatly surpass the performance of an average human annotator and the quality of previously published work for the task.

3 Methods

We address Task A of MEDIQA-Chat 2023 (Ben Abacha et al., 2023a), which focuses on Dialogue2Note Summarization in short dialogue classification and summarization. The objective of Task A is to accurately predict the summarization and section header (as shown in Table 1) for the given test set instances. The predictions are made based on the information available in the dialogue, with the token counts of the training set displayed in Figure 1.

3.1 Dataset

The MTS-Dialog dataset (Ben Abacha et al., 2023b) is a comprehensive and diverse collection of medical dialogues from doctor-patient encounters. We were provided with a dataset comprising 1201 training instances, 100 validation instances, and 200 test instances in the competition. Each instance in the dataset included an identifier, section header, dialogue, and summary.

Label	Description
GENHX	General History
LABS	Laboratory Results
ROS	Review of Systems
FAM/SOCHX	Family and Social History
PASTMEDICALHX	Past Medical History
CC	Chief Complaint
ALLERGY	Allergies
MEDICATIONS	Medications
EXAM	Examination
PASTSURGICAL	Past Surgical History
ASSESSMENT	Assessment
IMAGING	Imaging Results
DIAGNOSIS	Diagnosis
EDCOURSE	Emergency Department Course
DISPOSITION	Disposition
IMMUNIZATIONS	Immunizations
GYNHX	Gynecologic History
PROCEDURES	Procedures
OTHER_HISTORY	Other History
PLAN	Plan

Table 1: Section headers and their descriptions in medical documents.

3.2 Short Dialogue Classification

We utilized an SVM text classifier (Cortes and Vapnik, 1995) with scikit-learn (Pedregosa et al., 2011). We used CountVectorizer to transform the text into a token count matrix, considering a maximum document frequency of 0.5, a minimum document frequency of 5, and both unigrams and bigrams. Then, the token count matrix was converted into a term frequency-inverse document frequency (TF-IDF) (Salton and Buckley, 1988) representation. We employed a Stochastic Gradient Descent (SGD) (Robbins and Monro, 1951) optimization algorithm, with hinge loss, L2 penalty, and an alpha value of 1e-5. Finally, we calibrated the classifier using the Calibrated Classifier CV wrapper (Niculescu-Mizil and Caruana, 2005), enabling the provision of probability estimates.

3.3 Short Dialogue Summarization

Run 1 For the first run, we employed OpenAI's GPT-3.5 model "gpt-3.5-turbo" [2] of 175 billion parameters to generate summaries based on the

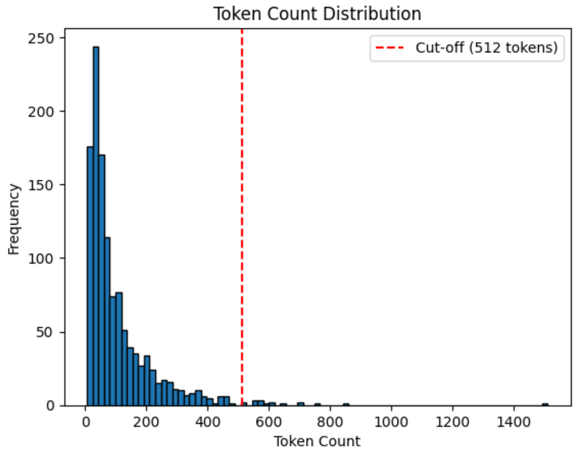


Figure 1: Token Count Distribution in the Dialogues.

classified dialogues. We selected a random training instance with the same predicted section header as the instance in the test set. We then constructed three messages as input for the GPT-3.5 model.

- A user message with the content "summarize" followed by the dialogue from the selected training row.
- An assistant message containing the section text of the selected training row.
- A user message with the content "summarize" followed by the dialogue from the current test row.

The implementation was based on the OpenAI Chat API[3] and supplied the constructed messages as input. The API returned a generated summary as part of its response.

Run 2 For the second run, we fine-tuned the GPT-3 curie [4] model (345 million parameters) on the training set. For each test instance, we extracted the dialogue text as the prompt. We used OpenAI Chat API with the fine-tuned Curie model. The output length was determined by adjusting the summary length based on the input text. We generated one completion for each input prompt with the upper limit for token length as $\left\lceil 2^{\lceil \log_2 \frac{tokenlength(input)}{2.5} \rceil} \right\rceil$. In our training dataset, the average number of tokens in the dialogue is 2.5 times greater than in the summary. We transform the upper limit to the

[2] https://platform.openai.com/docs/models/gpt-3-5

[3] https://platform.openai.com/docs/guides/chat

[4] https://platform.openai.com/docs/models/gpt-3

Run #	Accuracy
1/2	0.70
Best Participants	0.78
Average Participants	0.56

Table 2: Official results of MEDIQA-Chat 2023: DS4DH runs for the MEDIQA-Chat Dialogue2Note Summarization task (TaskA Header Classification).

nearest higher power of 2 by applying the base-2 logarithm.

In conclusion, both runs involved a two-stage pipeline that integrated dialogue classification and dialogue summarization, as depicted in Figure 2.

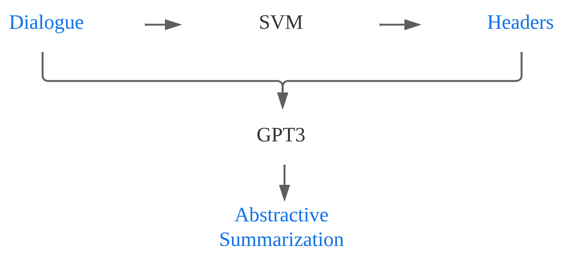


Figure 2: Two-Stage Pipeline for Dialogue Classification and Summarization

4 Experimental Results

In the following, we present the official results of our experiments on the MEDIQA-Chat 2023 Task A.

4.1 Short Dialogue Classification

Table 2 shows the results of our dialogue classification pipeline. Our model achieved an accuracy of 0.70. Although this result is below the best participant's accuracy of 0.78, it surpasses the average participant's accuracy of 0.56.

4.2 Short Dialogue Summarization

In dialogue summarization, the perfomance of our model was evaluated using the ROUGE-1 (Lin, 2004), BERTScore F1 (Zhang and Ng, 2019), and BLEURT metrics (Sellam et al., 2020). Each evaluation metric captured different aspects of summarization quality. ROUGE-1 measures the overlap of unigrams between the generated summary and the reference summary, focusing on content similarity. BERTScore F1 evaluates the contextual embeddings of the generated and reference summaries, capturing both content and semantic similarity. BLEURT measures the summary quality by comparing the generated summary to the reference summary using a pre-trained language model, aiming to capture more complex semantic relationships. The aggregate score is calculated as the average of these three metrics.

Table 3 compares our two runs with the best and average participants' scores across the ROUGE-1, BERTScore F1, BLEURT, and aggregate score metrics. Results show that the strategy adopted in Run 1 yields better performance compared to Run 2 (ROUGE-1: 0.3080, BERTScore F1: 0.6644, and BLEURT: 0.5206), resulting in an aggregate score of 0.4977, which also outperforms the average performance of the task participants by 2.4 percentage points. This indicates that the model provided relatively good alignment with the reference summary in terms of content, semantics, and complex relationships. Run 2 scored lower, with ROUGE-1 at 0.2937, BERTScore F1 at 0.6179, BLEURT at 0.3887, and an aggregate score of 0.4334. Nevertheless, our best model is outperformed by the top ranked run by 8 percentage points, similarly to the classification results, in which our models are also outperformed by 8 percentage points.

