Creating Ontology-annotated Corpora from Wikipedia for Medical Named-entity Recognition

Johann Frei and **Frank Kramer**
IT-Infrastructure for Translational Medical Research
University of Augsburg
<firstname>.<lastname>@informatik.uni-augsburg.de

Abstract

Acquiring annotated corpora for medical NLP is challenging due to legal and privacy constraints and costly annotation efforts, and using annotated public datasets may do not align well to the desired target application in terms of annotation style or language. We investigate the approach of utilizing Wikipedia and WikiData jointly to acquire an unsupervised annotated corpus for named-entity recognition (NER). By controlling the annotation ruleset through WikiData's ontology, we extract custom-defined annotations and dynamically impute weak annotations by an adaptive loss scaling. Our validation on German medication detection datasets yields competitive results. The entire pipeline only relies on open models and data resources, enabling reproducibility and open sharing of models and corpora. All relevant assets are shared on GitHub[1].

1 Introduction

A major reoccurring pain point in natural language processing (NLP) remains the issue of lacking resources regarding text corpora, including adequate annotation information in particular. Especially in medical and clinical NLP environments, the situation is notoriously difficult due to privacy and legal restrictions on data use and sharing; rendering efforts to share open datasets from the clinical domain complex and cumbersome. Yet there are notable and important attempts to provide text resources close to the clinical domain to the public research domain, e.g. MIMIC-III/-IV (Pollard and Johnson, 2016; Johnson et al., 2023). While these resources are extremely valuable, a single annotated dataset is governed by pre-defined parameters: First, the actual language, which may not align with the desired target language, and may pose another challenge regarding cross-lingual transfer. Second, the domain-dependent linguistic text properties and styles may vary from corpus to corpus. Third, the provided annotation data are usually tied to a certain ontology in entity linking, or label classes in named-entity recognition (NER). Forth, even if the label classes appear identical across multiple corpora, the underlying annotation guidelines that were employed as rulesets for annotation decision-making are usually not identical or consistent with annotation guidelines from other datasets. While the language and the text style are inherently fixed, new annotation data for a given corpus could be manually created if a custom annotation guideline is needed. However, a manual annotation is costly, resource-intensive and may remain non-reproducible to a certain degree due to the human-provided input. Hence, creating such an alternative annotation layer is not feasible in practice. In this work, we investigate the practicality of applying a fully unsupervised approach for annotated data acquisition in the medical context, yielding a corpus that is subsequently used as training material for an NER model. To achieve this, we compose several steps to obtain our final results. Our approach combines two public knowledge sources, Wikipedia[2] and WikiData (Vrandečić and Krötzsch, 2014)[3], to extract text data and annotation information, whereas the core annotation ruleset can be defined by leveraging the graph-like ontology structure of WikiData. The crafted dataset is used to train a conventional NER model. To demonstrate the feasibility of our approach, we evaluate our trained NER models on several external public datasets. Since our approach is in particular of interest for medium- to low-resource languages, we choose German as our non-English target language, but it is also motivated by the fact that external annotated datasets are available in that

[1] https://github.com/frankkramer-lab/WikiOntoNERCorpus

[2] https://www.wikipedia.org/ (accessed July 5th, 2024)

[3] https://www.wikidata.org/ (accessed July 5th, 2024)

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 570–579

language for a final evaluation.

Due to its simple availability, quality and text size, Wikipedia has been subject to NLP research in the past decade, in particular exploited further to obtain NER corpora. Therefore, it has been applied in numerous works for generic NER (Ghaddar and Langlais, 2017; Ryu et al., 2017; Nothman et al., 2008; Nothman et al., 2013; Hahm et al., 2014; Kim et al., 2012; Richman and Schone, 2008; Ni and Florian, 2016; Krishnan et al., 2021; Tsai et al., 2016; Alves et al., 2021), mostly using Wikipedia text or features, while others also include structured knowledge bases like DBpedia (Mendes et al., 2012) or FreeBase (Bollacker et al., 2008). Similar to our work, Jiang et al. 2021 also cover English biomedical NER for weak annotation data by modifying the loss function to be "noise-aware" and applying annotation imputation. Yet they also include a set of small, manually fully-annotated labels, and use PubMed texts with automatic dictionary-based label synthesis instead of Wikipedia resources. Regarding open domain NER, Liang et al. 2020 tackle the challenge by a similar two-stage self-training approach using WikiData, but do not further cover any custom Wikipedia parsing. To the best of our knowledge, no work using Wikipedia and WikiData has been reported so far in the German, medical domain.

