Leveraging Wikidata for Biomedical Entity Linking in a Low-Resource Setting: A Case Study for German

Faizan E Mustafa
QUIBIQ GmbH

Corina Dima
University of Stuttgart

Juan G. Diaz Ochoa
PerMediQ GmbH
QUIBIQ GmbH

Steffen Staab
University of Stuttgart
University of Southampton

Abstract

Biomedical Entity Linking (BEL) is a challenging task for low-resource languages, due to the lack of appropriate resources: datasets, knowledge bases (KBs), and pre-trained models. In this paper, we propose an approach to create a biomedical knowledge base for German BEL using UMLS information from Wikidata, that provides good coverage and can be easily extended to further languages. As a further contribution, we adapt several existing approaches for use in the German BEL setup, and report on their results. The chosen methods include a sparse model using character n-grams, a multilingual biomedical entity linker, and two general-purpose text retrieval models. Our results show that a language-specific KB that provides good coverage leads to most improvement in entity linking performance, irrespective of the used model. The finetuned German BEL model, newly created UMLS$_{Wikidata}$ KB as well as the code to reproduce our results are publicly available[1].

1 Introduction

BEL is the task of disambiguating text spans by linking them to a unique identifier in a biomedical knowledge base (French and McInnes, 2023). For instance, the UMLS (Bodenreider, 2004) entity having the Concept Unique Identifier (CUI) `C0007765` is usually mentioned in English under the name *cerebellum* but can be mentioned in German using either *Kleinhirn, Cerebellum* or *Zerebellum*. Each *entity* in the KB has an *entity name* and one or multiple *aliases* associated with it, in multiple languages, as shown in Fig. 1. In this work, we refer to such names as *entity mentions*. The task of biomedical entity linking is to recover the unambiguous entity identifier from a KB given either of the names that can be used to refer to an entity. The task can be performed *with context* - where the name is provided together with the surrounding text, or *without context* - where only the name itself is provided for the disambiguation. In this paper we tackle the problem of BEL for entity mentions without context.

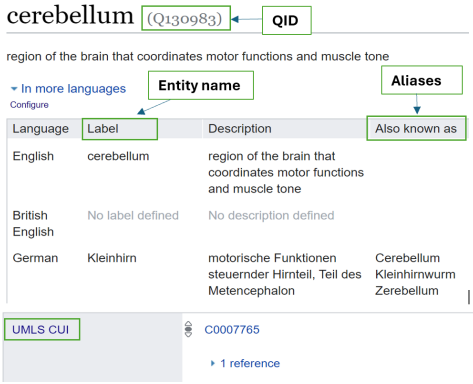

Figure 1: QID for an entity in Wikidata

A wide range of models using ruled-based and deep learning approaches for BEL have been proposed for English, for which many data resources are available (Shi et al., 2023). However, the in-domain BEL datasets, KBs, and models are scarce for low-resource languages. Multilingual biomedical models such as SapBERT (Liu et al., 2021a) have been proposed and evaluated on cross-lingual BEL benchmarks like XL-BEL (Liu et al., 2021b). This benchmark, however, is only intended for evaluation purposes, as it includes only 1,000 samples per language.

Wang et al. (2023) proposed a comprehensive German BEL benchmark, WikiMed-DE-BEL, which has, however, not yet been used for evaluating BEL models. We adapt several models from the literature to BEL on German, and evaluate them on this new benchmark.

A problematic aspect when training a BEL model for German is the lack of a biomedical KB with entity names and descriptions in German. The Unified Medical Language System (UMLS) (Bo-

[1] German-Bio-Entity-Linking GitHub Repository

denreider, 2004), the most comprehensive biomedical thesaurus available to date, which is the standard KB in BEL for English, only contains 1.6% entities with German descriptions (Liu et al., 2021b). We propose a solution to this problem by building a German biomedical KB using UMLS information harvested from Wikidata (Vrandečić and Krötzsch, 2014), an approach that leads to better entity coverage and can be extended to further languages.