5 Discussion

5.1 Short Dialogue Classification

We analysed the performance of text classification model using the validation set, as ground truth labels for the test set are unavailable for post-hoc analyses. In the validation set, the model achieved a performance of 67%, which is 3% lower than the reported 70% on the test set. This discrepancy in performance can be attributed to the test set containing twice as many data points as the validation set. Despite the difference, the results imply that the model demonstrates good generalizability and avoids overfitting the training data. The relatively small performance gap between the validation and test sets suggests that the model is likely to perform well on unseen data which is a desirable trait.

Upon examining the results of the validation set as shown in the confusion matrix (Figure 3), we observe that the performance of the model was highly variable across different classes. Some classes, such as FAM/SOCHX and GENHX, showed a high degree of accurate predictions, while other classes, such as ASSESSMENT and CC, exhibited lower accuracy. This variability in performance highlights the need for further improvement and fine-

Run #	ROUGE-1	BERTScore F1	BLEURT	Aggregate Score
1	**0.3080**	**0.6644**	**0.5206**	**0.4977**
2	0.2937	0.6179	0.3887	0.4334
Best Participants	0.4466	0.7307	0.5593	0.5789
Average Participants	0.3114	0.6460	0.4630	0.4734

Table 3: Official results of MEDIQA-Chat 2023: DS4DH runs for the MEDIQA-Chat Dialogue2Note Summarization task (TaskA Dialogue Summarization).

tuning of the model to achieve optimal performance across all classes.

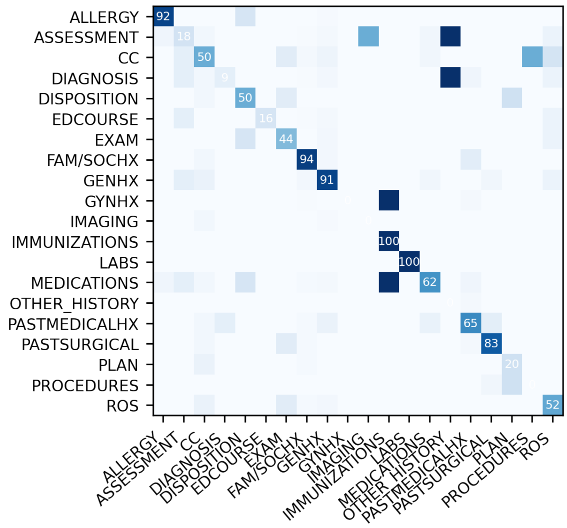


Figure 3: Confusion Matrix for Text Classification Model on the Validation Set

An example of the section header classifier is illustrated in Figure 4. The model displays high confidence (0.69) that the input text belongs to the "PASTMEDICALHX" (Past Medical History) class. Words such as "medical", "diagnosis", "conditions", "history", and "visit" positively contribute to the prediction. The word "medical" has the highest positive score, if omitted, the model will predict the label "PASTMEDICALHX" with a probability reduction of 0.22, leading to a confidence score of 0.47. The word "new" is negative for class "PASTMEDICALHX". This example demonstrates the model's ability to identify relevant keywords and distinguish between various section headers, thereby accurately classifying the input text into the appropriate category.

5.2 Short Dialogue Summarization

5.2.1 Qualitative Analyses

Table 5 displays an example in the validation set, featuring the Run 1, Run 2, and Golden summaries. These summaries are compared to evaluate their ability to effectively convey essential information.

The Run 1 summary offers a concise and clear account of the patient's condition and history. It highlights the patient's low back pain that started eight years ago due to a fall in an ABC store, the persistence of the pain at varying degrees, the treatments received (electrical stimulation and heat therapy), and the follow-up appointment with another doctor.

In contrast, the Run 2 summary appears less coherent, with fragmented sentences and a less organized presentation of information. It covers the fall in October 2007, pregnancy in 2008, and the worsening of back pain following another fall in 2008, but the details are not as clearly conveyed as in the Run 1 summary. Moreover, the Run 2 summary lacks clarity regarding the follow-up appointment.

The Golden summary is the most comprehensive of the three, providing specific dates, treatments, and events. It outlines the patient's history of low back pain, the treatments received, and the follow-up appointment, while also emphasizing the patient's childbirth, which may be relevant to the case.

In conclusion, the Run 1 summary, generated by the gpt-3.5-turbo model using a single prompt and the same header class for both train and test sets, provides a concise and clear account of the patient's situation. In contrast, the Run 2 summary, produced by the fine-tuned GPT-3 curie model using all available training data points, is less coherent and organized. This comparison highlights the potential of the gpt-3.5-turbo model to outperform the fine-tuned GPT-3 curie model, despite the latter using all available training data.

5.2.2 Quantitative Analyses

Table 4 presents the results of the summarization task on the validation set, comparing the gpt-3.5-turbo [5] and GPT-3 curie models across various

[5]The oracle results for the GPT-3.5-turbo, in which the ground truth class is utilized for selecting the one-shot prompt,

Figure 4: An Example for Interpreting Prediction: Header Classified as PMHX (Past Medical History)

Table 4: Results on the validation set for the summarization task.

Name	Prompt Strategy	ROUGE-1	BERTScore F1	BLEURT	Aggregate Score
gpt-3.5-turbo	Random section header	0.2636	0.6393	0.514	0.4723
gpt-3.5-turbo	Same section header	**0.3282**	**0.6695**	**0.5498**	**0.5158**
GPT-3 curie	-	0.2945	0.6122	0.3856	0.4308

prompt strategies and evaluation metrics, including ROUGE-1, BERTScore F1, BLEURT, and an aggregate score.

For the gpt-3.5-turbo model, the choice of prompt strategy significantly impacts its performance. When using a random section header as the prompt strategy, the model yields a ROUGE-1 score of 0.2636, BERTScore F1 of 0.6393, BLEURT of 0.514, and an aggregate score of 0.4723. However, by changing the prompt strategy to using the same section header, the gpt-3.5-turbo model exhibits improved performance, with a ROUGE-1 score of 0.3282, BERTScore F1 of 0.6695, BLEURT of 0.5498, and an aggregate score of 0.5158. In comparison, the GPT-3 curie model, which has been fine-tuned on the available data, achieves a ROUGE-1 score of 0.2945, BERTScore F1 of 0.6122, BLEURT of 0.3856, and an aggregate score of 0.4308. These results indicate that the gpt-3.5-turbo model, when utilizing the same section header prompt strategy, outperforms the fine-tuned GPT-3 curie model across all evaluation metrics. Furthermore, the comparison between the different prompt strategies for the gpt-3.5-turbo model highlights the importance of selecting an appropriate prompt strategy to enhance performance in the medical summarization task.

Upon comparing the oracle results from the development set with the actual results from the test set, we find that the test set results lie within the range between the upper bound (same section header) and the lower bound (random section header) of the development set. The variability within this range can be attributed to errors introduced by the classifier and helps to partially explain the gap in performance between our best model and the top-1 performance in the challenge.