2 Methods

2.1 Mapping Wikipedia and WikiData

Throughout this work, we consider Wikipedia as a language-dependent set of text documents identified by unique titles. The text documents are encoded in WikiText, a domain-specific language in markup style, which we mainly treat as plain text sequences along with span-oriented text references to other Wikipedia documents. In contrast, WikiData is a language-agnostic knowledge base which encodes its knowledge in a graph structure with typed, directed edges ("statements") between individual nodes. Each node is a WikiData entity, uniquely identified by its QID number, and either represents an actual concept (e.g. *cancer* (Q12078)) and therefore may be referenced to its corresponding language-specific Wikipedia page, or is part of a virtual concept that encodes certain ontology-inspired hierarchy structures (e.g. *class of disease* (Q112193867)). Note that the correspondences between the WikiData entities to their language-specific Wikipedia pages are bijective in most cases. We utilize these references to establish a mapping between WikiData and Wikipedia entries.

2.2 Extracting Annotations from Wikipedia

The WikiText markup language, which is used to encode the Wikipedia pages, facilitates the use of references to other Wikipedia pages from the same language. These kinds of references are eventually rendered as hyperlinks in web browsers, and are used to link certain terms within a Wikipedia page text to pages that address the mentioned concepts as their main topic. Given a set of concepts we are interested in, defined as a set of Wikipedia pages in practice, we parse the language-specific Wikipedia dump to extract all sentences that contain references to our set of concepts of interest while we retain the reference information of each of the extracted sentences with regard to the text span of the link and its target page. Note that for each extracted sentence, we also keep the information on references that were *not* part of our set of concepts as *negative* mentions. Finally, we obtain an annotated corpus in a certain language that contains annotation information for mentions of concepts of our interest. However, since not every mentioned concept is usually referenced in the WikiText and concepts worth referencing are usually only reference at the first occurrence on a page, the obtained corpus only contains weakly-annotated labels. Using the corpus as training resource to directly train an NER model therefore is expected to yield a model with high precision, yet very low recall scores due to the weak annotation.

2.3 Graph-defined Entity Selection using SPARQL

To enable the use of the WikiData ontology to define a set of concepts of interest, we leverage the SPARQL interface[4], an RDF query language, to determine all WikiData entities of interest. By these means, we can make use of more complex queries that take full advantage of the WikiData ontology structures, and thus it yields an explainable and well-defined output. We further resolve the WikiData entities into their language-specific Wikipedia pages by the mapping we established before, and extract all relevant sentences with annotations, as described earlier. The entire process is illustrated in Figure 1.

[4] https://query.wikidata.org/ accessed May 3rd, 2024

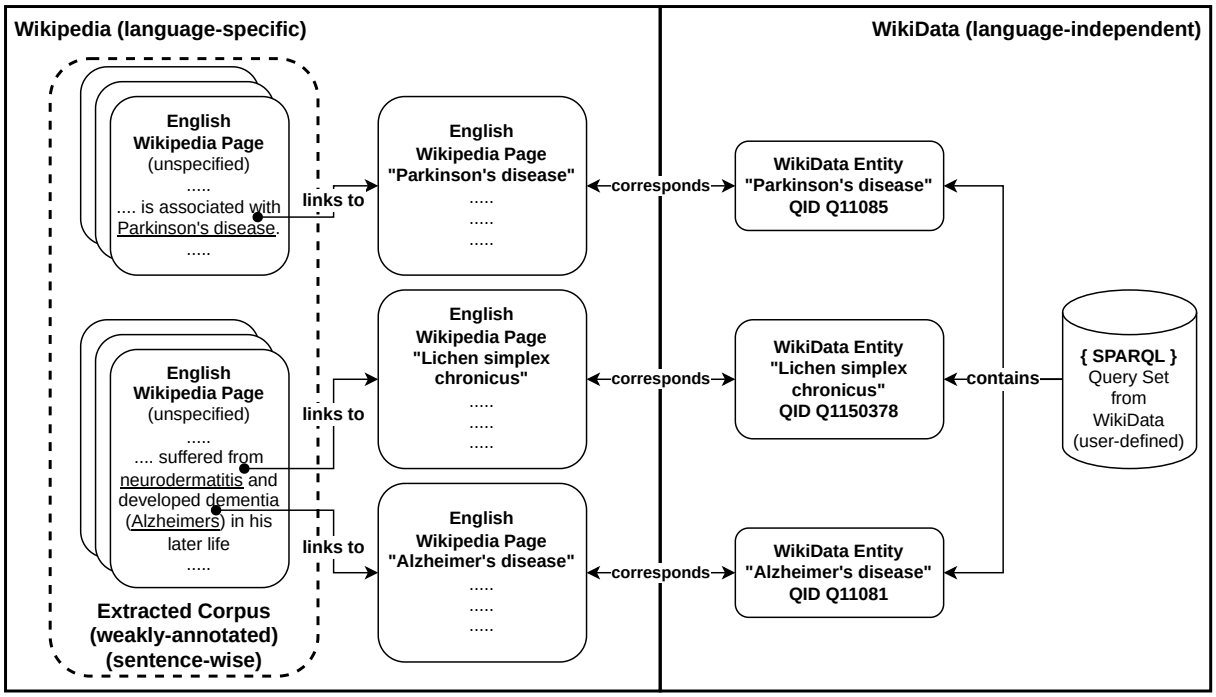

Figure 1: Conceptual design of the weakly-annotated dataset creation for a certain SPARQL query on WikiData.