2 Knowledge Bases

2.1 UMLS

UMLS (Bodenreider, 2004) is a metathesaurus integrating information from multiple biomedical vocabularies with the aim of improving interoperability. The terminology utilized across vocabularies is standardized by assigning a unique identifier, called the Concept Unique Identifier (CUI) to the same entities modeled in different vocabularies and across multiple languages. The latest UMLS Metathesaurus release, 2023AB, contains approximately 3.36 million concepts and 15.9 million unique concept names from 185 source vocabularies[2].

2.2 Wikidata

Wikidata (Vrandečić and Krötzsch, 2014) is a collaborative knowledge base providing the data for many Wikimedia projects, including the multilingual Wikipedia. Wikidata currently consists of more than 100M items that have been edited over 2 billion times by Wikidata users[3]. A defining trait of Wikidata is that it serves as a hub for integrating knowledge from different domains, including the biomedical domain. Wikidata entities can be connected, for example, to the UMLS, to the Disease Ontology or to many other biomedical vocabularies through pre-defined properties.

3 BEL Datasets for German

3.1 WikiMed-DE-BEL

WikiMed-DE (Wang et al., 2023) is a silver-standard biomedical entity linking dataset for the German language. It was built starting from German Wikipedia articles with hyperlinked text, where the hyperlinks are considered to be entity mentions and are linked to the corresponding Wikidata unique item identifiers (QIDs). The QIDs were then used to assign unique concept IDs from several biomedical vocabularies including UMLS. The annotations for each article include the article's title, text, QID, biomedical vocabularies concept IDs as well as a list of mentions, each assigned an unique QID as well as biomedical concept IDs. The creators of the WikiMed-DE dataset released a high-quality subset named WikiMed-DE-BEL which we use as a benchmark. WikiMed-DE-BEL includes 53,981 articles from the German Wikipedia. The `train`, `test` and `dev` splits follow the 80/10/10 rule.

We post-process WikiMed-DE-BEL as follows: for each data split, we only keep unique (mention, CUI) pairs. To increase the number of available pairs we create pairs both from the article title and the CUI assigned to the whole article, as well as from the entity mentions inside the article together with their assigned CUI. The `train`, `dev`, and `test` sets contain 42,679, 13,017, and 13,019 unique CUIs and 79,904, 19,561, and 19,203 (CUI, mention) pairs, respectively.

3.2 XL-BEL

XL-BEL (Liu et al., 2021b) is a cross-lingual biomedical entity linking evaluation benchmark that covers 10 languages, including German. Entity mentions from Wikipedia articles in the target languages were linked to language-agnostic UMLS CUIs using the methodology proposed by (Vashishth et al., 2021). The dataset samples are (sentence, mention, CUI) triples extracted from these Wikipedia articles. A number of 1,000 samples were retained for each language, making sure that each surface form appears only once in the sampled examples. We use the German subset of XL-BEL for evaluation purposes.

4 Models for German BEL

To the best of our knowledge, there are no existing dedicated models for German BEL that are publicly available. We therefore selected several models that could be adapted to German. Because we perform BEL without context, we also report on results obtained using embedding models trained for text retrieval. In this case, the evaluation is based on the nearest neighbour search, using the mention as an input query.

ScispaCy. Neumann et al. (2019) introduce ScispaCy, a Python library for biomedical text processing. One of the provided models creates sparse vector representations of the entity names and aliases

[2] 2023AB UMLS® Release Notes and Bugs.

[3] Wikidata statistics, accessed March 18th, 2024.

from the KB by representing them in terms of the TF-IDF scores of character 3-grams which are part of ten or more entities from the given KB. An entity mention is similarly modeled in terms of its character 3-grams and is linked to the KB by retrieving the k nearest neighbours from the KB. This process is language-agnostic and can easily be applied to other languages as long as there is an available KB in the target language.

SapBERT. Liu et al. (2021a) propose SapBERT, a pre-training scheme for self-aligning transformer-based representations to KBs based on synonymy relations. The training process takes as input a list of (mention, CUI) pairs from the KB, where the mention could be either the entity name or one of its aliases (see Fig. 1). The authors use a dedicated mining process to discover informative training examples: within a mini-batch they look for triples of the form (x_a, x_p, x_n) where x_a is an *anchor*, a random mention from the mini-batch, x_p is a *positive match* for x_a and x_n is a *negative match* for x_a. A positive match has the same CUI as the anchor mention, whereas a negative match has a different CUI than the anchor. For the example in Fig. 1 a triplet could be (*Kleinhirn*, *Cerebellum*, *Gehirn*), where the first two items in the triplet refer to the same CUI, `C0007765`, whereas *Gehirn*, German for *brain*, refers to a different CUI, `C0006104`. Each triplet contributes a positive pair and a negative pair towards the training data. The model is then trained using an adapted version of multi-similarity loss (Wang et al., 2019). The goal is to bring the representations of positive pairs closer to each other while pushing the negative pairs far from each other. Each mention is represented using the output `[CLS]` token resulting from feeding the mention text through the base transformer model.

