Generative Biomedical Entity Linking via Knowledge Base-Guided Pre-training and Synonyms-Aware Fine-tuning

Hongyi Yuan ∗ Zheng Yuan ∗ Sheng Yu †
Tsinghua University
{yuanhy20,yuanz17}@mails.tsinghua.edu.cn
syu@tsinghua.edu.cn

Abstract

Entities lie in the heart of biomedical natural language understanding, and the biomedical entity linking (EL) task remains challenging due to the fine-grained and diversiform concept names. Generative methods achieve remarkable performances in general domain EL with less memory usage while requiring expensive pre-training. Previous biomedical EL methods leverage synonyms from knowledge bases (KB) which is not trivial to inject into a generative method. In this work, we use a generative approach to model biomedical EL and propose to inject synonyms knowledge in it. We propose KB-guided pre-training by constructing synthetic samples with synonyms and definitions from KB and require the model to recover concept names. We also propose synonyms-aware fine-tuning to select concept names for training, and propose decoder prompt and multi-synonyms constrained prefix tree for inference. Our method achieves state-of-the-art results on several biomedical EL tasks without candidate selection which displays the effectiveness of proposed pre-training and fine-tuning strategies. The source code is available at Github.com/Yuanhy1997/GenBioEL.

1 Introduction

Biomedical entity linking (EL) refers to mapping a biomedical mention (i.e., entity) in free texts to its concept in a biomedical knowledge base (KB) (e.g., UMLS (Bodenreider, 2004)). This is one of the most concerned tasks of research in medical natural language processing which is highly related to high-throughput phenotyping (Yu et al., 2015), relation extraction (Li et al., 2016), and automatic diagnosis (Yuan and Yu, 2021).

Recent methods in biomedical EL mainly used neural networks to encode mentions and each concept name into the same dense space, then linked mentions to corresponding concepts depending on embedding similarities (Sung et al., 2020; Lai et al., 2021; Bhowmik et al., 2021; Ujiie et al., 2021; Agarwal et al., 2021). Synonyms knowledge has been injected into these similarity-based methods by contrastive learning (Liu et al., 2021a; Yuan et al., 2022). For example in UMLS, concept C0085435 has synonyms: Reiter syndrome, Reactive arthritis and ReA which help models to learn different names of a concept entity. However, similarity-based methods require large memory footprints to store representation for each concept which are hard to deploy.

In general domain, GENRE (Cao et al., 2021a) viewed EL as a seq2seq task which inputs mentions with contexts and outputs concept names token-by-token. Mentions, contexts and concepts can be mingled due to the power of transformers, and the model does not need to store representations for each concept during inference. GENRE pre-trained on Wikipedia EL datasets to boost performances. However, directly implementing GENRE on biomedical EL cannot harvest satisfying results. The gap occurs in two aspects: (1) There are no such large-scale human-labeled biomedical EL datasets for pre-training. (2) Biomedical concepts may have multiple synonyms. We find the results are sensitive to the synonyms selection for training, and simply using a 1-to-1 mapping between names and concepts as Cao et al. (2021a,b) may hurt performances.

To address the above issues, we propose KB-guided pre-training and synonyms-aware fine-tuning to improve generative EL. For pre-training, we construct pre-training samples using synonyms and definitions collected from KBs and sentence templates. KB-guided pre-training has the same format as seq2seq EL, which fills the gap of loss of pre-training corpus. Compared to the method introduced in Cao et al. (2021a), ours performs better in biomedical EL with fewer resources. For fine-
tuning, we propose decoder prompts to highlight mentions. We find the model tends to generate textually similar names to mentions. Hence textual similar criterion is proposed for selecting concept names during fine-tuning. During inference, we propose a multi-synonyms constrained prefix tree, which results in significantly improved performance. The overview of our approach is illustrated in Figure 1.

We conduct experiments on various biomedical EL datasets and achieve SOTA results on COMETA, BC5CDR, and AskAPatient (AAP) even without candidate selection. Extensive studies show the effectiveness of our proposed pre-training and fine-tuning schemes.