2.4 Dataset Imputation of Weak Annotations

Since NER can be framed as a token classification task, we can stratify our evidence about a token class into three cases: First, if the token is an actual part of a reference to a page from our set of concepts, it is considered a *positive* token. Second, if the token is also part of a reference, but does not link to our desired concept set, we consider it a *negative* token. Any other token without any reference is declared as *unknown*, for which we assume the token to not belong to any named entity class similar to *negative* tokens, yet due to the weak annotation its actual class remains unclear. To mitigate the weak label signal during training, we dynamically scale the gradient loss for each token by the factor ω according to the following schema:

$$L_{scaled} = \begin{cases} \omega_{pos} L & \text{if token} \in \text{positive} \\ \omega_{neg} L & \text{if token} \in \text{negative} \\ \omega_{unk} L & \text{if token} \in \text{unknown} \end{cases} \quad (1)$$

We balance the loss scaling weights ω accordingly.

$$\omega_{pos} = \frac{\#tokens_{neg}}{\#tokens_{pos}}, \omega_{neg} = \frac{\#tokens_{pos}}{\#tokens_{neg}} \quad (2)$$

ω_{unk} remains choosable as a hyperparameter. Given the actual positive tokens, the negative tokens serving as contrastive samples, as well as the unknown token samples, we can train an NER model while maintaining the dynamic loss scaling at each token position. Given the trained NER model, it can be re-applied to the weakly-annotated corpus in order to impute missing annotation spans to subsequently obtain a silver standard, fully-annotated corpus as shown in Figure 2.

3 Results

Regarding our implementation, significant portions are re-purposed from existing work (Frei et al., 2022). To assess our proposed approach in medical, non-English NER, we mimic a medication detection task in German texts due to the availability of public datasets with annotations including label classes semantically related to drug or medication, namely BRONCO150 (Kittner et al., 2021), CARDIO:DE (Richter-Pechanski et al., 2023), GPTNERMED (Frei and Kramer, 2023), GERNERMED++ (Frei et al., 2023), and GGPOnc 2 (Borchert et al., 2022) (with short, fine annotation layer). To address the medication detection task, our simple entity selection strategy hereby filters all WikiData entries that have an ATC code assigned through the WikiData property *P267* to eventually obtain a weakly-annotated corpus. Based on this corpus, we fine-tune an NER model with the Huggingface Transformers (Wolf et al., 2020) library for different ω_{unk} scalars while ω_{pos}/ω_{neg}

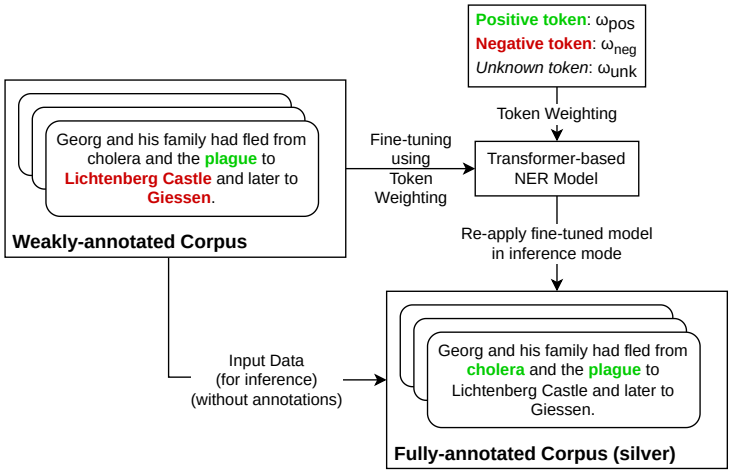

Figure 2: Illustration of the dataset imputation process for the annotation data using dynamic token loss scaling.

remains balanced (1.337/0.748). In order to retain a generic, non-medical setup for the dataset imputation, we use GottBERT (Scheible et al., 2020), a non-domain-specific German RoBERTa model as encoder model for our NER token classifier. To obtain a fully-annotated corpus, we apply the dataset imputation on the weakly-annotated corpus using the fine-tuned NER model. The statistics of the corpus in various configurations are provided in Table 1. Finally, we simulate a traditional but domain-aware NER fine-tuning setup that assumes an ordinary, fully-annotated corpus by training on the imputed corpus and use the default NER training setup of SpaCy (Montani et al., 2023) with GerMedBERT's *medBERT.de* (Bressem et al., 2024), a German medical language model, as BERT encoder without applying any token loss scaling. The scores on the external datasets are evaluated using *BratEval*[5] in overlap mode, and shown in Table 2. The results for all ω_{unk} values are provided in Table 3 in the appendix.