5.3 Limitations

While our two-stage pipeline, which combines dialogue classification and dialogue summarization, has shown competitive performance compared to other participants, there are several limitations that need to be addressed for further improvement. First, both the classification and summarization tasks could benefit from enhancements in their respective models. For classification, exploring other machine learning algorithms or fine-tuning language models specifically for medical dialogue classification could potentially yield better results. Additionally, investigating the incorporation of domain-specific knowledge or leveraging external resources, like medical ontologies, might improve classification accuracy. Regarding summarization, refining the prompt strategies and experimenting with different configurations could lead to more coherent and informative summaries. This may involve exploring various prompt templates, incorporating more context from dialogues, or applying chain-of-thought reasoning to extract relevant informa-

as opposed to a predicted class.

Run	Summarization
Run 1	The patient has been experiencing low back pain for the last eight years. The pain started after a fall in an ABC store. The patient has received only electrical stimulation and heat therapy from physical therapy. The pain has been continuous but at varying degrees. The patient has a child and fell four or five days ago, worsening the low back pain. The patient has seen another doctor for this issue and has a follow-up appointment scheduled.
Run 2	Fall in October 2007 resulted in low back pain for 8 years. Pregnancy in 2008 and low back pain since. She was evaluated by a specialist in early January 2009 and is scheduled to return in 2/10/09. Back pain began when she fell in an ABC store in fall 2007; had been good periodically since. She had back pain suddenly worsened when she fell 4/10/08. Since back in October 2008, most recently
Golden	The patient is a 26-year-old female, referred to Physical Therapy for low back pain. The patient has a history of traumatic injury to low back. The patient stated initial injury occurred eight years ago, when she fell at a ABC Store. The patient stated she received physical therapy, one to two visits and received modality treatment only, specifically electrical stimulation and heat pack per patient recollection. The patient stated that she has had continuous low-back pain at varying degrees for the past eight years since that fall. The patient gave birth in August 2008 and since the childbirth, has experienced low back pain. The patient also states that she fell four to five days ago, while mopping her floor. The patient stated that she landed on her tailbone and symptoms have increased since that fall. The patient stated that her initial physician examination with Dr. X was on 01/10/09, and has a followup appointment on 02/10/09.

Table 5: Example Summarizations: Run 1, Run 2, and Golden Summary Comparison

tion. Furthermore, fine-tuning the language model on a domain-specific corpus or using multi-task learning that incorporates related tasks, such as question-answering or information extraction, may contribute to better summarization performance. Finally, the evaluation metrics used in this study may not fully capture the quality of the generated summaries. It is important to acknowledge that automated evaluation metrics, like ROUGE-1, BERTScore F1, and BLEURT, might not be fully aligned with human judgments. Therefore, conducting user studies with medical professionals could provide valuable insights into the utility and accuracy of the generated summaries in real-world clinical settings.

6 Conclusion

Our study demonstrates the effectiveness of combining traditional machine learning techniques, such as SVM, with advanced language models, like GPT-3.5, for medical dialogue summarization. This hybrid methodology has the potential to improve documentation procedures during patient care and facilitate informed decision-making for healthcare professionals by classifying medical dialogues and generating concise summaries.

For future work, we plan to address the limitations identified in this study. For classification, we will experiment with model configurations and explore alternative machine learning algorithms. For summarization, we will refine prompt strategies, incorporate domain-specific knowledge, and investigate various fine-tuning techniques. Lastly, conducting user studies with medical professionals will provide valuable feedback to assess the utility and accuracy of our generated summaries in real-world clinical settings and further refine our approach.

References

N. S. Altman. 1992. An introduction to k-nearest neighbour classification. *Journal of Classification*, 9(1):1–27.

Asma Ben Abacha, Wen wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen. 2023a. Overview of the mediqa-chat 2023 shared tasks on the summarization and generation of doctor-patient conversations. In *ACL-ClinicalNLP 2023*.

Asma Ben Abacha, Wen wai Yim, Yadan Fan, and Thomas Lin. 2023b. An empirical study of clinical note generation from doctor-patient encounters. In *EACL 2023*.

Leo Breiman. 1984. Classification and regression trees. *Wadsworth International Group*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020b. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Xiaoyan Cai, Sen Liu, Libin Yang, Yan Lu, Jintao Zhao, Dinggang Shen, and Tianming Liu. 2022. Covidsum: A linguistically enriched scibert-based summarization model for covid-19 scientific papers. *Journal of Biomedical Informatics*, 127:103999 – 103999.

Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2021. Medically aware gpt-3 as a data generator for medical dialogue summarization. In *Proceedings of the 6th Machine Learning for Healthcare Conference*, volume 149 of *Proceedings of Machine Learning Research*, pages 354–372. PMLR.

Tom Christensen and Anders Grimsmo. 2008. Instant availability of patient records, but diminished availability of patient information: a multi-method study of gp's use of electronic patient records. *BMC medical informatics and decision making*, 8(1):1–8.

Corinna Cortes and Vladimir Vapnik. 1995. *Support Vector Networks*. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sohrab Ferdowsi, Nikolay Borissov, Julien Knafou, Poorya Amini, and Douglas Teodoro. 2021. Classification of hierarchical text using geometric deep learning: the case of clinical trials corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 608–618.

Sohrab Ferdowsi, Jenny Copara, Racha Gouareb, Nikolay Borissov, Fernando Jaume-Santero, Poorya Amini, and Douglas Teodoro. 2022. On graph construction for classification of clinical trials protocols using graph neural networks. In *Artificial Intelligence in Medicine: 20th International Conference on Artificial Intelligence in Medicine, AIME 2022, Halifax, NS, Canada, June 14–17, 2022, Proceedings*, pages 249–259. Springer.

Sohrab Ferdowsi, Julien Knafou, Nikolay Borissov, David Vicente Alvarez, Rahul Mishra, Poorya Amini, and Douglas Teodoro. 2023. Deep learning-based risk prediction for interventional clinical trials based on protocol design: A retrospective study. *Patterns*, 4(3).

Som Gupta and Sanjai Kumar Gupta. 2019. Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications*, 121:49–65.

Jochen Hartmann, Juliana Huppertz, Christina Schamp, and Mark Heitmann. 2019. Comparing automated text classification methods. *International Journal of Research in Marketing*, 36(1):20–38.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Quincy John and Pat Langley. 1995. Estimating continuous distributions in bayesian classifiers. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345.

A Joshi, N Katariya, X Amatriain, and A Kannan. 2020a. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3755–3763, Stroudsburg, PA, USA. Association for Computational Linguistics.

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020b. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3755–3763, Online. Association for Computational Linguistics.

Ben-Tzion Karsh, Richard J Holden, Samuel J Alper, and CKL Or. 2006. A human factors engineering paradigm for patient safety: designing to support the performance of the healthcare professional. *BMJ Quality & Safety*, 15(suppl 1):i59–i65.

Julien Knafou, Quentin Haas, Nikolay Borissov, Michel Counotte, Nicola Low, Hira Imeri, Aziz Mert Ipekci, Diana Buitrago-Garcia, Leonie Heron, Poorya Amini, et al. 2023. Ensemble of deep learning language models to support the creation of living systematic reviews for the covid-19 literature. *bioRxiv*, pages 2023–01.

Huan Yee Koh, Jiaxin Ju, Ming Liu, and Shirui Pan. 2022. An empirical survey on long document summarization: Datasets, models, and metrics. *ACM Comput. Surv.*, 55(8).

Panagiotis Kouris, Georgios Alexandridis, and Andreas Stafylopatis. 2021. Abstractive text summarization: Enhancing sequence-to-sequence models using word sense disambiguation and semantic content generalization. *Computational Linguistics*, 47(4):813–859.

