Efficient Biomedical Entity Linking: Clinical Text Standardization with Low-Resource Techniques

Akshit Achara * **Sanand Sasidharan** † **Gagan N** ‡

GE Healthcare

Abstract

Clinical text is rich in information, with mentions of treatment, medication and anatomy among many other clinical terms. Multiple terms can refer to the same core concepts which can be referred as a clinical entity. Ontologies like the Unified Medical Language System (UMLS) are developed and maintained to store millions of clinical entities including the definitions, relations and other corresponding information. These ontologies are used for standardization of clinical text by normalizing varying surface forms of a clinical term through Biomedical entity linking. With the introduction of transformer-based language models, there has been significant progress in Biomedical entity linking. In this work, we focus on learning through synonym pairs associated with the entities. As compared to the existing approaches, our approach significantly reduces the training data and resource consumption. Moreover, we propose a suite of context-based and context-less reranking techniques for performing the entity disambiguation. Overall, we achieve similar performance to the state-of-the-art zero-shot and distant supervised entity linking techniques on the Medmentions dataset, the largest annotated dataset on UMLS, without any domain-based training. Finally, we show that retrieval performance alone might not be sufficient as an evaluation metric and introduce an article level quantitative and qualitative analysis to reveal further insights on the performance of entity linking methods.

1 Introduction and Related Work

Medical text consists of a diverse vocabulary derived from various nomenclatures including varying surface forms corresponding to terms like diagnosis, treatment, medications, etc. This diversity poses a challenge for effective communication across medical institutions and organizations. One of the techniques to mitigate this inherent diversity present in multiple references to the same term is entity linking. Entity linking is used to map these references to standardized codes. These codes are curated and maintained by medical organizations for standardization of medical nomenclature.

Given a corpus, *entity linking* includes the mapping of a mention m which is a span of k words, to an entity ϵ, where the entity belongs to a knowledge base such as Wikipedia. In the biomedical domain, the textual phrases are linked with the corresponding concepts from a knowledge base constructed using the medical ontologies like UMLS (Bodenreider, 2004), SNOMED (El-Sappagh et al., 2018), etc. The UMLS ontology comprises of a broad range of clinical entities along with rich information for each entity like synonyms, definitions, etc. Traditional approaches for entity linking, such as Support Vector Machines (Cristianini and Shawe-Taylor, 2000) and Random Forests (Breiman, 2001), rely heavily on hand-crafted features, thereby restricting generalization to diverse data. Neural networks have emerged as a prominent technique for entity linking due to their ability to learn semantic representations from textual data.

Alias matching based techniques like (Aronson, 2001; Neumann et al., 2019; Liu et al., 2020) have been proposed where an input mention is mapped to an alias associated with an entity in the knowledge-base. However, these techniques require large amount of training data. Contextualized entity linking approaches (Zhang et al., 2021) utilize the semantic similarity between contextualized mentions. This approach requires a list of entities in advance and includes distant-supervision on articles containing examples of these entities. Generating medical codes using large language models can be error prone (Soroush et al., 2024). In (Yuan et al., 2022b), the authors use a seq2seq model to map a

*akshit.achara@gehealthcare.com
†sanand.sasidharan@gehealthcare.com
‡gagan.n@gehealthcare.com

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 493–505

mention to its canonical entity name. This method is resource intensive and requires generation of synthetic examples for pretraining, utilizing entity definitions and synonyms. In (Kong et al., 2021), the authors propose a zero-shot entity linking approach by leveraging synonym and graph based tasks. However, the approaches require training samples from UMLS for both these tasks. Moreover, entity disambiguation has not been explored in the work.

Efficient student models like MiniLM (Wang et al., 2020) can be used to perform contrastive learning on synonyms of entities. This results in a significantly less embedding size (384) as compared to the approaches like SAPBERT (Liu et al., 2020) with an embedding size of 768. The predicted candidates in alias based techniques are ranked based on the cosine similarity score. However, there are ambiguous cases where multiple entities have similar scores for a common mention. Therefore, there is a requirement to disambiguate these candidates through reranking. Cross-Attention based reranking approaches utilize supervised training on the concatenated mention and candidate representations as inputs (Zhang et al., 2021). More recent approaches utilize homonym disambiguation (Garda and Leser, 2024) and have shown to improve the performance of autoregressive approaches like GenBioEL.

In comparison to the discussed techniques, we propose an efficient and low resource zero-shot biomedical entity linking approach along with a suite of disambiguation techniques. Furthermore, we introduce an article level similarity analysis to obtain further insights. This also allows us to conduct a qualitative analysis without manually going through all the articles manually.

Our contributions are as follows:

- **Data**: We show that the impact of training is negligible on a finetuned MiniLM model[1] as compared to the pretrained MiniLM model. Moreover, the pretrained MiniLM model when finetuned on all UMLS synonym pairs has worse performance than the all-MiniLM model.
- **Disambiguation** We show that reranking on entity-level semantic information provided in UMLS can be highly effective for entity disambiguation. We further propose a parametric reranking technique that is beneficial for alias-based entity linking solutions.
- **Evaluation** We propose a comprehensive evaluation of entity linking which utilizes the semantic representation of articles coupled with the strict matching and related matching of predicted and gold standard entities. This evaluation is used to highlight issues related to the annotation granularity, missing context and surface form bias (for abbreviations) without the need of going through all the articles.

[1] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

2 Datasets

In this work, we explore entity linking on the Medmentions (Mohan and Li, 2019) dataset which consists of titles and abstracts from 4392 English biomedical articles. These articles comprise of textual spans annotated with mentions of UMLS 2017AA entities. The dataset provides two versions: a full version containing 34724 unique entities and an st21pv version with 25419 unique entities, the latter being recommended by the authors for information retrieval. Further details about the dataset versions are discussed in Table 9 in (Kartchner et al., 2023).