The results on the BRONCO150, GERNERMED++, and GPTNERMED datasets indicate best and robust F1 scores in cases of $\omega_{unk} < 0.8$, whereas for CARDIO:DE the scores are evidently rather modest, to rather poor in case of GGPOnc2. The impact of the token loss scaling ω_{unk} is, as expected, clearly noticeable as it consistently increases the recall rates at lower ω_{unk} values at the expense of minor precision losses. A preliminary lookup into the annotation disagreements in CARDIO:DE and GGPOnc reveals that the models tend to perform poorly on corpora with sparse, less frequent annotations, especially on samples without any annotation. However, another apparent factor remains the conceptual disagreements, for instance "Thrombozyten" was detected due to its assigned ATC code in its WikiData entity *Q101026*, however it was not annotated by the Gold standard annotation.

Property	Value
#Samples	84478
#Sents	90918
#Tokens	2023187
#Labels (pos), raw	105207
#Labels (neg), raw	145492
#Labels (pos), $\omega_{unk} = 0.01$	135795
#Labels (pos), $\omega_{unk} = 0.05$	127950
#Labels (pos), $\omega_{unk} = 0.1$	128157
#Labels (pos), $\omega_{unk} = 0.2$	120973
#Labels (pos), $\omega_{unk} = 0.5$	114339
#Labels (pos), $\omega_{unk} = 0.8$	110237
#Labels (pos), $\omega_{unk} = 1.0$	110024

Table 1: Statistics of the obtained datasets for different ω_{unk} settings. The corpora are based on the SPARQL query that identifies all WikiData entities with assigned ATC codes. The query is given in the appendix.

4 Discussion and Limitations

While a conclusive verdict is not feasible based on the pure evaluation scores due to the diverse and incoherence issues of the external datasets, the fact that the entire data process operates solely on open data yet can perform surprisingly well, even on external corpora, encourages further efforts to foster open NLP resources and models, especially for more sparse, non-English domains. Comparing our results with related work, the GGPOnc 2 baseline

[5] https://github.com/READ-BioMed/brateval/tree/v0.3.2 (Commit c4f5fff) accessed May 3rd, 2024

ω_{unk}	Dataset	Pr	Re	F1
0.01	BRONCO150	0.8103	0.7505	0.7792
0.2	(Kittner et al., 2021)	0.8014	0.7538	0.7768
1.0	[MEDICATION]	0.8537	0.5983	0.7035
0.01	GERNERMED++	0.8104	0.7897	0.7999
0.2	(Frei et al., 2023)	0.8453	0.7526	0.7963
1.0	[Drug]	0.8831	0.6841	0.771
0.01	GPTNERMED	0.8002	0.8802	0.8383
0.2	(Frei and Kramer, 2023)	0.8553	0.8537	0.8545
1.0	[Medikation]	0.8336	0.8172	0.8253
0.01	CARDIO:DE	0.5402	0.7266	0.6197
0.2	(Richter-Pechanski et al., 2023)	0.5352	0.7107	0.6106
1.0	[DRUG, ACTIVEING]	0.5634	0.5924	0.5775
0.01	GGPOnc 2	0.1908	0.7257	0.3021
0.2	(Borchert et al., 2022)	0.2324	0.6635	0.3442
1.0	[Clinical_Drug] (short, fine)	0.2425	0.5702	0.3402

Table 2: Performance scores on external datasets using BratEval in *overlap* mode for **Pr**ecision, **Re**call and **F1** score for different ω_{unk} values. See Table 3 in the appendix for all ω_{unk} values. The harmonized label classes are given in square brackets.

NER model is reported to achieve .91 F1-score on its test set on the *Clinical_Drug* label class, likewise CARDIO:DE achieves .85/0.81 F1-scores for the *ACTIVEING/DRUG* label classes. However, major disagreements are reported in cross-corpus model transfer. For instance, (Richter-Pechanski et al., 2023) report a .21 F1-score for the DRUG class when applying the GGPOnc 2 NER model to the held-back part of the CARDIO:DE corpus, highlighting certain innate limitations when comparing F1-scores across different datasets and annotation guidelines, as well as the need for the use of multiple datasets during evaluation. As for another, less severe instance, (Frei et al., 2023) and (Frei and Kramer, 2023) report .73/.72. F1-scores respectively on the BRONCO150 corpus for the *MEDICATION* label class, hinting towards more consistent mutual annotation agreements.