2 Approach

Define a set of concepts $\mathcal{E}$ as target concepts (i.e. concepts from target KBs). For each concept $e \in \mathcal{E}$, we have a set of synonyms names $f(e) = \{s^e_i|i \in \{1, \ldots, n_e\}\}$. All the synonyms forms a name set $\mathcal{S} = \bigcup_{e \in \mathcal{E}}\{s^e_i|i \in \{1, \ldots, n_e\}\}$. Names-to-concept mappings can be defined by: $\sigma(s^e_i) = e$ where $s^e_i = f^{-1}$. For mention $m$ with left and right contexts $c_l$ and $c_r$, which gold label is $e_m \in \mathcal{E}$, we need to find the target concept $e_m \in \mathcal{E}$. $m$, $c$, and $s$ comprise a sequence of tokens.

2.1 Seq2seq EL

Our model applies an encoder-decoder transformer architecture following Cao et al. (2021a). The encoder input is: $[\text{BOS}] q [\text{ST}] m [\text{ET}] c_l [\text{EOS}]$, where [ST] and [ET] are the special tokens marking $m$. For the decoder side, unlike GENRE decoding target names directly, we use simple prefix prompts $P_m = \langle m \rangle$ to strengthen the interaction between mentions and make the decoder side output resemble a natural language sentence: $[\text{BOS}] m$ is $s$, where $s$ is a target name belonging to label concept $e$. The training objective of Seq2Seq EL is to maximize the likelihood:

$$p_\theta(s | P_m, c, m) = \prod_{i=1}^{N_s} p_\theta(y_i | y_{<i}, P_m, c, m),$$

where $N_s$ is the number of tokens of $s$ and $y_i$ indicates the $i$th token. The inference of Seq2seq EL applies beam search (Sutskever et al., 2014) with targets constrained to the name set $\mathcal{S}$ by a prefix tree (constructed by name set $\mathcal{S}$). Unlike mGENRE Cao et al. (2021b) using provided candidates to decrease the size of the prefix tree, we use the whole name set $\mathcal{S}$ instead.

2.2 KB-Guided Pre-training

As the training data of EL is tiny compared to the vast number of concepts in KB, thus it makes EL for some mentions zero-shot problems. We want to leverage synonyms knowledge from KB to enhance the model’s performance. Injecting synonyms knowledge to the encoder-only models can be done by contrastive learning (Liu et al., 2021a; Yuan et al., 2022). However, such a paradigm cannot directly apply to encoder-decoder architecture as entities are not represented by dense embeddings. To mitigate this problem, we construct a pre-training task that shares a similar form as Section 2.1. We manually define a set of clause templates to splice with synonyms and definitions in KB to form input synthetic language discourses. Concretely, we select two synonyms $s^a_e$ and $s^b_e$ and definition $c_e$ of a concept $e \in \mathcal{E}$. Then we randomly pick a template to concatenate them to form the encoder input, here we give two examples:

$$[\text{BOS}] [\text{ST}] s^a_e [\text{ET}] \text{ is defined as } c_e [\text{EOS}]$$
$$[\text{BOS}] c_e \text{ describes } [\text{ST}] s^a_e [\text{ET}] [\text{EOS}]$$

For the decoder: $[\text{BOS}] \quad s^a_e \text{ is } s^b_e, c_e$ is the simulated context and $s^b_e$ is for model to predict. If definitions are absent in KB, we will use other synonyms to construct $c_e$. All templates we used can be found in Appendix E.2.

2.3 Synonyms-Aware Fine-tuning

We propose and validate by experiments in Section 4 that seq2seq EL is profoundly influenced by the textual similarity between mentions and concept names. It tends to generate textually similar names. We select the target name by calculating the character 3-gram TF-IDF similarity (Neumann et al., 2019) between mention $m$ and all synonyms $\{s^e_i\}$ and choosing the most similar one as

$$s = \underset{s \in \{s^e_i\}}{\text{arg max}} \cos(\text{TFIDF}(m), \text{TFIDF}(s)).$$

By the textual similarity criterion, we manually reduce the difficulty of fine-tuning. We do not use this criterion for pre-training since we want it to learn various synonyms to improve generalization.
Different from GENRE using only one canonical name for each concept for prefix tree, we use multi-synonym names (i.e. $\mathcal{S}$) to construct prefix tree. During inference, we apply prefix tree constrained beam search to decode the name $\hat{s}_m$, and map to the concept $\hat{e}_m = \sigma(\hat{s}_m)$ via N-to-1 names-to-concept mapping $\sigma$.