2.1 Preprocessing

We replace the abbreviations with their corresponding full forms using Ab3p (Sohn et al., 2008). The abbreviation expansion using Ab3p has shown to significantly improve the entity linking performance across different approaches (Kartchner et al., 2023). Prior to creating synonym pairs for training, we remove all the suppressed entities, deleted entities and deprecated entities. Some deprecated entities have also been merged with other entities having a synonymous relation. We map these deprecated entities to the corresponding active entities with a synonymous relation.

	st21pv	full
merged	181	280
deleted	49	60
non-synonymous	226	348

Table 1: The table shows the details of Medmentions entities annotated with UMLS 2017AA version that are deprecated in UMLS 2023AB version.

Some annotations in Medmentions (prepared with UMLS 2017AA) are deprecated in the UMLS

2023AB (see details in table 1). Therefore, approaches utilizing UMLS 2023AB version may want to use an updated version of Medmentions. Furthermore, the prototype space (feature vector space) consisting of UMLS entities will have to be updated to the remove deprecated entities. This would help in avoiding deprecated entities to be predicted as candidates.

3 Methodology

In this work, we create a prototype vector space comprising of the encodings (feature vectors) associated with the canonical name of each entity in the UMLS ontology. To obtain meaningful encodings for constructing this prototype space, we train an encoder-based transformer (Vaswani et al., 2017) model on pairs of canonical names of entity synonyms. This is similar to the training approaches utilized in (Kong et al., 2021) and (Liu et al., 2020). The prototype space constructed using this trained model is used for performing semantic search, where the query encoding is obtained by passing the mention through the same model. This step is known as candidate generation. The candidate generation may lead to ambiguous results where multiple predicted entities have equal similarity scores. This is addressed through the reranking approaches discussed in section 3.3. Finally, we utilize both semantic similarity and retrieval performance for our quantitaive and qualitative evaluation. The comprehensive structure of our proposed approaches is depicted in the figure 1.

The following sub-sections discuss the individual components used in our work:

3.1 Training

We construct a training dataset by taking all the canonical names for each entity from UMLS and create pairs of canonical names corresponding to the same entity. Each pair is of the form (ϵ_i, ϵ_i^*), where ϵ_i^* represents the canonical name of a synonym of entity ϵ_i. The preprocessing steps are discussed in the section 2.1. We use this dataset to finetune a sentence-transformer (Reimers and Gurevych, 2019) model using Multiple Negatives Ranking loss (Henderson et al., 2017). We use MiniLM (Wang et al., 2020) which is a distilled version of BERT$_{BASE}$ model obtained using an effective knowledge distillation approach outperforming other lightweight models like TinyBERT and DistillBERT. We also utilize a finetuned all-MiniLM[2] model for training/finetuning on this dataset. The all-MiniLM model is obtained by training the MiniLM model on a 1B sentence pairs dataset using a contrastive learning objective. The corresponding MiniLM and all-MiniLM models trained/finetuned on k examples are hereafter referred as MiniEL$^*_k$ and MiniEL$_k$ respectively. For example, the all-MiniLM model finetuned on 10 pairs/examples is referred as MiniEL$_{10}$.

The Multiple Negatives Ranking loss function is defined as:

$$L(x, y, \theta) = \frac{1}{B} \sum_{j=1}^{B} \log P(y_j|x_j) \quad (1)$$

Here, θ represents the network parameters, (x, y) represents a pair of phrases and B represents the batch size. The parameters details for training are provided in section in Appendix in the section A.1.

3.2 Candidate Generation

A prototype space is prepared for the UMLS 2017AA version comprising of the encodings of canonical names of each entity and its synonyms. These encodings are computed using the MiniEL* and MiniEL models. The prototype space is used for performing semantic search where the queries are formed using the labeled mentions from the Medmentions dataset. The top-k concepts are retrieved based on the cosine similarity of the query and entity encodings. These candidates are referred as generated candidates.

3.3 Disambiguation

The candidate generation solely relies on the cosine similarity score between the mention and prototype space candidate encodings. However, there may be cases where multiple candidates have similar scores or the scores alone may not be sufficient to rank the candidates. Therefore, there is a need to rerank the candidates. We propose the following reranking approaches that to perform the entity disambiguation:

3.3.1 Parametric Reranking

In this section, we propose a parametric approach to rerank the generated candidates. We consider three parameters based on the prototype space and our training framework for disambiguation namely, cosine similarity score (CSS), representative alias

[2] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

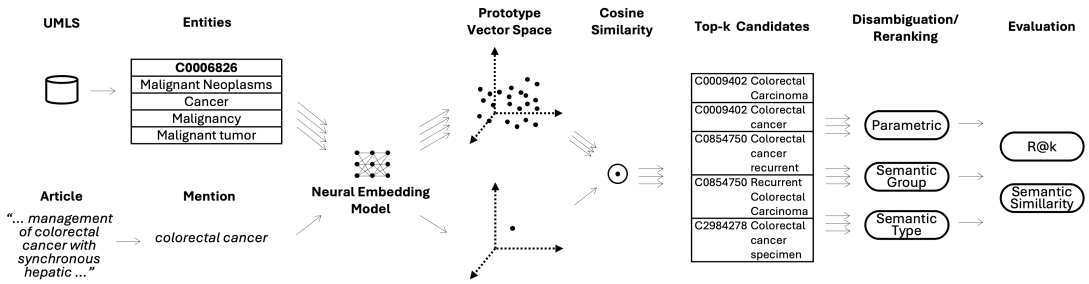

Figure 1: This figure illustrates the sequential flow of our proposed approaches. Starting from the left, we begin with leveraging a neural embedding model to create a prototype space on the UMLS entities. The cosine similarity metric is used to perform semantic search on the queries given the input mentions. The resultant top-k candidates are reranked using the listed methods for disambiguation and finally a comprehensive evaluation comprising of the retrieval performance and semantic similarity is performed.

score (RAS) and the candidate entity frequency score (CEFS) as our parameters. The parameters have the corresponding coefficients a, b and c respectively. These parameters are used to compute a new ranking score for each candidate. The equation below shows the updated score ($\delta^*(.,.)$) computation for reranking each of the generated candidates.

$$\delta^*(q,v) = a * \delta(q,v) + b * \frac{1}{n}\sum_{j=1}^{n}\delta(q,v_j) + c * n \quad (2)$$

Here, q is the query encoding, v is a generated candidate encoding and n is the number of aliases of v in the generated candidates.