Other factors in NER are not further addressed in this work, such as efforts towards support for nested entities or discontinuous annotations or the support for label classes beyond medication detection. The latter aspect may be achieved by the use of an updated SPARQL query definition for certain label classes that align well to the annotation schema from Wikipedia articles but may fail for other label classes like *strength-* or *frequency-*related entities which may not be covered well by Wikipedia or WikiData. In this regard, potential limitations within the WikiData knowledge base are not investigated that may influence the quality of our results in other domains. Other limiting factors such as the quality of the pre-trained language models are not quantified in isolation. However, in general, the effective use of more sophisticated SPARQL queries for entity selection may unlock further potential gains, as well as its application in other languages and domains since our method is not inherently bound to the medical field. Yet, these aspects only serve as motivation for future work.

5 Conclusion

In this work, we demonstrated an unsupervised approach for creating an annotated dataset for medical NER for the German language defined by the WikiData ontology structure using exclusively open data resources. The proof of concept of the proposed method in practical scenarios was shown on a set of external datasets, yielding surprisingly well results. We also discussed further potential but currently underexplored factors such as improved SPARQL queries as future work. Relevant assets are published on GitHub[6], including a web interface that enables external users to create new corpora from custom SPARQL queries for independent experiments.

Acknowledgments

We want to thank Dr. Lisa Raithel for proofreading the manuscript and for providing valuable feedback and support. We also thank all three anonymous reviewers for their constructive and positive feedback.

References

Diego Alves, Gaurish Thakkar, and Marko Tadic. 2021. Building and evaluating universal named-entity recognition english corpus. In *Proceedings of the 2nd International Workshop on Cross-lingual Event-centric Open Analytics co-located with the 30th The Web Conference (WWW 2021), Ljubljana, Slovenia, April 12, 2021 (online event due to COVID-19 outbreak)*, volume 2829 of *CEUR Workshop Proceedings*, pages 2–16. CEUR-WS.org.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, SIGMOD '08, pages 1247–1250. Association for Computing Machinery.

Florian Borchert, Christina Lohr, Luise Modersohn, Jonas Witt, Thomas Langer, Markus Follmann,

[6] https://github.com/frankkramer-lab/WikiOntoNERCorpus (accessed July 5th, 2024)

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 Language Resources and Evaluation Conference*, pages 3650–3660. 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, Moritz Augustin, Lennart Grosser, Marcus R. Makowski, Hugo JWL Aerts, and Alexander Löser. 2024. MEDBERT.de: A comprehensive german BERT model for the medical domain. *Expert Systems with Applications*, 237:121598.

Johann Frei, Ludwig Frei-Stuber, and Frank Kramer. 2023. GERNERMED++: Semantic annotation in german medical NLP through transfer-learning, translation and word alignment. *Journal of Biomedical Informatics*, 147:104513.

Johann Frei and Frank Kramer. 2023. Annotated dataset creation through large language models for non-english medical NLP. *Journal of Biomedical Informatics*, 145:104478.

Johann Frei, Iñaki Soto-Rey, and Frank Kramer. 2022. Drnote: An open medical annotation service. *PLOS Digital Health*, 1(8):1–18.

Abbas Ghaddar and Phillippe Langlais. 2017. WiNER: A wikipedia annotated corpus for named entity recognition. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 413–422. Asian Federation of Natural Language Processing.

Younggyun Hahm, Jungyeul Park, Kyungtae Lim, Youngsik Kim, Dosam Hwang, and Key-Sun Choi. 2014. Named entity corpus construction using wikipedia and DBpedia ontology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2565–2569. European Language Resources Association (ELRA).

Haoming Jiang, Danqing Zhang, Tianyu Cao, Bing Yin, and Tuo Zhao. 2021. Named entity recognition with small strongly labeled and large weakly labeled data. 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 1775–1789. Association for Computational Linguistics.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Liwei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1. Publisher: Nature Publishing Group.

Sungchul Kim, Kristina Toutanova, and Hwanjo Yu. 2012. Multilingual named entity recognition using parallel data and metadata from wikipedia. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 694–702. Association for Computational Linguistics.

Madeleine Kittner, Mario Lamping, Damian T Rieke, Julian Götze, Bariya Bajwa, Ivan Jelas, Gina Rüter, Hanjo Hautow, Mario Sänger, Maryam Habibi, Marit Zettwitz, Till de Bortoli, Leonie Ostermann, Jurica Ševa, Johannes Starlinger, Oliver Kohlbacher, Nisar P Malek, Ulrich Keilholz, and Ulf Leser. 2021. Annotation and initial evaluation of a large annotated german oncological corpus. *JAMIA Open*, 4(2):ooab025.

Aravind Krishnan, Stefan Ziehe, Franziska Pannach, and Caroline Sporleder. 2021. Employing wikipedia as a resource for named entity recognition in morphologically complex under-resourced languages. In *Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021)*, pages 28–39. INCOMA Ltd.