3 Experiments

3.1 Datasets and KBs

Pre-training. We use a subset of UMLS st21pv (Mohan and Li, 2019) as $\mathcal{E}$ for pre-training. It contains 2.37M concepts, where 160K concepts contain definitions and 1.11M concepts have multiple synonyms. While pre-training, we iterate concepts and synonyms to construct inputs and outputs.

Fine-tuning. We evaluate our model on BC5CDR (Li et al., 2016), NCBI (Doğan et al., 2014), COMETA (Basaldella et al., 2020) and AAP (Limsopatham and Collier, 2016). These benchmarks focus on different entity types, including disease, chemicals, and colloquial terms. For fair comparison, we follow Varma et al. (2021) in BC5CDR, Lai et al. (2021) in NCBI and COMETA, and Limsopatham and Collier (2016) in AAP to construct dataset splits and target KB concepts $\mathcal{E}$. Name set $\mathcal{S}$ is constructed by synonyms from UMLS and original KB which is detailed in Appendix E.1. Datasets summaries are shown in Appendix A. We use recall@$1/0.5$ as metrics for performance illustration, and recall@$5$ results are listed in Appendix C.

3.2 Implementation Details

We use BART-large (Lewis et al., 2020) as the model backbone. We pre-train and fine-tune our model using teacher forcing (Williams and Zipser, 1989) and label smoothing (Szegedy et al., 2016) which are standard in seq2seq training. The hyperparameters can be found in Appendix E.3.

3.3 Main Results

Table 1 and 2 compare the recall@$1$ with state-of-the-art (SOTA) methods. Our method with KB-guided pre-training exceeds the previous SOTA on BC5CDR, COMETA and AAP, which is up to 1.4 on BC5CDR, 1.3 on COMETA and 0.3 on AAP. Our method also shows superiority over previous SOTA on BC5CDR and COMETA without pretraining. Besides, our method shows competitive results on NCBI compared to SOTA, and we further analyse the results of NCBI through case studies in Appendix D.

4 Discussion

Does pre-training help? On all datasets, KB-guided pre-training improves fine-tuning consis-
Table 3: Recall@1 for BC5CDR and COMETA test dataset using different initial checkpoints. We only use left context for BART in decoding and omit decoding prompts for GENRE when not fine-tuning which are consistent with their pre-training.

| Init | BC5CDR | COMETA |
|------|--------|--------|
| Not Fine-tune | | |
| BART | 6.2 | 4.1 |
| GENRE | 38.3 | 23.8 |
| KB-guided | 42.4 | 33.1 |
| Synonyms-aware Fine-tune | | |
| BART | 92.5 | 80.9 |
| GENRE | 92.9 | 80.8 |
| KB-guided | **93.3** | **81.4** |

Table 4: Recall@1 for BC5CDR and COMETA test dataset using different $s$ for training, different $S$ for inference and applying decoder prompts or not.

| $s$ | $S$ | Prompt | BC5CDR | COMETA |
|-----|-----|--------|--------|--------|
| TF-IDF | All | ✓ | **93.3** | **81.4** |
| Shortest | All | ✓ | 87.2 | 80.6 |
| Sample | All | ✓ | 86.9 | 80.8 |
| Shortest | Shortest | ✓ | 76.3 | 77.5 |
| Sample | Sample | ✓ | 72.4 | 77.8 |
| TF-IDF | All | × | 93.0 | 80.9 |

Decoding similarity sample groups on COMETA trained with using all synonyms as $S$ and TF-IDF for selecting target names $s$. From Figure 2, we find the distribution of TF-IDF similarity between mentions and selected names is polarized, and the accuracy increases along with textual similarities which prove textually similar targets are easy to generate. This phenomenon validates the advantage of selecting textually similar names for fine-tuning.