K Krishna, S Khosla, J Bigham, and ZC Lipton. 2021. Generating soap notes from doctor-patient conversations using modular summarization techniques. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4958–4972, Stroudsburg, PA, USA. Association for Computational Linguistics.

Md Tahmid Rahman Laskar, Enamul Hoque, and Jimmy Xiangji Huang. 2022. Domain adaptation with pre-trained transformers for query-focused abstractive text summarization. *Computational Linguistics*, 48(2):279–320.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Tinghuai Ma, Qian Pan, Huan Rong, Yurong Qian, Yuan Tian, and Najla Al-Nabhan. 2022. T-bertsum: Topic-aware text summarization based on bert. *IEEE Transactions on Computational Social Systems*, 9(3):879–890.

Juan Manuel and Torres Moreno. 2014. Automatic text summarization. *DOI*, 10:9781119004752.

C J McDonald. 1976. Protocol-based computer reminders, the quality of care and the non-perfectability of man. *N Engl J Med*, 295(24):1351–1355.

Clement J McDonald, Fiona M Callaghan, Arlene Weissman, Rebecca M Goodwin, Mallika Mundkur, and Thomson Kuhn. 2014. Use of internist's free time by ambulatory care electronic medical record systems. *JAMA internal medicine*, 174(11):1860–1863.

Hans Moen, Laura-Maria Peltonen, Juho Heimonen, Antti Airola, Tapio Pahikkala, Tapio Salakoski, and Sanna Salanterä. 2016. Comparison of automatic summarisation methods for clinical free text notes. *Artificial Intelligence in Medicine*, 67:25–37.

S Molenaar, L Maas, V Burriel, F Dalpiaz, and S Brinkkemper. 2020. *Medical Dialogue Summarization for Automated Reporting in Healthcare*, pages 76–88.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. *Proceedings of the 22nd International Conference on Machine Learning*, pages 625–632.

Jihad S Obeid, Paul M Heider, Erin R Weeda, Andrew J Matuskowitz, Christine M Carr, Kevin Gagnon, Tami Crawford, and Stephane M Meystre. 2019. Impact of De-Identification on clinical text classification using traditional and deep learning classifiers. *Stud Health Technol Inform*, 264:283–287.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Rimma Pivovarov and Noémie Elhadad. 2015. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association*, 22(5):938–947.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Mayank Ramina, Nihar Darnay, Chirag Ludbe, and Ajay Dhruv. 2020. Topic level summary generation using bert induced abstractive summarization model. In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 747–752.

Herbert Robbins and Sutton Monro. 1951. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

George E. Sakr, Maria Mokbel, Ahmad Darwich, Mia Nasr Khneisser, and Ali Hadi. 2016. Comparing deep learning and support vector machines for autonomous waste sorting. In *2016 IEEE International Multidisciplinary Conference on Engineering Technology (IMCET)*, pages 207–212.

Gerard Salton and Chris Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

Tariq Sellam, Colin Raffel, Wei Liu, and Ashish Vaswani. 2020. Bleurt: Learning robust metrics for text generation. In *International Conference on Learning Representations*.

WILLIAM W Stead and H Lin. 2009. Committee on engaging the computer science research community in health care informatics. *Computational technology for effective health care: immediate steps and strategic directions*.

Ming-Hsiang Su, Chung-Hsien Wu, and Hao-Tse Cheng. 2020. A two-stage transformer-based approach for variable-length abstractive summarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2061–2072.

Douglas Teodoro, Julien Gobeill, Emilie Pasche, P Ruch, and D Vishnyakova. 2010. Automatic ipc encoding and novelty tracking for effective patent mining. In *The 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access*.

Douglas Teodoro, Julien Knafou, Nona Naderi, Emilie Pasche, Julien Gobeill, Cecilia N Arighi, and Patrick Ruch. 2020. Upclass: a deep learning-based classifier for uniprotkb entry publications. *Database*, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010. Curran Associates Inc.

Mengqian Wang, Manhua Wang, Fei Yu, Yue Yang, Jennifer Walker, and Javed Mostafa. 2021. A systematic review of automatic text summarization for biomedical literature and ehrs. *Journal of the American Medical Informatics Association*, 28(10):2287–2297.

Zhengjue Wang, Zhibin Duan, Hao Zhang, Chaojie Wang, Long Tian, Bo Chen, and Mingyuan Zhou. 2020. Friendly topic assistant for transformer based abstractive summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 485–497, Online. Association for Computational Linguistics.

Adhika Pramita Widyassari, Supriadi Rustad, Guruh Fajar Shidik, Edi Noersasongko, Abdul Syukur, Affandy Affandy, et al. 2022. Review of automatic text summarization techniques & methods. *Journal of King Saud University-Computer and Information Sciences*, 34(4):1029–1046.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

L Zhang, R Negrinho, A Ghosh, et al. 2021. Leveraging pretrained models for automatic summarization of doctor-patient conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3693–3712, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tianyi Zhang and See-Kiong Ng. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

GersteinLab at MEDIQA-Chat 2023: Clinical Note Summarization from Doctor-Patient Conversations through Fine-tuning and In-context Learning

Xiangru Tang[♠] **Andrew Tran**[♠] **Jeffrey Tan**[♠] **Mark Gerstein**[♠]
[♠] Yale University, New Haven, CT 06520, USA

{xiangru.tang, a.tran, jeffrey.tan, mark.gerstein}@yale.edu

Abstract

This paper presents our contribution to the MEDIQA-2023 Dialogue2Note shared task, encompassing both subtask A and subtask B. We approach the task as a dialogue summarization problem and implement two distinct pipelines: (a) a fine-tuning of a pre-trained dialogue summarization model and GPT-3, and (b) few-shot in-context learning (ICL) using a large language model, GPT-4. Both methods achieve excellent results in terms of ROUGE-1 F1, BERTScore F1 (deberta-xlarge-mnli), and BLEURT, with scores of 0.4011, 0.7058, and 0.5421, respectively. Additionally, we predict the associated section headers using RoBERTa and SciBERT based classification models. Our team ranked fourth among all teams, while each team is allowed to submit three runs as part of their submission. We also utilize expert annotations to demonstrate that the notes generated through the ICL GPT-4 are better than all other baselines. The code for our submission is available [1].

1 Introduction

The field of medical AI has witnessed significant advancements in recent years, fueled by its promise to transform clinical documentation procedures (Beltagy et al., 2019; Alsentzer et al., 2019; Huang et al., 2019; Si et al., 2019; Lee et al., 2020; Gu et al., 2021). Extracting clinical notes from doctor-patient interactions is a crucial aspect of maintaining medical records, as it fosters effective communication among healthcare practitioners. By automating this process, healthcare professionals can shift their focus toward patient care and minimize the time dedicated to administrative duties (Jain et al., 2022; Navarro et al., 2022). The development of efficient and accurate algorithms for summarizing these conversational notes is therefore of paramount importance, as it has the potential to improve overall healthcare quality and efficiency (Quiroz et al., 2020; Krishna et al., 2021; Menon et al., 2021; Michalopoulos et al., 2022; Tang et al., 2023b).