Chen Liang, Yue Yu, Haoming Jiang, Siawpeng Er, Ruijia Wang, Tuo Zhao, and Chao Zhang. 2020. BOND: BERT-assisted open-domain named entity recognition with distant supervision. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, pages 1054–1064. Association for Computing Machinery.

Pablo Mendes, Max Jakob, and Christian Bizer. 2012. DBpedia: A multilingual cross-domain knowledge base. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1813–1817. European Language Resources Association (ELRA).

Ines Montani, Matthew Honnibal, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, and Henning Peters. 2023. explosion/spaCy: v3.7.2: Fixes for APIs and requirements.

Jian Ni and Radu Florian. 2016. Improving multilingual named entity recognition with wikipedia entity type mapping. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1275–1284. Association for Computational Linguistics.

Joel Nothman, James R. Curran, and Tara Murphy. 2008. Transforming wikipedia into named entity training data. In *Proceedings of the Australasian Language Technology Association Workshop 2008*, pages 124–132.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175.

Tom J Pollard and Alistair EW Johnson. 2016. The MIMIC-III clinical database.

Alexander E. Richman and Patrick Schone. 2008. Mining wiki resources for multilingual named entity recognition. In *Proceedings of ACL-08: HLT*, pages 1–9. Association for Computational Linguistics.

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. *Scientific Data*, 10(1):207. Publisher: Nature Publishing Group.

Seonghan Ryu, Hwanjo Yu, and Gary Geunbae Lee. 2017. Two-stage approach to named entity recognition using wikipedia and DBpedia. In *Proceedings of the 11th International Conference on Ubiquitous Information Management and Communication*, IMCOM '17, pages 1–4. Association for Computing Machinery.

Raphael Scheible, Fabian Thomczyk, P. Tippmann, V. Jaravine, and M. Boeker. 2020. GottBERT: a pure german language model. *ArXiv*.

Chen-Tse Tsai, Stephen Mayhew, and Dan Roth. 2016. Cross-lingual named entity recognition via wikification. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 219–228. Association for Computational Linguistics.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

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. Association for Computational Linguistics.

A Complete NER Results

The results for all ω_{unk} values are provided in Table 3.

ω_{unk}	Dataset	Pr	Re	F1
0.01	BRONCO150 (Kittner et al., 2021) [MEDICATION]	0.8103	0.7505	0.7792
0.05		0.8209	0.7302	0.7729
0.1		0.7762	0.7727	0.7745
0.2		0.8014	0.7538	0.7768
0.5		0.8381	0.6728	0.7464
0.8		0.8618	0.5865	0.698
1.0		0.8537	0.5983	0.7035
0.01	GERNERMED++ (Frei et al., 2023) [Drug]	0.8104	0.7897	0.7999
0.05		0.8363	0.7383	0.7843
0.1		0.7627	0.7654	0.764
0.2		0.8453	0.7526	0.7963
0.5		0.8518	0.7152	0.7776
0.8		0.8773	0.6876	0.771
1.0		0.8831	0.6841	0.771
0.01	GPTNERMED (Frei and Kramer, 2023) [Medikation]	0.8002	0.8802	0.8383
0.05		0.8286	0.8638	0.8458
0.1		0.741	0.8787	0.804
0.2		0.8553	0.8537	0.8545
0.5		0.83	0.8413	0.8356
0.8		0.8145	0.8151	0.8148
1.0		0.8336	0.8172	0.8253
0.01	CARDIO:DE (Richter-Pechanski et al., 2023) [DRUG, ACTIVEING]	0.5402	0.7266	0.6197
0.05		0.5219	0.6774	0.5895
0.1		0.5786	0.7278	0.6447
0.2		0.5352	0.7107	0.6106
0.5		0.5856	0.6506	0.6163
0.8		0.6262	0.5731	0.5985
1.0		0.5634	0.5924	0.5775
0.01	GGPOnc 2 (Borchert et al., 2022) [Clinical_Drug] (short, fine)	0.1908	0.7257	0.3021
0.05		0.2386	0.6944	0.3552
0.1		0.1855	0.7026	0.2935
0.2		0.2324	0.6635	0.3442
0.5		0.2085	0.6265	0.3128
0.8		0.2143	0.5805	0.3131
1.0		0.2425	0.5702	0.3402

Table 3: Performance scores on external datasets using BratEval in *overlap* mode for **Pr**ecision, **Re**call and **F1** score for all different ω_{unk} values. The harmonized label classes are given in square brackets.

B Setup Parameters

B.1 SPARQL Query for Entity Selection

The SPARQL query that has been used for the entity selection. The query selects all WikiData entities with an assigned ATC code.

```
# Anything that has an assigned ATC code
SELECT ?item
WHERE
{
    ?item wdt:P267 ?atccode .
}
```

The query was performed on the WikiData SPARQL query service[7] on April 17th, 2024.