The optimal selection of coefficients a, b and c corresponding to each of these parameters is performed through a grid search on a subset of manually defined bounds. Further details on the grid search and the impact of the these coefficients are discussed in appendix in the section A.2.

3.3.2 With UMLS Semantic Information

UMLS comprises of additional classification associated with individual entities, grouping them based on their semantic types and semantic groups. Each semantic type and semantic group has a canonical name. In this section, we calculate the cosine similarity between the mention's semantic type or semantic group canonical name encoding and the corresponding canonical names of the top-k candidates. This similarity score is added to the initial candidate generation score to rerank the top-k candidates.

1. **Assuming Availability of Gold Standard Information**: In this case, we assume that the gold standard semantic type and semantic group information is available for each mention. We rerank the candidates by utilizing the following methods:

   (a) **Semantic Type Based Disambiguation**: In this method, calculate the cosine similarity between canonical name encodings of semantic types of a mention and each of its top-k candidates. The updated score is computed as follows:

   $$\delta^*(q,v) = \delta(q,v) + \delta(TUI(q), TUI(v)) \quad (3)$$

   Here, $TUI(.)$ maps the input mention/entity to the encoding of corresponding semantic type canonical names.

   (b) **Semantic Group Based Disambiguation**: In this method, calculate the cosine similarity between canonical name encodings of semantic groups of a mention and each of its top-k candidates. The updated score is computed as follows:

   $$\delta^*(q,v) = \delta(q,v) + \delta(SG(q), SG(v)) \quad (4)$$

   Here, SG(.) maps the input mention/entity to the encoding of the corresponding semantic group canonical names.

2. **Semantic Type/Group Prediction**: In scenarios where the semantic type/group information of the mentions is not available, the methods proposed in (Le et al., 2022) and (Mao et al., 2023) can be used to predict the semantic type or group based on the input mentions. This can be followed by the computational methods discussed in the section 3.3.2.

4 Results and Discussion

We obtain the retrieval performance for the discussed approaches by considering the top-k closest candidates (that include aliases) from the prototype space. We observe that the retrieval performance (considering top-128 candidates) of all the miniEL and miniEL$_{1000}$ approaches is around 87% for the st21pv version and 88% for the full version of the Medmentions dataset.

4.1 Quantitative Analysis

In this section, we present the quantitive analysis associated with candidate generation (see section 4.1.1 and tables 2, 3) and reranking (see section 4.1.2, figure 2 and tables 4). Furthermore, the intricate analysis on the distribution of exact, related and missed candidate matches are discussed in the section 4.1.3.

4.1.1 How much data do we need?

In this section, we discuss the candidate generation performance of our approaches trained using varying number of examples. It can be seen that the performance of miniEL has a negligible training impact and the performance is stable across different number of examples (see tables 2 and 3). However, the miniEL* approach improves consistently with increasing number of training examples.The miniEL approach without any finetuning still outperforms the miniEL* approach trained on all the training examples.

	miniEL*		miniEL	
Training Samples	R@1	R@5	R@1	R@5
0	0.401	0.594	0.553	0.756
10	0.427	0.622	0.552	0.758
1000	0.499	0.693	0.557	0.766
10000	0.518	0.717	0.553	0.76
ALL	0.534	0.736	0.556	0.756

Table 2: This table shows the R@1 and R@5 candidate generation performance of the approaches on the Medmentions (st21pv) dataset. The models are trained with varying number of training samples used to train/finetune the MiniEL* and MiniEL models.

In comparison, our approach outperforms generative methods like BioBART (Yuan et al., 2022a) and BioGenEL (Yuan et al., 2022b) that are resource intensive. Since these approaches use the Medmentions training set to finetune the models, we only compare the test set performance. The R@1 candidate generation performance of MiniEL is 0.552 as compared to the overall performance of 0.496 and 0.520 of BioBART and BioGenEL respectively (the results are taken from (Kartchner et al., 2023)).

	MiniEL*		MiniEL	
Training Samples	R@1	R@5	R@1	R@5
0	0.462	0.657	0.567	0.782
10	0.477	0.676	0.565	0.783
1000	0.525	0.728	0.569	0.789
10000	0.537	0.747	0.568	0.788
ALL	0.556	0.761	0.568	0.783

Table 3: This table shows the R@1 and R@5 candidate generation performance of the approaches on the Medmentions (full) dataset. The models are trained with varying number of training samples used to train/finetune the MiniEL* and MiniEL models.

4.1.2 Reranking Performance

In the following subsections, we discuss the candidate reranking results. The results corresponding to the parametric approach and those corresponding to the semantic disambiguation approaches are discussed in the following subsections.

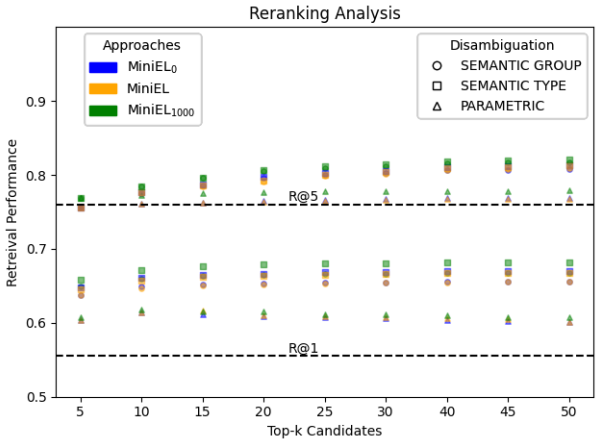

Figure 2: This figure highlights the trends associated with the retrieval performance improvement over varying top-k candidates using MiniEL$_0$, MinEL and MiniEL$_{1000}$ models. The improvement in R@1 is more significant as compared to that in R@5 for all the models and reranking methods. It can be observed that the retrieval performance of *PARAMETRIC* reranking decreases with increase in the top-k (k>15) whereas the performance of *SEMANTIC GROUP* and *SEMANTIC TYPE* reranking is consistent across the top-k.