M3 Embeddings. Chen et al. (2024) proposed a multilingual, hybrid text retrieval approach that can model input texts of up to 8192 tokens. A self-knowledge distillation framework is used to jointly learn three retrieval methods (dense, sparse, multi-vector) which reinforce each other. The model can be used for query-based text retrieval in more than 100 languages, including German.

Jina Embeddings. Mohr et al. (2024) developed an German-English bilingual model by pre-training a BERT-based language model on bilingual text. The model is then trained as an embedding retrieval model using contrastive learning by fine-tuning on text pairs (q, p) consisting of a query string q and a target string p. The evaluation indicate a considerable improvement in German-English cross-lingual retrieval performance when compared to multilingual models.

5 Creating a German Biomedical Knowledge Base

Liu et al. (2021b) report that 69.6% of the names of the UMLS entities in release 2020AA are in English, but only 1.6% are in German. The multilingual UMLS subset they use to evaluate SapBERT, $\text{UMLS}_{SapBERT}$, is provided by the SapBERT authors in their GitHub repository[4]. It contains 399,931 entity names or aliases assigned to 62,094 unique CUIs. Most of the names are in English, with only a small fraction being in German. The number of unique (entity, CUI) pairs amounts to 260,633.

We create a large German biomedical KB, $\text{UMLS}_{Wikidata}$, by leveraging Wikidata information. We first obtain a list of Wikidata QIDs that are annotated with CUIs by querying Wikidata using the official SPARQL endpoint[5] to fetch items that have the *UMLS CUI* property (`P2892`). The QIDs are further used to obtain the German label, description and alias(es) using the Python package *qwikidata* [6]. The resulting KB has 599,330 unique CUIs and 671,797 unique (entity name, CUI) pairs, where all the entity names are in German. Table 1 shows the statistics of the two KBs. $\text{UMLS}_{Wikidata}$ KB is made publicly available for further use[7].

	Unique CUIs	**Unique (CUI, Entity) Pairs**
$\text{UMLS}_{Wikidata}$	599,330	671,797
$\text{UMLS}_{SapBERT}$	62,094	260,633

Table 1: KB Statistics

6 Methodology

The first step in the evaluation of each of the selected models is to create vector representations for all KB entities using each model in turn and then store the obtained entity representations in a Faiss index (Johnson et al., 2019) for efficient retrieval.

The linking step for all the models involves first creating a vector representation for the entity mention using the selected model and then finding the k nearest neighbors from the KB by comparing the

[4] SapBERT UMLS subset.
[5] Official Wikidata SPARQL endpoint.
[6] `https://pypi.org/project/qwikidata/`
[7] https://zenodo.org/records/11003203

mention vector to the KB vector representations stored in the corresponding Faiss index using cosine similarity. Mentions are linked to the 5 nearest neighbors for all the models.

We further fine-tune the SapBERT-UMLS model[8], which is already trained on multilingual UMLS pairs, on $\text{UMLS}_{Wikidata}$. We use the same procedure as described in Section 4 and train for 5 epochs using a batch size of 256. The fine-tuned model is available on Hugging Face Model Hub[9]

The only hyperparameter of the ScispaCy model is the size of the character n-grams to be used. We use the 3-grams that appear in 10 or more entities in the target KB. We only use the dense representations from the M3 embedding model. For performance reasons, the maximum sequence length of all the embedding models is set to 40 tokens.