B.2 Training Configuration

B.2.1 First-stage NER with Dynamic Loss Scaling

We use the Transformers (Wolf et al., 2020) library to train the first NER model used for dataset imputation during successive steps. The entire dataset (for ATC) was split into train+dev set (90%) and test set (10%), and the train+dev set was split into train set (80%) and dev set (20%). The following Huggingface Transformers parameters were used for training.

```
"evaluation_strategy": "epoch",
"per_device_train_batch_size": 32,
"per_device_eval_batch_size": 32,
"gradient_accumulation_steps": 1,
"learning_rate": 5e-05,
"weight_decay": 0.0,
"adam_beta1": 0.9,
"adam_beta2": 0.999,
"adam_epsilon": 1e-08,
"max_grad_norm": 1.0,
"num_train_epochs": 3,
"lr_scheduler_type": "linear",
"warmup_ratio": 0.0,
"warmup_steps": 0,
"save_strategy": "epoch",
"seed": 42,
"load_best_model_at_end": true,
"metric_for_best_model": "loss",
"optim": "adamw_torch",
```

As a pre-trained encoder, we used `uklfr/gottbert-base` (Scheible et al., 2020) from the Huggingface Hub. The final model was picked according to the lowest loss on the dev set.

B.2.2 Output Token Decoding for Dataset Imputation

For the decoding of the output probabilities from the first-stage NER model for the dataset imputation, we used a greedy decoding strategy to predict the IOB2 labels (*O*, *B-LABEL*, *I-LABEL*). However, invalid outputs were set to -inf prior to the final token decoding.

B.2.3 Second-stage NER on Imputed Dataset

For the training on the imputed dataset using SpaCy (Montani et al., 2023), the initial default configuration was created with the CLI command:

[7] https://query.wikidata.org/

`python3 -m spacy init config base.cfg -l de -p ner -G -o accuracy`. The following modifications were made to the base configuration:

- In `[components.transformer.model]`, set `name = "GerMedBERT/medbert-512"`
- In `[training.optimizer.learn_rate]`, set `initial_rate = 5e-5`
- In `[training]`, set `max_epochs = 10`
- In `[training]`, set `max_steps = -1`
- In `[training]`, set `seed = 0`

The final configuration was created with the CLI command: `python3 -m spacy init fill-config base.cfg final.cfg`. Similar to the first-stage NER training, the entire dataset was split into train+dev set (90%) and test set (10%), and the train+dev set was split into train set (80%) and dev set (20%). After training, the best model was picked according to the best (internal) F1 score on the dev set, as this is the default SpaCy approach.

C Data Versioning

C.1 NLP Tools

The Python libraries from pypi.org in the following versions were used for the experiments:

- **Huggingface Transformers**: `transformers`: 4.36.2
- **SpaCy**: `spacy`: 3.7.4
- **SpaCy-Transformers**: `spacy-transformers`: 1.3.4

C.2 Wikipedia and WikiData Dumps

The dumps for WikiData and Wikipedia were accessed by the web references at the following time:

- **Wikipedia / German**: https://dumps.wikimedia.org/dewiki/latest/dewiki-latest-pages-meta-current.xml.bz2 on February 22, 2024.
- **Wikipedia / English**: https://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-meta-current.xml.bz2 on February 26, 2024.
- **Wikipedia / French**: https://dumps.wikimedia.org/frwiki/latest/frwiki-latest-pages-meta-current.xml.bz2 on February 26, 2024.
- **Wikipedia / Spanish**: https://dumps.wikimedia.org/eswiki/latest/eswiki-latest-pages-meta-current.xml.bz2 on February 26, 2024.
- **WikiData**: https://dumps.wikimedia.org/wikidatawiki/entities/ on February 23, 2024.

D Annotated Text Samples

To visualize the effect of the annotation imputation stage, several samples from the datasets are shown in Table 4. The datasets are based on the SPARQL query which applies the ATC code assignment filter. While in most instances, the added entities can be considered correct, some ambiguities persist even after manual inspection. For instance, the words "calcium-" and "magnesiumhaltigen" may refer to "calcium carbonate" (Q23767) and "magnesium carbonate" (Q407931) and ATC codes are assigned to their corresponding WikiData entities. However, the correspondent WikiData items for "calcium" (Q706) and "magnesium" (Q660) lack any ATC code.