1. **Parametric Reranking**: The top-k candidates selected based on the parametric approach discussed in section 3.3.1 and the corresponding results are shown in figure 2 and

	st21pv		full	
Reranking	top-5	top-10	top-5	top-10
PARAMETRIC	0.604	0.614	0.620	0.630
GROUP	0.638	0.649	0.659	0.670
TYPE	0.648	0.661	0.681	0.697

Table 4: The table shows the R@1 performance of the MiniEL$_0$ model after applying the listed reranking methods using the top-5 and top-10 candidates. It can be seen that there is a significant improvement in the performance as compared to the results in Tables 2 and 3.

table 4. It can be seen that the retrieval performance improves by from 0.553 to 0.614 for the st21pv version and from 0.567 to 0.630 for the full version of Medmentions. The a, b and c values used to obtain these results have the proportion $a : b : c \propto 50 : 2 : 1$ (see section A.2 for more details).

2. **With UMLS Semantic Information**: In this section, we discuss the retrieval performance improvements after the reranking using the semantic type and group information. The details of these methods are discussed in section 3.3.2.

   Figure 2 and table 4 show the R@1 performance of the MiniEL$_0$ model after applying these reranking strategies. The performance improves from 0.553 to 0.649 for semantic group and to 0.661 for semantic type reranking for the st21pv version of Medmentions. Similar observations can be made for the full version of Medmentions. Moreover, the retrieval performance does not deviate significantly with the increase in the top-k candidates used for reranking (see figure 2 for details).

The improvement in candidate ranking is approached in two ways. Firstly, to maximize the R@1 performance by reranking the generated candidates (see details in section 3.3) and secondly, to include context for addressing the context based ambiguity (see details in Appendix in section A.3).

4.1.3 How should the performance be evaluated?

In the retrieval-based evaluation strategy, we compute the retrieval performance on gold standard and predicted entity matches. However, there are cases where the most similar candidate is related to the gold standard entity. It can be seen in the table 5

Approach	Exact	Related	Missed
MiniEL$_0$	0.553	0.220	0.227
MiniEL$_0$ + PARAMETRIC	0.614	0.172	0.214
MiniEL$_0$ + GROUP	0.649	0.188	0.163
MiniEL$_0$ + TYPE	0.661	0.176	0.163

Table 5: This table shows the R@1 retrieval performance distributed into the exact matches, related matches and missed matches. The top-10 candidates are used for reranking. Here, we use the st21pv version of Medmentions.

that about 77% entities are exacting matching or are related to the gold standard entity. The details of each type of relation we have considered are provided by UMLS.[3]

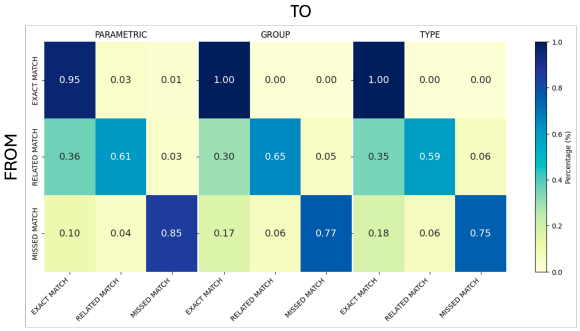

Figure 3: This heatmap illustrates the percentage changes in the number of initial exact, related and missed matches for the MiniEL$_0$ model. The performance preceding the changes is labeled 'FROM' for the rows, while the subsequent performance is denoted by 'TO' for the columns. The experiments are performed on the st21pv version of Medmentions.

Figure 3 shows that the effect of parametric reranking is directed primarily towards converting related matches to exact matches, coverting 36% of related matches into exact matches. The semantic group and semantic type based reranking approaches convert both missed and related matches into exact matches.

The following analysis is focused on the further evaluation of related and missed matches. In this article level analysis, we replace a mention with the closest generated candidate's canonical name for each mention in the article where the closest candidate is a related match or a missed match respectively. This results in an article A_P. We compute the cosine similarity between the original article A and the modified aricle A_P called S_P using a

[3] https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/abbreviations.html#mrdoc_REL

PubmedBERT-base (Gu et al., 2020) model[4] fine-tuned using sentence transformers (Reimers and Gurevych, 2019) on biomedical data. Similarly, we also replace the mentions with the gold standard canonical names to create an article A_G. This is followed by computation of cosine similarity between A and A_G called S_G. We focus on scenarios where S_P and S_G deviate significantly as compared to the mean deviation of the articles. These are highlighed in the figure 4. This forms a base for our qualitative analysis where we use this deviation to provide insights on the granularity of gold standard predictions as well as highlight current issues in the approach.

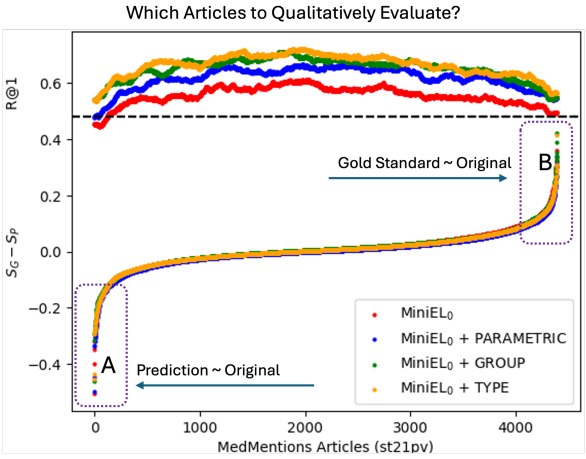

Figure 4: This figure illustrates the disparity in similarity scores ($S_G - S_P$) at the article level (4392 articles), alongside the smoothed retrieval performance (R@1) per article using a moving average with a window size of 200. The region A consists of semantically closer predictions and B consists of semantically farther predictions.