The MEDIQA-Chat 2023 challenge[2] was established to promote the development of novel summarization techniques, specifically targeting the automatic generation of clinical notes from doctor-patient conversations (Ben Abacha et al., 2023a). The Dialogue2Note and Note2Dialogue shared tasks are designed to stimulate research and innovation in this field, addressing the summarization of medical conversations for clinical note creation and the generation of synthetic doctor-patient conversations for data creation and augmentation. The Dialogue2Note Summarization task entails converting a doctor-patient conversation into a clinical note containing one or multiple note sections, such as Assessment, Past Medical History, or Past Surgical History. This task is subdivided into two subtasks: (A) generating specific sections from conversations (Ben Abacha et al., 2023b), and (B) generating complete notes from conversations (wai Yim et al., 2023), two examples are shown in Figure. 1.

In this paper, we discuss our submission to both subtask A and subtask B of the shared task: For **subtask A**, we first focused on section classification and explored two methods: (1) using RoBERTa (Liu et al., 2019) and SciBERT (Beltagy et al., 2019) with a classification head, and (2) fine-tuning OpenAI's Davinci model [3]. Subsequently, we investigated generating specific sections using a fine-tuned pre-trained dialogue summarization model. We employed the CONFIT model (Tang et al., 2022b), which proposes a training strategy that enhances the factual consistency and overall quality of summaries through a

[1] https://github.com/gersteinlab/MEDIQA-Chat-2023

[2] https://sites.google.com/view/mediqa2023/clinicalnlp-mediqa-chat-2023

[3] https://platform.openai.com/docs/guides/fine-tuning

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 546–554
July 14, 2023

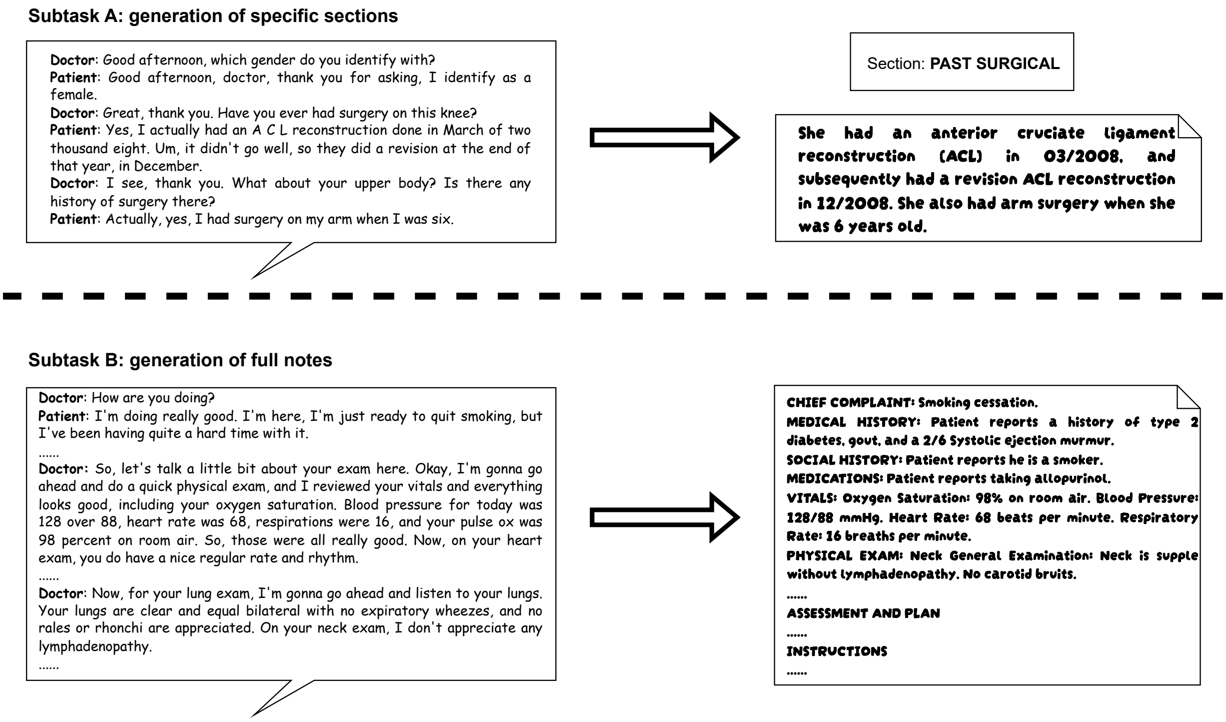


Figure 1: Two examples of our tasks, which include both subtask A and subtask B. In subtask A, the goal is to generate a summary and the corresponding section name for a specific section, while subtask B aims to generate a complete note.

novel contrastive fine-tuning approach. For **subtask B**, we examined how to utilize large language models (LLMs) like GPT. We (1) fine-tuned OpenAI's Davinci model and (2) explored in-context learning (Dong et al., 2022) with GPT-4 [4]. We achieved promising results on automated metrics (ROUGE, BERTScore (Zhang et al., 2019), and BLEURT (Sellam et al., 2020)), and our outcomes were also assessed manually. Although the GPT-based model scored slightly lower on automated metrics, it received high scores in human evaluations. We believe that for zero-shot models, existing automated metrics may not be the most appropriate evaluation method, suggesting a potential direction for future research.

2 Tasks

2.1 Task Formulation

In this paper, we focus solely on the Dialogue2Note Summarization task of MEDIQA-Chat Tasks @ ACL-ClinicalNLP 2023. The main tasks include:

- **Dialogue2Note Summarization**: Given a doctor-patient conversation, participants are required to generate a clinical note summarizing the conversation, including one or multiple note sections (e.g., Assessment, Past Medical History, Past Surgical History). This task comprises two subtasks:
  - **Subtask A**: Generating specific sections from doctor-patient conversations (Ben Abacha et al., 2023b).
  - **Subtask B**: Generating full notes from doctor-patient conversations (wai Yim et al., 2023).
- **Note2Dialogue Generation**: Participants are tasked with generating a synthetic doctor-patient conversation based on the information described in a given clinical note (wai Yim et al., 2023).

For **subtask A**, the training set consists of 1,201 pairs of conversations and associated section headers and contents, while the validation set includes 100 pairs of conversations and their summaries. A full list of normalized section headers is provided in the paper.

As for **subtask B**, the training set is composed of 67 pairs of conversations and full notes, and the

[4]https://openai.com/research/gpt-4

validation set includes 20 pairs of conversations and clinical notes.

Lastly, the **Note2Dialogue Generation task**'s training set comprises 67 pairs of full doctor-patient conversations and notes, with the validation set containing 20 pairs of full conversations and clinical notes. The Task-A training and validation sets (1,301 pairs) could be used as additional training data.

Thus, we could formally define the tasks as follows. Given a doctor-patient conversation $C = \{c_1, c_2, ..., c_n\}$, where c_i represents the ith utterance in the conversation and n denotes the total number of utterances, the goal of the Dialogue2Note Summarization task is to generate a clinical note summarizing the conversation.

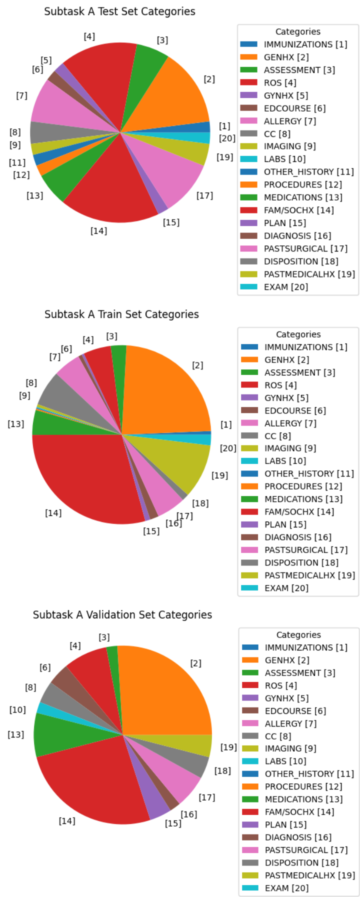


Figure 2: The proportion of section categories in subtask A.