**Decoder Prompting** One difference between GENRE and ours is using prompt tokens on the decoder side. Prompting has shown improvement on various NLP tasks (Liu et al., 2021b). Here, prompt tokens serve as informative hints by providing additional decoder attention queries and making the outputs resemble language models’ pre-training tasks. We test dropping the prompt tokens, and Table 4 shows degraded performances (-0.3 on BC5CDR and -0.5 on COMETA). The results illustrate the improvement brought by decoder prompting.

**Sub-population Analysis** We list several sub-populations of the BC5CDR benchmark to illustrate the model’s performance on different fine-grained categories of mentions. The details of sub-populations are shown in Appendix C.2.

Our model’s performance on different sub-populations of BC5CDR is shown in Table 9. Through the results, we have several findings:

1. Compared with Varma et al. (2021), our method shows superiority over most of the sub-populations of BC5CDR. Our method without pre-training outperforms the data-augmented version of Varma et al. (2021).

2. Our method surpasses Varma et al. (2021) by the largest margin on Unseen Concepts. One possible explanation is that our generative method learns the linkage between mentions and contextual information better, thus gaining superior zero-shot-shot performance.
Table 5: Accuracy over sub-populations on BC5CDR of our proposed methods and Varma et al. (2021). The best results are presented in bold letters and the second best results are highlighted by underlines.

| Subset                | Varma et al. (2021) | Ours       | Sample Size |
|-----------------------|----------------------|------------|-------------|
|                       | Baseline | Full | PT | PT + FT |             |
| Overall               | 89.5 | 91.3 | 92.5 | **93.4** | 9.65k         |
| Single Word Mentions  | 91.4 | 92.9 | 95.8 | **96.6** | 7.04k         |
| Multi-Word Mentions   | 84.5 | **86.8** | 85.3 | 84.8 | 2.62k         |
| Unseen Mentions       | 75.3 | 79.6 | 81.8 | **83.3** | 3.28k         |
| Unseen Concepts       | 69.7 | 77.5 | 86.3 | **86.9** | 2.16k         |
| Not Direct Match      | 89.4 | **91.9** | 83.0 | 84.3 | 3.83k         |
| Top 100               | 97.3 | 97.2 | **97.9** | **98.1** | 3.31k         |

3. Our KB-guided pre-training gains improvement universally on most subsets. This reflects the effectiveness of the pre-training task.

4. Popular concepts and single-word mentions are more easily resolved compared to unseen mentions or concepts and multi-word mentions, respectively. Unsurprisingly, mentions with less training resources and longer textual forms are more difficult. The lengths of mentions may challenge the generative model.

5 Related Work

Biomedical EL is an important task in biomedical NLP. Classification-based methods used a softmax layer for classification (Limsopatham and Collier, 2016; Miftahutdinov and Tutubalina, 2019) which consider concepts as category factors and lost information of concept names. Recent methods (Sung et al., 2020; Liu et al., 2021a; Lai et al., 2021; Bhowmik et al., 2021; Ujiie et al., 2021; Agarwal et al., 2021; Yuan et al., 2022) encoded mentions and names into a common space and disambiguated mentions by nearest neighbors. Angell et al. (2021) and Varma et al. (2021) adopted a retrieve-and-rerank framework to boost performances. Varma et al. (2021) emphasized the lack of training samples in EL and augmented data using Wikipedia and PubMed, while our pre-training corpus constructed by KB and templates can serve as good supplementary training data.

Generative EL Cao et al. (2021a) proposed to view EL as a seq2seq problem that got rid of hard negative sampling during training and required less memory at inference. The shortage is it demanded vast training sources (11 GB training data and 32 GPU days) to achieve competitive performance. Cao et al. (2021b) explored dealing synonyms in multilingual generative EL by adding language identifiers that cannot be directly implemented in biomedical EL.

6 Conclusion

To the best of our knowledge, our work is the first to explore generative EL in the biomedical domain. We inject synonyms and definition knowledge into the generative language model by KG-guided pre-training. We emphasize the synonym selection issue and propose synonyms-aware fine-tuning by considering the textual similarity. Decoding prompts are also introduced to improve the model’s performance. Our model sets new state-of-the-art on different biomedical EL benchmarks. GENRE shows that well-selected candidate sets can improve seq2seq EL, and we believe this will further boost our performances.