Setup	Text Sample
raw (neg)	Dabei wird das Kollagen des Fleischbindegewebes durch **Säuren**, Tannine und weitere Bestandteile des Weins angegriffen, gelockert und teilweise gelatiniert, wodurch das Fleisch zarter wird und Geschmack freigesetzt wird.
raw (pos)	Dabei wird das **Kollagen** des Fleischbindegewebes durch Säuren, Tannine und weitere Bestandteile des Weins angegriffen, gelockert und teilweise gelatiniert, wodurch das Fleisch zarter wird und Geschmack freigesetzt wird.
ω_{unk} = 0.01 (imp)	Dabei wird das **Kollagen** des Fleischbindegewebes durch Säuren, Tannine und weitere Bestandteile des Weins angegriffen, gelockert und teilweise **gelatiniert**, wodurch das Fleisch zarter wird und Geschmack freigesetzt wird.
ω_{unk} = 0.2 (imp)	Dabei wird das **Kollagen** des Fleischbindegewebes durch Säuren, Tannine und weitere Bestandteile des Weins angegriffen, gelockert und teilweise **gel**atiniert, wodurch das Fleisch zarter wird und Geschmack freigesetzt wird.
ω_{unk} = 1.0 (imp)	Dabei wird das **Kollagen** des Fleischbindegewebes durch Säuren, Tannine und weitere Bestandteile des Weins angegriffen, gelockert und teilweise gelatiniert, wodurch das Fleisch zarter wird und Geschmack freigesetzt wird.
raw (neg)	Die klinische Entwicklung bei Depression wurde jedoch eingestellt, da Rolipram im Vergleich zu herkömmlichen Antidepressiva keinen Zusatznutzen zeigen konnte.
raw (pos)	Die klinische Entwicklung bei Depression wurde jedoch eingestellt, da Rolipram im Vergleich zu herkömmlichen **Antidepressiva** keinen Zusatznutzen zeigen konnte.
ω_{unk} = 0.01 (imp)	Die klinische Entwicklung bei Depression wurde jedoch eingestellt, da **Rolipram** im Vergleich zu herkömmlichen **Antidepressiva** keinen Zusatznutzen zeigen konnte.
ω_{unk} = 0.2 (imp)	Die klinische Entwicklung bei Depression wurde jedoch eingestellt, da Rolipram im Vergleich zu herkömmlichen **Antidepressiva** keinen Zusatznutzen zeigen konnte.
ω_{unk} = 1.0 (imp)	Die klinische Entwicklung bei Depression wurde jedoch eingestellt, da Rolipram im Vergleich zu herkömmlichen **Antidepressiva** keinen Zusatznutzen zeigen konnte.
raw (neg)	Durch die Gabe von calcium- und magnesiumhaltigen Antacida nach oraler Überdosierung von Ofloxacin kann die Resorption infolge Bildung schwerlöslicher Komplexe verzögert werden.
raw (pos)	Durch die Gabe von calcium- und magnesiumhaltigen **Antacida** nach oraler Überdosierung von Ofloxacin kann die Resorption infolge Bildung schwerlöslicher Komplexe verzögert werden.
ω_{unk} = 0.01 (imp)	Durch die Gabe von **calcium-** und magnesiumhaltigen **Antacida** nach oraler Überdosierung von **Ofloxacin** kann die Resorption infolge Bildung schwerlöslicher Komplexe verzögert werden.
ω_{unk} = 0.2 (imp)	Durch die Gabe von **cal**cium- und magnesiumhaltigen **Antacida** nach oraler Überdosierung von **Oflo**xacin kann die Resorption infolge Bildung schwerlöslicher Komplexe verzögert werden.
ω_{unk} = 1.0 (imp)	Durch die Gabe von calcium- und magnesiumhaltigen **Antacida** nach oraler Überdosierung von Ofloxacin kann die Resorption infolge Bildung schwerlöslicher Komplexe verzögert werden.
raw (neg)	Bei einer **Überdosierung** von Fenetyllin werden große Mengen der **Neurotransmitter** Noradrenalin und Dopamin aus den Speichervesikeln im **zentralen Nervensystem** freigesetzt.
raw (pos)	Bei einer Überdosierung von Fenetyllin werden große Mengen der Neurotransmitter **Noradrenalin** und **Dopamin** aus den Speichervesikeln im zentralen Nervensystem freigesetzt.
ω_{unk} = 0.01 (imp)	Bei einer Überdosierung von **Fenetyllin** werden große Mengen der Neurotransmitter **Noradrenalin** und **Dopamin** aus den Speichervesikeln im zentralen Nervensystem freigesetzt.
ω_{unk} = 0.2 (imp)	Bei einer Überdosierung von **Fenetyllin** werden große Mengen der Neurotransmitter **Noradrenalin** und **Dopamin** aus den Speichervesikeln im zentralen Nervensystem freigesetzt.
ω_{unk} = 1.0 (imp)	Bei einer Überdosierung von Fenetyllin werden große Mengen der Neurotransmitter **Noradrenalin** und **Dopamin** aus den Speichervesikeln im zentralen Nervensystem freigesetzt.

Table 4: Original text samples (raw) and their annotation-imputed instances for certain ω_{unk} values. The text in **bold** denotes the annotated entities. The samples were chosen for illustration purposes. The annotation granularity reflects the token structure from the subword tokenizer of the GottBERT model.