For **subtask A**, the objective is to generate a specific section summary S_j and its corresponding section header H_j for a given conversation C. The output can be represented as a tuple (H_j, S_j). For **subtask B**, the goal is to generate a complete clinical note $N = \{S_1, S_2, ..., S_m\}$, where S_i represents the ith summary section and m denotes the total number of sections. Each section S_i is associated with a section header H_i. We use a combination of various evaluation metrics such as ROUGE, BERTScore, and BLEURT.

2.2 Data Analysis

For the section classification task in Subtask A, we created pie charts (See Figure. 2) representing the proportions of different sections in the train, test, and validation sets to analyze the distribution differences among them. We observed that there is no significant gap in the section categories across the data splits. However, there is a considerable disparity in the number of instances among different categories 3, with some sections having very few data points. This may lead to insufficient training and poor performance for those underrepresented categories.

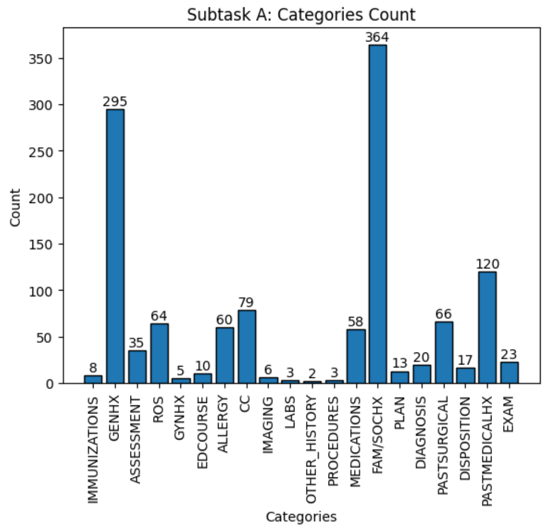


Figure 3: The numbers of section categories in subtask A. Here are the totals on the train and validation sets.

In Figure. 4, we illustrate the length of input dialogues and the length of output summaries. And Figure. 5 shows the number of utterances. In the picture. 6, the length of the dialogue is plotted against the length of the summary for each data entry. The graphs show a noticeable positive correlation, thus indicating that longer dialogues do have significantly longer summaries. Additionally, we can see that dialogue lengths increase at about twice the rate of summary lengths, so our summaries should be about one-third to one-half the dialogue length.

3 Methodology

In this section, we will describe how we employed various approaches to perform subtask A and subtask B. We will discuss each subtask in detail, outlining the methods used and the rationale behind our choices to achieve optimal results.

3.1 Subtask A

We define this task as a dialogue summarization problem and, therefore, we selected a state-of-the-art dialogue summarization model, CONFIT (Tang et al., 2022b), as our foundation to fine-tune. Firstly, CONFIT is based on BART and has been fine-tuned on the SAMSum dialogue summarization dataset. We utilized the model fine-tuned on SAMSum and further fine-tuned it on MEDIQA subtask A data to generate notes. It is worth noting that ideally, we should have added an additional step, fine-tuning the model on PubMed summarization data before fine-tuning on subtask A data, which would enable the model to better understand the clinical summarization task. We set max-input-length to 1024 and keep output max-length at 128. For generating section names, we

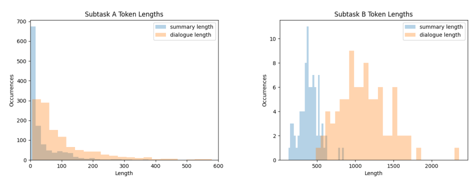


Figure 4: Histogram of token lengths for subtask A train and validation sets in subtask A and B.

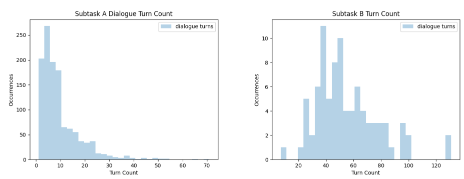


Figure 5: Histogram of utterance numbers for subtask A train and validation sets in subtask A and B.

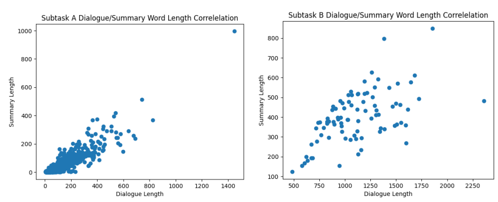


Figure 6: Dialogue Length plotted against Summary Length for each data entry. The graphs show a noticeable positive correlation.

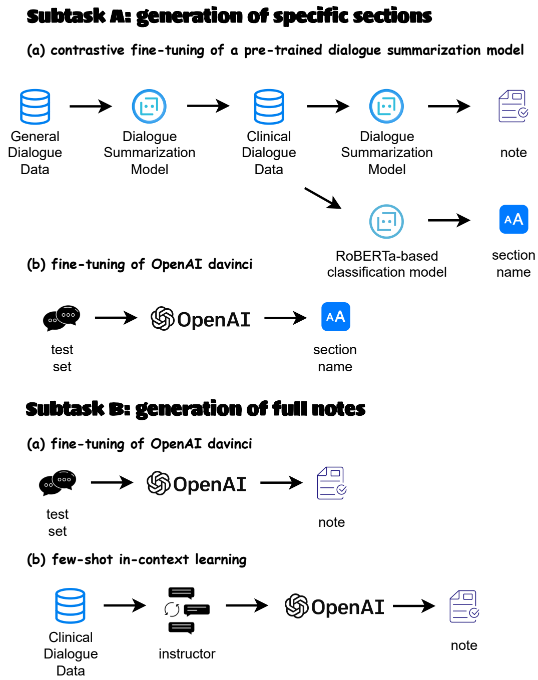


Figure 7: This diagram illustrates the pipeline we employed for both subtask A and subtask B. For subtask A, we fine-tune a dialogue summarization model to generate notes, and to generate section names, we use a RoBERTa model and fine-tuning of OpenAI's Davinci model. For subtask B, we perform fine-tuning as well as few-shot in-context learning to achieve our desired results.

used a RoBERTa model and fine-tuning of OpenAI's Davinci model. We fine-tuned RoBERTa to classify text based on 20 predefined categories. Additionally, we invoked OpenAI's API and used a customized model (davinci:ft-personal-2023-03-15-08-22-14) to classify the dialogues. This model classifies text according to the 20 predefined categories, with the classification implemented through the OpenAI API.