4.2 Qualitative Analysis

We perform a qualitative analysis on the entity linking predictions to highlight the difference in the granularity of the gold standard and predicted entities.

In this section, we qualitatively evaluate the articles displayed in the regions A and B of figure 4. The region A consists of articles where the predicted article A_P is semantically more similar to the original article A as compared to the gold standard articleA_G. Whereas, the region B consists of articles where A_G is more similar to A as compared to A_P.

[4] https://huggingface.co/NeuML/pubmedbert-base-embeddings

MENTION: *"Vitamin D Receptor Activator Use and Cause-specific...Vitamin D receptor activators (VDRA) may exert...5,635 VDRA users were matched...that VDRA use was"* **GOLD**: Biologically Active Substance (C0574031) **PREDICTION**: VDR protein, human (C3657722) with parent entity Vitamin D3 Receptor (C0108082)
MENTION: *"Influence of Sinus Floor Configuration....the sinus floor configuration...osteotome sinus grafting procedure...into the sinus area...sinus floor configuration...sinus floor profile...flat sinus group...maxillary sinus following...predictable in sinuses with a concave..."* **GOLD**: Anatomical space structure (C0229984) **PREDICTION**: Nasal sinus (C0030471)
MENTION: *"...effectiveness of disc synoptoscope on patients...effectiveness of disc synoptoscope on binocularity...therapy with disc synoptoscope in...with disc synoptoscope is effective...disc synoptoscope could serve as an..."* **GOLD**: Medical Devices (C0025080) **PREDICTION**: Synoptophores (C0183765)
MENTION: *"...performance of the Afirma gene expression classifier...the Afirma gene expression classifier (GEC)...on which GEC was performed...GEC testing was performed...atypia of undetermined significance (AUS)...the AUS cases...the AUS group...patients with AUS...value of GEC decreased from...suspicious GEC result...value of GEC in indeterminate...suspicious GEC result...suspicious GEC result..."* **GOLD**: Research Activities (C0243095), Finding (C0242481) **PREDICTION**: Gene Expression Profiling (C0752248), Atypical cells of undetermined significance (C0522580)
MENTION: *"including the cytoplasmic tails of integrins and components of the actin cytoskeleton"* **GOLD**: CytoPlasmic (C0521449) **PREDICTION**: Cytoplasmic Domain (C1511625) with alias 'Cytoplasmic Tail'.

Table 6: The table shows qualitative examples selected from the region A in the figure 4.

Table 6 shows the qualitiative examples from region A where it can be observed that our approach is penalized for granular or highly related predictions. For example, The mention *gene expression classifier* has a gold standard entity *Research Activities* as compared to the more granular prediction *Gene Expression Profiling*. Similarly, the mention *cytoplasmic tails* has a gold standard entity *CytoPlasmic* as compared to the more granular prediction *Cytoplasmic Domain*.

Table 7 shows the qualitatuve examples corresponding to the region B where it can be seen that the gold standard annotation is based on the context of mention in the article. More specifically, the mention *mice* has a gold standard entity: *Laboratory mice* based on the article context. However, this context is missing in the mention surface form. Therefore, to address these kind of cases, we need to provide the necessary context in the query. We utilize three different disambiguation techniques and show examples of the corresponding predictions. We observe that additional context from the articles may result in granular predictions.

However, the results are highly sensitive to the context and overall retrieval performance drops significantly (see section A.3 for more details).

We also observe an inconsistency in the granularity of gold standard entities in these examples. The mention *experimental mice* has a gold standard entity *Animals, Laboratory* as compared to the more granular prediction *Laboratory mice*.

MENTION: *"....iron accumulation in the substantia nigra (SN) of mice.....the substantia nigra of experimental mice treated with MPTP."* **GOLD**: Laboratory mice (C0025929), Animals, Laboratory (C0003064) **PREDICTION**: House mice (C0025914), Laboratory mice (C0025929)
MENTION: *"...mRNA N6-methyladenosine methylation of postnatal...mRNA m6A methylation during...outcomes of mRNA m6A methylation...levels of m6A methylation and...by m6A methylation at...higher m6A methylation and...differential m6A methylation may..."* **GOLD**: mRNA methylation (C2611689) **PREDICTION**: Methylation (C0025723)

Table 7: The table shows qualitative examples selected from the region B in the figure 4.

5 Conclusion

Biomedical entity linking has been an active area of research with various approaches being proposed to improve medical text standardization (see details in section 1). We propose a multi-stage approach where the first stage retrieves candidates with a high recall ($\sim 87\%$ for top-128 candidates). This is followed by application of the proposed reranking approaches focused on improving the R@1 retrieval performance. The reranking improves the performance by more than 10% (see figure 2 and table 4). We investigate the misses in R@1 and segregate the candidates into related and missed matches. Following this, we compute the article level semantic similarity together with the article level retrieval performance. This analysis highlights qualitative examples that can be used to obtain further insights about the framework. The semantic analysis is used to select the following types of qualitative examples: a) low retrieval performance and high similarity and, b) low retrieval performance and low similarity. The former can be highlight issues pertaining to granularity of gold standard entities and the latter can be used to highlight issues pertaining to the retrieval performance. Overall, the proposed techniques are highly effective in entity linking and have negligible training, prototype-space creation and inference costs (see table 9 for more details).

5.1 Future Scope

We believe that there is a significant scope for future developments in biomedical entity linking across different components of existing deep learning solutions. Firstly, there can be multiple biomedical normalizations for a mention or surface form. However, there is no method to determine the "closeness" of a prediction to a surface form as opposed to the binary matching. We believe that there should be a partial scoring instead of a binarized computation in order to accomodate the quality of predictions in the evaluation. Moreover, semantic similarities can also determined by experts to provide a ranking that could be used across biomedical entity linking for disambiguation.