Acknowledgements

We thank Shengxuan Luo and Chuanqi Tan for fruitful discussions and advises, and Xixi Mo for drawing the overview figure. We appreciate the anonymous reviewers for their helpful comments and suggestions. This work was supported by the National Natural Science Foundation of China (Grant No. 12171270), the Natural Science Foundation of Beijing Municipality (Grant No. Z190024), and the International Digital Economy Academy.
References

Dhruv Agarwal, Rico Angell, Nicholas Monath, and Andrew McCallum. 2021. Entity linking and discovery via arborescence-based supervised clustering.

Rico Angell, Nicholas Monath, Sunil Mohan, Nishant Yadav, and Andrew McCallum. 2021. Clustering-based inference for biomedical entity linking. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2598–2608, Online. Association for Computational Linguistics.

Marco Basaldella, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. 2020. COMETA: A corpus for medical entity linking in the social media. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 3122–3137, Online. Association for Computational Linguistics.

Rajarshi Bhowmik, Karl Stratos, and Gerard de Melo. 2021. Fast and effective biomedical entity linking using a dual encoder. arXiv preprint arXiv:2103.05028.

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

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021a. Autoregressive entity retrieval. In International Conference on Learning Representations.

Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhonov, Luke Zettlemoyer, Nicola Cancelda, Sebastian Riedel, and Fabio Petroni. 2021b. Multilingual autoregressive entity linking. In arXiv pre-print 2103.12528.

Allan Peter Davis, Thomas C Wiegers, Michael C Rosenstein, and Carolyn J Mattingly. 2012. Medic: a practical disease vocabulary used at the comparative toxicogenomics database. Database, 2012.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Nebi disease corpus: a resource for disease name recognition and concept normalization. Journal of biomedical informatics, 47:1–10.

Tuan Lai, Heng Ji, and ChengXiang Zhai. 2021. BERT might be overkill: A tiny but effective biomedical entity linker based on residual convolutional neural networks. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1631–1639, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. Database, 2016.

Nut Limsopatham and Nigel Collier. 2016. Normalising medical concepts in social media texts by learning semantic representation. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1014–1023, Berlin, Germany. Association for Computational Linguistics.

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.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021b. Pretrain, prompt, and predict: A systematic survey of prompting methods in natural language processing. arXiv preprint arXiv:2107.13586.

Zulfat Miftahutdinov and Elena Tutubalina. 2019. Deep neural models for medical concept normalization in user-generated texts. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 393–399, Florence, Italy. Association for Computational Linguistics.

Sunil Mohan and Donghui Li. 2019. Medmentions: A large biomedical corpus annotated with [umls] concepts. In Automated Knowledge Base Construction (AKBC).

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.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC ’20. IEEE Press.

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

We pre-process the datasets by the following procedures: (1) the abbreviations in the texts are expanded using AB3P (Sohn et al., 2008); (2) the texts are lower-cased, and the beginning and ending of a mention are marked by two special tokens \[\text{ST}\] and \[\text{ET}\]; (3) the overlapping mentions and mentions absent from the target KB are discarded.

BC5CDR (Li et al., 2016) is a benchmark for biomedical entity recognition and disambiguation. The dataset annotates 1500 PubMed article abstracts with 4409 chemicals, 5818 diseases entities, and 3116 chemical-disease interactions. All the annotated entities are mapped to MeSH ontology, which is a smaller medical vocabulary that comprises a subset of UMLS (Bodenreider, 2004). In this work, in consideration of fairness, we follow two most recent works (Angell et al., 2021; Varma et al., 2021) that use MeSH contained in UMLS 2017 AA release to construct the target knowledge base.