3.2 Subtask B

We explored two approaches for utilizing OpenAI's large-scale language models. First, we defined a function to shorten the dialogue, ensuring it does not exceed the maximum token length of 1200. Next, we used another function to call OpenAI's API and employ a customized model (davinci:ft-personal-2023-03-23-05-58-11) to generate the corresponding notes. We fine-tuned the Davinci model and, during this process, implemented multiple generation attempts to adjust the maximum token length, setting max tokens to 800. We experimented with temperatures of 0.0 and 0.2. We used

Models	R_1	R_2	R_L	R_{Lsum}	$BERTScore_p$	$BERTScore_r$	$BERTScore_{f1}$	BLEURT
CONFIT	0.3882	0.1966	0.3214	0.3214	0.7037	0.7065	0.7	0.5294
$CONFIT_{dynamic}$	0.4011	0.2147	0.3322	0.3322	0.7115	0.7102	0.7058	0.5421

Table 1: Comparison of performance between the CONFIT model and the CONFIT model with dynamic max length on various evaluation metrics.

the following prompt: "Please summarize the following dialog between doctors and patients from the perspective of the doctor, and be sure to include all important details about the patient."

Moreover, we employed in-context learning, choosing GPT-4 [5] and designing a prompt that included natural language instructions and context examples. We limited the input prompt length to 6000 tokens and the output length to 2000 tokens. We used two contexts per instance and set the temperature parameter to 0.2. In our prompt template, we incorporate three main components: instructions, in-context examples, and test input dialogue. The final "FULL NOTE:" indicates the output the model needs to generate. Our instructions are as follows: "Write a clinical note for this doctor-patient dialogue. Use the example notes below to know the different sections." This guides the model to generate a clinical note based on the given doctor-patient dialogue, taking into account the structure and sections observed in the provided examples.

4 Results

We report the accuracy of the section classification. For note generation in subtask A, we present the following metrics: Rouge-1, Rouge-2, Rouge-L, Rouge-Lsum, BERTScore precision (Zhang et al., 2019), BERTScore recall (Zhang et al., 2019), BERTScore F1 (Zhang et al., 2019), BLEURT (Sellam et al., 2020), and aggregate score. The aggregate score is the arithmetic mean of ROUGE-1 F1, BERTScore F1, and BLEURT-20 (Pu et al., 2021).

Additionally, we implemented a dynamic max length feature.For the CONFIT model and the CONFIT model with dynamic max length, we obtained aggregate scores of 0.5392 and 0.5497, respectively. The results demonstrate that the dynamic max length approach slightly improves the overall performance of the model in the summarization task. Initially, it encodes the input dialogue into a vector using a pre-trained tokenizer. Subsequently, it dynamically calculates the maximum length of the summary to be generated, based on the length of the dialogue. The formula employed here is $max_{length} = round(0.55 * dialog_{len} + 18)$, which determines the maximum length of the summary according to the dialogue's word count. Then, the BART model generates the summary while controlling the max length of the output. Ultimately, the generated summary is decoded into text.

[5] https://platform.openai.com/docs/models/gpt-4

Models	Accuracy
OpenAI davinci	0.745
SciBERT	0.710
RoBERTa	0.700

Table 2: Accuracy of section classification for different models.

The results in Table 1 show a comparison between the performance of the CONFIT model and the CONFIT model with dynamic max length on various evaluation metrics, including ROUGE scores (R_1, R_2, R_L, and R_{Lsum}), BERTScore ($BERTScore_p$, $BERTScore_r$, and $BERTScore_{f1}$), and BLEURT. The $CONFIT_{dynamic}$ model achieves better performance on most evaluation metrics, suggesting that the incorporation of dynamic max length improves the overall quality of the generated summaries.

In Table 2, we present the accuracy of section classification for three different models, OpenAI davinci, SciBERT, and RoBERTa. The task involves classifying the sections into one of the 20 pre-defined categories. As shown in the table, OpenAI davinci achieves the highest accuracy of 0.745, outperforming both SciBERT and RoBERTa, which achieve accuracies of 0.710 and 0.700, respectively. This indicates that the OpenAI davinci model is more effective in classifying sections in this 20-class classification task. In addition to the RoBERTa classifier, we also utilized SciBERT to improve the initial classification performance. By incorporating the domain-specific knowledge embedded in SciBERT, we were able to enhance the accuracy of our section classification task.

Table 3 presents the evaluation of the ROUGE

Models	R_1	R_2	R_L	R_{Lsum}
ICL	0.5821	0.3209	0.4032	0.5443
Davinci (t=0.0)	0.5008	0.2506	0.3282	0.4668
Davinci (t=0.2)	0.5004	0.2502	0.3249	0.4675

Table 3: The evaluation of the ROUGE scores for subtask 2 full note summarization.

Models	Score
Subtask A BART	7.2
Subtask A T5	6.6
Subtask A PEGASUS	7.0
Subtask A CONFIT	7.2
Subtask A $\text{CONFIT}_{\text{dynamic}}$	7.8
Subtask B BART	3.5
Subtask B T5	3.8
Subtask B PEGASUS	3.2
Subtask B ICL	6.6
Subtask B Davinci (t=0.0)	5.3
Subtask B Davinci (t=0.2)	5.2

Table 4: Expert annotation of generated notes in a scale of 1-10 on subtask A and B.

scores for subtask 2 full note summarization. The table includes the ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum scores for each model. Higher ROUGE scores indicate better summarization quality, reflecting the extent to which the generated summaries capture the important information and coherence of the full notes. Except for our submission, we expanded our approach by incorporating in-context learning. As the capabilities of large language models (LLMs) continue to advance, in-context learning (ICL) has emerged as a novel paradigm in the field of natural language processing (NLP). In ICL, LLMs make predictions based on a limited set of examples, augmented with additional context. This approach allows LLMs to leverage the power of context in order to enhance their predictive abilities. By incorporating contextual information during the learning process, LLMs are able to generate more accurate and contextually relevant predictions. In the evaluation, the in-context model achieves the highest ROUGE scores, indicating that it generates summaries that have a higher overlap and alignment with the reference summaries of the full notes. Comparatively, the ”Davinci” model with a temperature of 0.0 achieves lower ROUGE scores, but still performs better than the ”Davinci” model with a temperature of 0.2 in terms of ROUGE scores.

Based on previous studies that have shown that automated metrics are not suitable for evaluating the results generated by zero-shot models (Goyal et al., 2022), we sought alternative evaluation methods. To address this, we enlisted the expertise of three medical students to manually rate 50 note summarization outputs. Following established practices (Tang et al., 2022a), we employed a scoring system of 1-10 to assess the quality of the generated summaries. This expert evaluation provides a more comprehensive and reliable assessment of the performance of our models in Subtask A and Subtask B, capturing nuanced aspects that automated metrics may not capture accurately. By incorporating expert assessments, we aim to enhance the evaluation process and gain deeper insights into the capabilities and limitations of our zero-shot models in the context of note summarization. In this study, we also implemented traditional fine-tuned summarization models such as BART (Lewis et al., 2019), T5 (Raffel et al., 2020), and PEGASUS (Zhang et al., 2020) as baselines. However, it is important to note that due to the absence of a test set reference, we were unable to compute automated metrics for evaluation. Therefore, we solely rely on the results of the manual assessments conducted by human evaluators. Although automated metrics are commonly used to evaluate summarization models, the absence of a reference necessitates a shift towards expert judgments to assess the quality and effectiveness of the generated summaries. Our focus is on reporting the outcomes of the manual evaluations as a reliable measure of the performance of the models.

The results demonstrate that, in subtask A, CONFIT outperforms other baseline methods. In subtask B, it is evident that traditional models do not perform well, and we believe the main difference lies in the input length. As all the examples exceed the maximum input length limitation of traditional models, and the reference also significantly surpasses the maximum length limitation for generation, it is naturally challenging to generate ideal notes. Therefore, in terms of human evaluation, both fine-tuning OpenAI models and utilizing ICL perform far better than using traditional fine-tuned models. The superior performance is likely attributed to OpenAI's models, which provide longer input and output limitations.