5.2 Limitations

We observe that while an abbreviation preprocessing module is utilized in the proposed approaches, it doesn't convert all the abbreviations into their full forms. This causes a high amount of ambiguity in the results and often times the retrieval candidates do not consist of the correct entity. This drawback in positive pairs based learning has also been highlighted in (Zhang et al., 2021). Research addressed towards improving abbreviation expansion can help improve the recall of our candidate generation. Moreover, the region B in figure 4 highlights the examples where missing context in the surface form causes our framework to predict broader entities as the closest candidates. We utilize various approaches to include additional implicit and explicit context into our queries and analyze the corresponding retrieval performance (see details in Appendix section A.3).

References

Alan R Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In Proceedings of the AMIA Symposium, page 17. American Medical Informatics Association.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research, 32(suppl_1):D267–D270.

Leo Breiman. 2001. Random forests. Machine learning, 45:5–32.

Nello Cristianini and John Shawe-Taylor. 2000. An introduction to support vector machines and other

kernel-based learning methods. Cambridge university press.

Shaker El-Sappagh, Francesco Franda, Farman Ali, and Kyung-Sup Kwak. 2018. Snomed ct standard ontology based on the ontology for general medical science. BMC medical informatics and decision making, 18:1–19.

Samuele Garda and Ulf Leser. 2024. Belhd: Improving biomedical entity linking with homonoym disambiguation. arXiv preprint arXiv:2401.05125.

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.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. arXiv preprint arXiv:1705.00652.

David Kartchner, Jennifer Deng, Shubham Lohiya, Tejasri Kopparthi, Prasanth Bathala, Daniel Domingo-Fernández, and Cassie Mitchell. 2023. A comprehensive evaluation of biomedical entity linking models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 14462–14478, Singapore. Association for Computational Linguistics.

Luyang Kong, Christopher Winestock, and Parminder Bhatia. 2021. Zero-shot medical entity retrieval without annotation: Learning from rich knowledge graph semantics. arXiv preprint arXiv:2105.12682.

Linh Le, Guido Zuccon, Gianluca Demartini, Genghong Zhao, and Xia Zhang. 2022. Leveraging semantic type dependencies for clinical named entity recognition. In AMIA Annual Symposium Proceedings, volume 2022, page 662. American Medical Informatics Association.

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.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2020. Self-alignment pretraining for biomedical entity representations. arXiv preprint arXiv:2010.11784.

Yuqing Mao, Randolph A Miller, Olivier Bodenreider, Vinh Nguyen, and Kin Wah Fung. 2023. Two complementary ai approaches for predicting umls semantic group assignment: heuristic reasoning and deep learning. Journal of the American Medical Informatics Association, 30(12):1887–1894.

Sunil Mohan and Donghui Li. 2019. Medmentions: A large biomedical corpus annotated with umls concepts. arXiv preprint arXiv:1902.09476.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: fast and robust models for biomedical natural language processing. arXiv preprint arXiv:1902.07669.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.

Sunghwan Sohn, Donald C Comeau, Won Kim, and W John Wilbur. 2008. Abbreviation definition identification based on automatic precision estimates. BMC bioinformatics, 9:1–10.

Ali Soroush, Benjamin S Glicksberg, Eyal Zimlichman, Yiftach Barash, Robert Freeman, Alexander W Charney, Girish N Nadkarni, and Eyal Klang. 2024. Large language models are poor medical coders—benchmarking of medical code querying. NEJM AI, page AIdbp2300040.

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.

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.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022a. Biobart: Pretraining and evaluation of a biomedical generative language model. arXiv preprint arXiv:2204.03905.

Hongyi Yuan, Zheng Yuan, and Sheng Yu. 2022b. Generative biomedical entity linking via knowledge base-guided pre-training and synonyms-aware fine-tuning. arXiv preprint arXiv:2204.05164.

Sheng Zhang, Hao Cheng, Shikhar Vashishth, Cliff Wong, Jinfeng Xiao, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Knowledge-rich self-supervision for biomedical entity linking. arXiv preprint arXiv:2112.07887.

A Appendices

A.1 Terminology and Parameters

This section includes the terminology details and training, inference or other parameters used in this work.

Term	Description
δ	Similarity function
m	mention
ϵ	Entity
q	Query
μ	Entity canonical name
$TUI(.)$	maps an entity to it's semantic type canonical name
$SG(.)$	maps an entity to it's semantic group canonical name
$R@n$	Retrieval performance on top-n unique candidate entities
top-k	top-k candidate entities including aliases

Table 8: The table shows the symbols used in our work and the corresponding descriptions.

Table 9 shows the memory consumption and carbon emissions associated with the MiniEL$_0$ approach. It can be seen that our proposed techniques is low resource and results in very low amount of carbon emissions.

Phase	Memory (MB)	Emissions (Kg. Eq. CO2)
Training	0	0
Prototype Space Creation	1906	0.1
Inference	938	0.04

Table 9: The table shows the memory and carbon emission details. We utilized a 16GB V100 GPU for our tasks. The Inference was performed on the st21pv version of Medmentions.

A.2 Ablation Studies

In this section, we discuss the influence of parameters used in the parametric disambiguation approach discussed in the section 3.3. Specifically, we consider the candidate generation results obtained by using the MiniEL$_0$ model and perform the reranking by removing b and c parameters respectively. To highlight the impact of changing the a, b and c values, we perform a grid search on a manually selected range of values.

Furthermore, considering the top-10 candidates for reranking, removal of the parameter b results in an R@1 of 0.611, removal of c results in 0.481. This can be compared to the baseline R@1 0.553 and the R@1 of 0.614 obtained using optimal a,b and c. The performance is computed on the st21pv version of Medmentions. Overall, the impact of parameter c is highly significant in the performance improvement.