NCBI (Doğan et al., 2014) contains a corpus of 793 PubMed abstracts. It consists of 6892 annotated disease mentions of 790 unique disease concepts. All the mentions are labeled with concepts in MELIC ontology (Davis et al., 2012). MELIC is a medical dictionary that merges the diseases concepts, synonyms, and definitions in MeSH and OMIM and is composed of 9700 unique diseases. In our work, we used the processed data and the target ontology provided by BioSyn (Sung et al., 2020) and ResCNN (Lai et al., 2021). We followed their works to construct our training, developing, and testing data.

COMETA (Basaldella et al., 2020) consists of 20k English biomedical entity mentions from publicly available and anonymous health discussions on Reddit. All the mentions are expert-annotated with concepts from SNOMEL CT. We use the “stratified (general)” split and follow the evaluation protocol of SapBert (Liu et al., 2021a) and ResCNN (Lai et al., 2021).

AskAPatient (Limsopatham and Collier, 2016) is a dataset containing 8,662 phrases of social media language. Each phrase can be mapped to one of the 1,036 medical concepts from SNOMEL-CT and AMT (the Australian Medicines Terminology). The samples in AskAPatient do not include contextual information and mentions can only be disambiguated by phrases per se. We follow the experimental settings from works of Sung et al. (2020) and Limsopatham and Collier (2016) and apply the 10-fold evaluation protocol.

Statistics of above-mentioned datasets are listed in Table 6.
Table 6: Basic statistics of NCBI-disease, BC5CDR, COMETA, AskAPatient datasets and the corresponding target knowledge bases.

| Dataset     | NCBI | BC5CDR | COMETA | AskAPatient |
|-------------|------|--------|--------|-------------|
| Target names | 108,071 | 809,929 | 904,798 | 3,381       |
| Target concepts | 14,967 | 268,162 | 350,830 | 1,036       |
| Train samples  | 5784 | 268,162 | 350,830 | 1,036       |
| Dev samples    | 787 | 9,515  | 2,176  | 1,663       |
| Test samples   | 960 | 9,654  | 4,350  | 1,712       |

Table 7: The Recall@5 results of our proposed method on different benchmarks.

| Dataset | Recall@5 Results |
|---------|------------------|
| FT      | 95.6±0.1         |
| PT+FT   | 96.3±0.1         |

B License and Availability of Resources

BC5CDR, NCBI, COMETA, and AskAPatients are all publicly available datasets on the Internet. Their target KBs MeSH, MELIC, and SNOMELCT are covered by UMLS Metathesaurus License. One can require such a license by signing up for a UMLS terminology services account to access KBs mentioned above.

C Additional Experiment Results

C.1 Recall@5 Results

We show the Recall@5 result of our method with and without pre-training in Table 7.

C.2 Sub-population Analysis

Following Varma et al. (2021), we split test samples into different sub-populations. The details of different sub-population categories we use is shown in Table 8.

D Case Study

We provide case studies on the NCBI-disease benchmark to give an insight and justification of the performance of our method. For the case of mention colorectal adenomas which is annotated with D018256 adenomatous polyp, the mention exists for 6 times in the test set (account for 0.63 score of Recall@1). Our model fails to correctly disambiguate all such mentions while linking the mention to other concepts D000236 colorectal adenomas or D003123 hereditary nonpolyposis colorectal cancer. In the training set, the mention colorectal adenomas exists for two times and is annotated with D003132 and D000236 respectively. Thus, through this case, we can see (1) our model learns the information contained in the training set; (2) such inconsistent test samples are hard to disambiguate correctly.

E Implementation Details

In this section, we provide more details of our experiments.

E.1 Knowledge Base Pre-processing

Given different knowledge bases, we pre-process their content by the following procedure:

1. For each concept, we include all its synonyms from the original target KB. We also expand the synonym set for each concept using UMLS. We use the 2017 AA Active Release of UMLS.

2. For synonyms in the expanded name set $S$, we lowercase them and remove the symbols (e.g., dash line - or comma ,).

3. There may exist a name as a synonym for multiple concepts. We de-duplicate these overlapped synonyms by removing the synonym from the concept with more other synonyms to avoid the unbalanced number of synonyms in each concept.

It is worth noticing that we do not de-duplicate the synonyms in the target KB of NCBI in consideration of comparison fairness. In the previous works, a mention link to multiple concepts, and correct disambiguation is claimed if the target concept is hit by one of the predicted concepts in NCBI.