5 Limitations and Discussion

We would like to discuss some of our findings and thoughts here. The MEDIQA-Chat 2023 challenge indeed presents an excellent opportunity for us to reflect on existing models, analyze their strengths and weaknesses, and investigate their performance on the task of clinical note generation from doctor-Patient conversations.

Evaluation Automatic metrics often do not have a strong correlation with the quality of summaries, especially as there are no readily available automatic evaluation metrics specifically designed for zero-shot or few-shot LLM evaluation. Almost all models on Subtask A have similar results of automatic evaluation. Therefore, human evaluation becomes an essential component for accurately assessing the quality of notes generated by LLMs. There is a need to develop more effective automatic evaluation methods for a zero-shot/few-shot generation or super-long output.

Length Limitation of Subtask A One significant limitation of our Transformer-based model is that it does not directly consider length during its generation process. This often results in the production of overly verbose summaries. In contrast, models like OpenAI's GPT-3 and GPT-4 have much longer input and output limitations, allowing them to handle more extensive text samples more effectively. It is worth noting that our model was trained on the SAMSum dataset, which has longer texts compared to subtask A. Consequently, our model struggles to adapt to the shorter length requirements of subtask A. Moreover, the training dataset for subtask A is relatively small, which further complicates the model's adaptation. Future exploration should look at how to constrain the conciseness of generated summaries, which may involve reconsidering the generation method chosen or examining other techniques to promote brevity. Developing methods to better control the length of generated summaries is essential to improve their relevance, coherence, and usability in real-world applications.

Length Limitation of Subtask B For Subtask B, it is challenging to achieve reasonable results using fine-tuned models. In reality, this task is more representative of real-world scenarios, where inputs and outputs are considerably long, and the output is expected to maintain a specific structure and format. Thus, we see the main advantage of using contextual examples lies in their ability to guide the structure, style, and length of the desired output. We believe that OpenAI's LLM is well-suited for similar real-life scenarios, provided that it is given an appropriate context. In such cases, its performance will significantly surpass that of fine-tuned Transformer-based models.

Factual inconsistentcy While our study did not specifically investigate the following issues, we noted several factual errors that occur in summaries. A previous study has shown that LLMs also exhibit a noticeable occurrence of attribute errors and misinterpretation errors (Tang et al., 2023a).

Prompting We find that prompt template and demonstration example selection both have a substantial impact on results. Using more prompt examples for demonstration improves significantly. We acknowledge that we did not explore different selection strategies, such as SemScore, LMScore, and TLength, which involve using top-ranked examples. These strategies have been shown to potentially improve the quality of the generated summaries by selecting more effective prompt examples. While our current approach did not incorporate these strategies, we recognize that exploring and incorporating better prompt examples could potentially yield improved results. This is an area that warrants further investigation and experimentation to enhance the performance of our models in future iterations of the study.

Data Privacy Both GPT-3 and GPT-4 are not local models; we utilize OpenAI's API to run these models, which actually violates data protection laws such as HIPAA. Ensuring data privacy during fine-tuning or testing is of paramount importance. We have not taken this aspect into consideration.

6 Conclusion

We have showcased our solution, submitted to the MEDIQA-Chat shared task, designed to generate clinical notes from doctor-patient dialogues. Our evaluation encompassed fine-tuned approaches utilizing models such as CONFIT, GPT-3, RoBERTa, and SciBERT, as well as an innovative method employing GPT-4. The results garnered are remarkable, evidenced by our team securing a fourth-place ranking amongst all participating teams, underscoring the efficacy of our methodologies. Of particular note is the fact that expert annotations were deployed to substantiate the comparative superiority of the ICL GPT-4 over other baseline models. This infusion of expert validation fortifies the credibility of our research and its ensuing findings, contributing a robust dimension of trustworthiness to our results.

7 Limitations and Future Work

For task A, the CONFIT model was utilized with dynamic length outputs, while task B placed significant emphasis on prompting-based techniques. The method hinges on API calls to OpenAI models, which are not open-source. This dependency could potentially induce instability.

We surmise that performance may have been further enhanced if initial training on PubMed data had preceded fine-tuning on the given dataset. Future endeavors could aim to provide theoretical substantiation or engage in ablation studies to affirm this hypothesis.

In task B, we employed a function to condense dialogues. However, due to time constraints, we did not explore the potential impact of this condensation on the performance, specifically, whether any information loss incurred during the shortening process might affect the results.

The results appear promising, particularly with the employment of CONFIT models with dynamic max length. This suggests an intriguing avenue for future exploration and research.

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Asma Ben Abacha, Wen wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen. 2023a. Overview of the mediqa-chat 2023 shared tasks on the summarization and generation of doctor-patient conversations. In *ACL-ClinicalNLP 2023*.

Asma Ben Abacha, Wen wai Yim, Yadan Fan, and Thomas Lin. 2023b. An empirical study of clinical note generation from doctor-patient encounters. In *EACL 2023*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Raghav Jain, Anubhav Jangra, Sriparna Saha, and Adam Jatowt. 2022. A survey on medical document summarization. *arXiv preprint arXiv:2212.01669*.

Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary C. Lipton. 2021. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4958–4972, Online. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Niharika G Menon, Avani Shrivastava, ND Bhavana, and Judy Simon. 2021. Deep learning based transcribing and summarizing clinical conversations. In *2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, pages 358–365. IEEE.

George Michalopoulos, Kyle Williams, Gagandeep Singh, and Thomas Lin. 2022. Medicalsum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4741–4749.

David Fraile Navarro, Mark Dras, and Shlomo Berkovsky. 2022. Few-shot fine-tuning sota summarization models for medical dialogues. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 254–266.

Juan C Quiroz, Liliana Laranjo, Ahmet Baki Kocaballi, Agustina Briatore, Shlomo Berkovsky, Dana Rezazadegan, and Enrico Coiera. 2020. Identifying relevant information in medical conversations to summarize a clinician-patient encounter. *Health Informatics Journal*, 26(4):2906–2914.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Yuqi Si, Jingqi Wang, Hua Xu, and Kirk Roberts. 2019. Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, 26(11):1297–1304.

Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin Rousseau, et al. 2023a. Evaluating large language models on medical evidence summarization. *medRxiv*, pages 2023–04.

Xiangru Tang, Arman Cohan, and Mark Gerstein. 2023b. Aligning factual consistency for clinical studies summarization through reinforcement learning. In *ACL-ClinicalNLP 2023*.

Xiangru Tang, Alexander Fabbri, Haoran Li, Ziming Mao, Griffin Adams, Borui Wang, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2022a. Investigating crowdsourcing protocols for evaluating the factual consistency of summaries. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5680–5692, Seattle, United States. Association for Computational Linguistics.

Xiangru Tang, Arjun Nair, Borui Wang, Bingyao Wang, Jai Desai, Aaron Wade, Haoran Li, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2022b. CONFIT: Toward faithful dialogue summarization with linguistically-informed contrastive fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5657–5668, Seattle, United States. Association for Computational Linguistics.

Wen wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. The aci demo corpus: An open dataset for benchmarking the state-of-the-art for automatic note generation from doctor-patient conversations. *Submitted to Nature Scientific Data*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Author Index