7 Results

The evaluation metric precision@k ($p@k$) indicates the percentage of samples where the correct entity is found in the top k KB entities predicted by a model. Tables 2 and 3 report the $p@1$ and $p@5$ obtained by the various models when linking against the $\text{UMLS}_{Wikidata}$ and the $\text{UMLS}_{SapBERT}$, respectively. As a general trend, the sparse, ScispaCy-based n-gram models score lower than the embedding models. The difference is more pronounced when using the $\text{UMLS}_{SapBERT}$ KB (in Table 3) because here the descriptions are mostly in English and thus the character 3-grams selected from the KB for the model have less overlap with the German mentions. The Jina embeddings outperform the rest of the embedding models when using the $\text{UMLS}_{Wikidata}$ knowledge base.

SapBERT fine-tuned on $\text{UMLS}_{Wikidata}$ offers good, consistent performance: it performs on par with the Jina model when using the $\text{UMLS}_{Wikidata}$ KB (see Table 2) and outperforms the rest of the models by a large margin, showing a 6 point improvement in $p@1$ score for XL-BEL when using the $\text{UMLS}_{SapBERT}$ KB (see Table 3). We hypothesize that this is due to the benefits of fine-tuning on the extra names contained in $\text{UMLS}_{Wikidata}$, as it allows the model to learn a better English-German cross-lingual mapping, as many medical terms are common between English and German.

Overall scores are much higher when using the $\text{UMLS}_{Wikidata}$ KB instead of the $\text{UMLS}_{SapBERT}$ KB because of its larger size and because it provides better coverage for the German entities in the two evaluation datasets. Moreover, the scores for WikiMed-DE-BEL are significantly higher than for XL-BEL when using the $\text{UMLS}_{Wikidata}$ KB but the opposite is true when using the $\text{UMLS}_{SapBERT}$ KB. The reason for this behaviour is discussed next.

[8] SapBERT-UMLS

[9] https://huggingface.co/permediq/SapBERT-DE

Model	Metrics	WikiMed-DE-BEL			XL-BEL DE
		Train	Dev	Test	
ScispaCy using $\text{UMLS}_{Wikidata}$ 3-grams	p@1	0.755	0.782	0.785	0.492
	p@5	0.824	0.847	0.851	0.590
SapBERT (Liu et al., 2021a)	p@1	0.756	0.783	0.785	0.462
	p@5	0.822	0.846	0.850	0.568
SapBERT fine-tuned on $\text{UMLS}_{Wikidata}$	p@1	0.774	0.796	0.80	0.485
	p@5	0.840	0.861	0.863	0.590
M3 embeddings	p@1	0.767	0.791	0.795	**0.499**
	p@5	0.836	0.857	0.860	0.604
Jina embeddings	p@1	**0.777**	**0.803**	**0.805**	0.495
	p@5	**0.840**	**0.861**	**0.864**	**0.605**

Table 2: Results using the $\text{UMLS}_{Wikidata}$ KB.

Model	Metrics	WikiMed-DE-BEL			XL-BEL DE
		Train	Dev	Test	
ScispaCy using $\text{UMLS}_{SapBERT}$ 3-grams	p@1	0.118	0.117	0.117	0.286
	p@5	0.141	0.141	0.142	0.359
SapBERT (Liu et al., 2021a)	p@1	0.139	0.147	0.146	0.346
	p@5	0.154	0.162	0.161	0.396
SapBERT fine-tuned on $\text{UMLS}_{Wikidata}$	p@1	**0.172**	**0.181**	**0.177**	**0.401**
	p@5	**0.197**	**0.206**	**0.204**	**0.473**
M3 embeddings	p@1	0.138	0.143	0.143	0.342
	p@5	0.155	0.160	0.160	0.401
Jina embeddings	p@1	0.141	0.148	0.149	0.338
	p@5	0.158	0.166	0.165	0.394

Table 3: Results using the $\text{UMLS}_{SapBERT}$ KB.

8 KB Coverage

The results obtained for the different dataset/KB combinations are drastically different. The precision is above 0.75 for WikiMed-DE-BEL using the $\text{UMLS}_{Wikidata}$ KB, but below 0.20 when using the $\text{UMLS}_{SapBERT}$ KB. If a mention's CUI is not present in KB then the model cannot link to it. Therefore, we check the upper limit for the metric scores by calculating the dataset coverage for the two KBs. Table 4 shows, for each dataset, the percentage of dataset CUIs that are present in the KB CUIs. It can be noticed that only 36% of the WikiMed training set CUIs are present in the $\text{UMLS}_{SapBERT}$ KB, in contrast to 98% coverage when using the $\text{UMLS}_{Wikidata}$ KB.