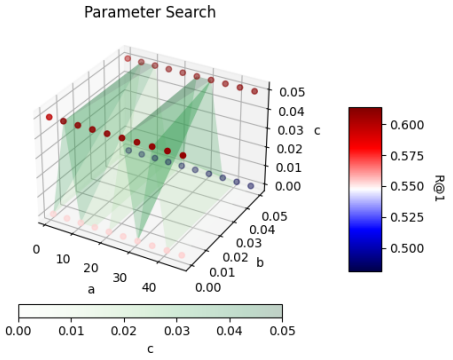

Figure 5: This figure shows the grid search on the parameters a, b and c for optimizing the R@1 performance of the MiniEL$_0$ model using the parametric approach discussed in the section 3.3. The optimal combination of a, b and c is found to be 5, 0.1 and 0.05, respectively.

A.3 Contextualized Queries

In our framework, the encoded representations of mentions are queried on the prototype space to get relevant candidates from UMLS. However, the mention spans alone may lack the necessary context to map the mention to their corresponding UMLS entities. In this section, we evaluate multiple techniques for incorporating context in the queries. Specifically, we use a running span based context addition, an implicit context addition and an attention based span context addition.

A.3.1 Neighboring Context

In this approach, we select a few words before and after the mention span to update the mention m and encode the updated mention to form a query.

Firstly, we add 5 neighbouring words before and after the mention and observe that the retrieval performance drops drastically ($R@1 \sim 10\%$). Therefore, we the number of words to 2 on both sides of the mention which results in a drastic drop in retrieval performance ($R@1 \sim 22\%$).

Overall, this context addition approach results in a significant drop in our retrieval performance and may not be suitable for contextual disambiguation.

A.3.2 Attention-Based Context

In this section, we perform experiments to identify the most influential words from the articles that

attend to the span in consideration. We modify the original mentions by adding these words as additional context. This is done by utilizing the attention mechanism of encoder based transformer models namely BioBERT (Lee et al., 2020). Firstly, the entire title and abstract text is tokenized and passed to these models. The corresponding attention outputs are obtained and passed to the mention enrichment algorithm.

Let k be the number of word-piece tokens obtained from the encoder model, for each head H of Layer L, the attention matrix can be mentioned as:

$$A = \begin{bmatrix} a_{11} & a_{12} & ... & a_{1k} \\ a_{21} & a_{22} & ... & a_{2k} \\ \vdots & \vdots & .... & \vdots \\ \vdots & \vdots & .... & \vdots \\ a_{k1} & a_{k2} & .... & a_{kk} \end{bmatrix} \quad (5)$$

The mention spans lie in the range $[c, d]$ where $0 \leq c < d \leq k$. Therefore, the matrix A can be shortened to a submatrix of interest B mentioned as:

$$B = \begin{bmatrix} a_{cc} & a_{c(c+1)} & ... & a_{cd} \\ a_{(c+1)c} & a_{(c+1)(c+1)} & ... & a_{(c+1)d} \\ \vdots & \vdots & .... & \vdots \\ \vdots & \vdots & .... & \vdots \\ a_{kc} & a_{k(c+1)} & .... & a_{kd} \end{bmatrix} \quad (6)$$

Equivalently,

$$B = \begin{bmatrix} b_c & b_{(c+1)} & ... & b_d \end{bmatrix} \quad (7)$$

where b_i represents a column of B. Next, the token corresponding to the maximum attention value of each column is obtained as $T(max(b_i))$ where $T(j)$ represents the token at index $j \in \{1, 2, ..., k\}$in the text spanning from 1^{st} to the k^{th} token. The resulting token vector from the attention head H_m and Layer L_n is represented as:

$$R_{nm} = \begin{bmatrix} T(max(b_c) & T(max(b_{(c+1)}) & ... & T(max(b_d) \end{bmatrix} \quad (8)$$

The *ENRICH* function discussed in the algorithm 1 return the enriched context for a given mention m, which is then modified as shown below:

$$m* = m : R_{mn}[0], R_{mn}[1] \quad (9)$$

Finally, stop words are removed from $R_{mn}[0]$ and $R_{mn}[1]$. An example mention cold can be modified as *cold:* ***severe,recent*** where, 'severe, recent' is the added context.

Algorithm 1 Enrichment Context Selection

procedure SORT$_{mcbl}$(V: 1D vector) {most common by length in descending order}
$C = \{x \mid count(x) = max(count(T)) \; \forall \; T \in V\}$
$C^* = \{x \mid x \in C \; and \; len(x) >= len(y) \; \forall \; y \in C\}$
return C^*
end procedure
{R_n denotes the representative token from all attention heads in Layer n}
{L_n denotes the representative token(s) from Layer n}
{M denotes the representative token(s) for the token$_k$ in mention M}
{E denotes the representative token context (E) for mention M}
procedure ENRICH(R_n: 1D vector) {enrich mention with context}
$C^* = SORT_{mcbl}(R_{nm})$
$R_n = C_1^*$ or $R_n = C^*(1)$
$L_n = \{R_1, R_2, ..., R_z\}$
$C^* = SORT_{mcbl}(L_n)$
$M_t = \{C_1^*, C_2^*\}$
$M = \{M_1, M_2, ..., M_k\}$
$C^* = SORT_{mcbl}(M)$
$E = \{C_1^*, C_2^*\}$
end procedure

A.3.3 Implicit Context

In this approach, we utilize mean-pooled embedding of the mention encodings taken from the entire article as an input. Firstly, the entire text is used as an input to obtain the tokenwise encodings from the model.

$$f(text, \theta) = \{E_{T_1}, E_{T_2}, ..., E_{T_n}\} \quad (10)$$

Here, E_T is encoding of token T and n are the number of tokens in the input text.

Given a span s, consisting of l tokens and tokens in the span $\{T_k, ..., T_{k+l}\}$, we take the corresponding encodings from the model outputs $\{E_{T_k}, ..., E_{T_{k+l}}\}$. We perform a mean pooling on these encodings to obtain the updated query representation $Q = \frac{1}{l}\sum_k^{k+l}\{E_{T_k}, ..., E_{T_{k+l}}\}$. The prototype space consists of the sentence encodings of the canonical names of all the entities in UMLS.