KB	WikiMed-DE-BEL			XL-BEL DE
	Train	Dev	Test	
$\text{UMLS}_{Wikidata}$	98.2%	97.5%	97.6%	81.0%
$\text{UMLS}_{SapBERT}$	36.4%	36.7%	37.1%	99.8%

Table 4: CUI coverage.

Another problematic setup is when the a particular name of an entity or alias is not present in the

KB, even its CUI is in KB. Therefore, we compute the (mention, CUI) pair coverage by looking at the percentage of (mention, CUI) pairs present in the respective KB. Table 5 shows that the pair coverage for XL-BEL wrt. to $\text{UMLS}_{SapBERT}$ is 11%, whereas for WikiMed-DE-BEL wrt. to $\text{UMLS}_{Wikidata}$ is 56% — which aligns better with the model performance reported in Tables 2 and 3.

KB	WikiMed-DE-BEL			XL-BEL DE
	Train	Dev	Test	
$\text{UMLS}_{Wikidata}$	56.5%	62.4%	63.0%	33.8%
$\text{UMLS}_{SapBERT}$	6.2%	5.9%	6%	11.8%

Table 5: (mention, CUI) pairs coverage.

9 Conclusion

The unavailability of knowledge bases, datasets and, subsequently, models makes BEL a challenging task for low-resource languages. To this end, we propose an approach to create a KB for German BEL, $\text{UMLS}_{Wikidata}$, using a methodology that can be easily applied to further low-resource languages. We further compare four different models with various representations and trained on different languages. Our results show that creating a dedicated, large-scale knowledge base in the target language leads to the most improvement for doing entity linking in that language, independently of the used model. The best BEL results for German are obtained using the language-specific $\text{UMLS}_{Wikidata}$ knowledge base.

Acknowledgements

This project was generously supported by the Ministry for Economics, Labor and Tourism from Baden-Württemberg, Germany via grant agreement number BW1_1456 (AI4MedCode). We would like to thank Klinikum Stuttgart for their valuable contribution and reviewers for their feedback.

References

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database-Issue):267–270.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation.

Evan French and Bridget T. McInnes. 2023. An overview of biomedical entity linking throughout the years. *Journal of Biomedical Informatics*, 137:104252.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

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, Online. Association for Computational Linguistics.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021b. Learning domain-specialised representations for cross-lingual biomedical entity linking. 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 565–574, Online. Association for Computational Linguistics.

Isabelle Mohr, Markus Krimmel, Saba Sturua, Mohammad Kalim Akram, Andreas Koukounas, Michael Günther, Georgios Mastrapas, Vinit Ravishankar, Joan Fontanals Martínez, Feng Wang, Qi Liu, Ziniu Yu, Jie Fu, Saahil Ognawala, Susana Guzman, Bo Wang, Maximilian Werk, Nan Wang, and Han Xiao. 2024. Multi-Task Contrastive Learning for 8192-Token Bilingual Text Embeddings.

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, Florence, Italy. Association for Computational Linguistics.

Jiyun Shi, Zhimeng Yuan, Wenxuan Guo, Chen Ma, Jiehao Chen, and Meihui Zhang. 2023. Knowledge-graph-enabled biomedical entity linking: a survey. *World Wide Web*, pages 1–30.

Shikhar Vashishth, Denis Newman-Griffis, Rishabh Joshi, Ritam Dutt, and Carolyn P. Rosé. 2021. Improving broad-coverage medical entity linking with semantic type prediction and large-scale datasets. *Journal of Biomedical Informatics*, 121:103880.

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

Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R. Scott. 2019. Multi-Similarity Loss With General Pair Weighting for Deep Metric Learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5017–5025.

Yi Wang, Corina Dima, and Steffen Staab. 2023. WikiMed-DE: Constructing a Silver-Standard Dataset for German Biomedical Entity Linking using Wikipedia and Wikidata. In *Proceedings of the Wikidata Workshop 2023 co-located with 22nd International Semantic Web Conference (ISWC 2023)*.