The R@1 candidate generation performance drops drastically in this setup where a drop of more than 30% is observed. Overall, we observe that these implicit contextual queries are not helpful in improvement of retrieval performance.

A.3.4 Evaluation

In this section, we perform the quantitative and qualitative analysis of our context based approaches on the Medmentions st21pv version. The qualitative examples shown below highlight the predictions provided by the proposed context based

MENTION: *"....iron accumulation in the substantia nigra (SN) of mice...."* **MENTION**^{AC}: *"...iron accumulation in the substantia nigra (SN) of mice: experiment...."* **PREDICTION**^{AC}: Laboratory mice (C0025929)) **PREDICTION**^{IC}: House mice (C0025914) **PREDICTION**^{NC}: Laboratory mice (C0025929)
MENTION: *Kindlin-1 is expressed primarily in epithelial cells, kindlin-2 is widely distributed and is particularly abundant in adherent cells, and kindlin-3 is expressed primarily in hematopoietic cells.* **MENTION**^{AC}: *Kindlin-1: kind,primarily is expressed primarily in epithelial cells, kindlin-2: distributed,Kind is widely distributed and is particularly abundant in adherent cells, and kindlin-3: expressed,Kind is expressed primarily in hematopoietic cells.* **PREDICTION**^{AC,IC,NC}: FERMT1 gene (C1423809), FERMT2 gene (C1423716), FERMT3 protein, human (C1311640) **PREDICTION**^{AC} **+ TYPE**: Fermitin Family Homolog 2, human (C3889282), Fermitin Family Homolog 2, human (C3889282), FERMT3 protein, human (C1311640)

Table 10: This table shows the qualitative analysis of the MiniEL$_0^{AC}$, MiniEL$_0^{IC}$ and MiniEL$_0^{NC}$ approaches on examples from Medmentions.

approaches. As discussed in the qualitative analysis of region B (see section 4.2), the surface forms have missing context resulting in an inaccurate prediction.

It can be observed in table 10 that the mention *mice* is correctly predicted as the entity *Laboratory mice* using the MiniEL$_0^{AC}$ and MiniEL$_{NC}$ reranking approaches. We also highlight the effect semantic type reranking approach though the example mentions *kindlin-2* and *kindlin-3* where the prediction semantic type changed from 'Gene' to the correct type 'Protein'. Here, the MiniEL$_0^{NC}$, MiniEL$_0^{AC}$ and MiniEL$_0^{IC}$ methods correspond to the results obtained using the Neighboring Context, Attention-based Context and Implicit Context approaches, respectively, utilizing MiniEL$_0$ as the base model.

It can be observed that the AC approach provides meaningful outputs as it includes the necessary context in the surface form. Similar outputs are provided by the NC approach. However, the neighbouring words may not necessarily contain the context and this can be seen in the following qualitative example listed in the table 11.

We observe that the attention span based context enrichment approach is sensitive to the context addition as it induces bias the surface form and the resulting candidates may be more similar to the bias term as compared to the base form. Therefore, to understand the impact of bias on the surface form, we observe the retrieval performance based on the number of words in the mention. The figure 6 shows that the performance of MiniEL$_0^{AC}$ is better

MENTION:"...inhibitor of T cell function....hypoxic conditions influence human T cell functions and found that..." **MENTION**^{AC}:"...inhibitor of T cell function: cell....hypoxic conditions influence human T cell functions: cell and found that..." **GOLD**: Cell physiology (C0007613), Cell physiology (C0007613) **PREDICTION**^{AC}: Cell physiology (C0007613), Cell physiology (C0007613) **PREDICTION**^{NC}: Cell physiology (C0007613), T cell differentiation (C1155013)

Table 11: This table shows the qualitative analysis of the MiniEL$_0^{AC}$ and MiniEL$_0^{NC}$ approaches on examples from Medmentions.

Approach	R@1	R@5
miniEL$_0$	0.553	0.756
miniEL$_0^{NC}$	0.219	0.405
miniEL$_0^{AC}$	0.384	0.642
miniEL$_0^{IC}$	0.161	0.359

Table 12: The table presents the candidate generation performance of the listed context based approaches. The performance is computed the st21pv version of Medmentions.

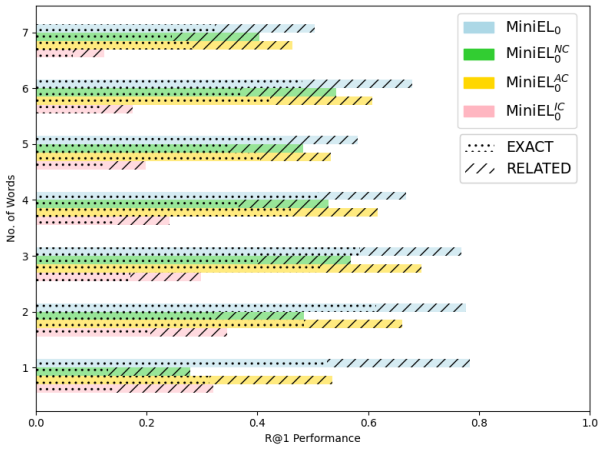

Figure 6: This figure presents the word-count level retrieval performance, measured in terms of exact and related matches, comparing the performance of the MiniEL$_0$ approach in comparison to its performance on applying the context based methods.

on mentions with higher length as compared to the mentions with lower lengths. A similar trend is observed for the MiniEL$_0^{NC}$ approach. This trend is not seen for the MiniEL$_0^{IC}$ approach where the performance drops with the increase in number of words in the mentions. However, the attention span based approach has better performance as compared to the neighboring context approach. For each specific mention word count, we select mentions with at least about a 100 examples for this

analysis.

To summarize, the quantitative and qualitative context enrichment analysis shows that the MiniEL$_0^{AC}$ approach outperforms the other approaches and is effective in context addition. However, the sensitivity in the encodings results in large deviations in the candidate generation (see table 12). Therefore, the robustness of this contextual approach needs to be improved.