E.2 Pre-training Clause Templates

We list pre-training clause templates we used in Table 10. For those concepts containing only 2 synonyms, $s^w_a$ and $s^w_b$ are the two synonyms respectively and $c^w_e$ is the same as $s^w_b$. For those concepts containing only 1 sole synonym, $s^A_a$, $s^B_b$ and $c^e_c$ are the same.
Table 8: Accuracy over sub-populations on BC5CDR of our proposed methods and Varma et al. (2021).

| Subset                        | Varma et al. (2021) | Ours | Sample Size |
|-------------------------------|----------------------|------|-------------|
|                               | Baseline             | Full | PT | PT + FT |
| Overall                       | 89.5                 | 91.3 | 92.5 | 93.4 | 9.65k |
| Single Word Mentions          | 91.4                 | 92.9 | 95.8 | 96.6 | 7.04k |
| Multi-Word Mentions           | 84.5                 | 86.8 | 85.3 | 84.8 | 2.62k |
| Unseen Mentions               | 75.3                 | 79.6 | 81.8 | 83.3 | 3.28k |
| Unseen Concepts               | 69.7                 | 77.5 | 86.3 | 86.9 | 2.16k |
| Not Direct Match              | 89.4                 | 91.9 | 83.0 | 84.3 | 3.83k |
| Top 100                       | 97.3                 | 97.2 | 97.9 | 98.1 | 3.31k |

Table 9: Accuracy over sub-populations on BC5CDR of our proposed methods and Varma et al. (2021). The best results are presented in bold letters and the second best results are highlighted by underlines.

Table 10: The templates used for constructing pre-training samples for different kinds of concepts. $s_a^e$ is the selected synonym as the input mention, $s_b^e$ is the synonym selected as the decoding target. $c_e$ is the contextual information comprised by definition, other synonyms, or mention itself. The template words are between <>, and we omit the special tokens for marking mentions for conciseness.
E.3 Experiment Parameters

Our model contains 406M parameters with 12-layer transformer encoders and 12-layer transformer decoders. We list the hyper-parameters of our model for KB-guided pre-training and synonyms-aware fine-tuning on different benchmarks in Table 11. For pre-training, we heuristically select our parameters. For fine-tuning, we tune training steps among \{20000, 30000, 40000\} on development set. For BC5CDR and COMETA, we use learning rate as $1e-5$ and warmup steps as 500. For AskAPatient and NCBI, we search learning rate among \{5e-6, 8e-7, 3e-7\}, and do not use warmup. We only evaluate our model at the end of training. For each benchmark, we run three times to calculate means and standard deviations.

E.4 Computational Resource

For our KB-guided pre-training, we implement our model on 6 A100 GPU with 40 GB memory with the help of DeepSpeed ZeRO 2 (Rajbhandari et al., 2020) and train for 1 day. For fine-tuning on different benchmarks, we implement our model on 1 A100 GPU.
Table 11: Hyper-parameters for KB-guided pre-training and synonyms-aware fine-tuning.

| Parameters            | Pre-train | BC5CDR | COMETA | AskAPatient | NCBI     |
|-----------------------|-----------|--------|--------|-------------|----------|
| Training Steps        | 80,000    | 20,000 | 40,000 | 30,000      | 20,000   |
| Learning Rate         | 4e-5      | 1e-5   | 1e-5   | 5e-6        | 3e-7     |
| Weight Decay          | 0.01      | 0.01   | 0.01   | 0.01        | 0.01     |
| Batch Size            | 384       | 8      | 8      | 8           | 8        |
| Adam $\epsilon$       | 1e-8      | 1e-8   | 1e-8   | 1e-8        | 1e-8     |
| Adam $\beta$          | (0.9, 0.999) | (0.9, 0.999) | (0.9, 0.999) | (0.9, 0.999) | (0.9, 0.999) |
| Warmup Steps          | 1,600     | 500    | 500    | 0           | 0        |
| Attention Dropout     | 0.1       | 0.1    | 0.1    | 0.1         | 0.1      |
| Clipping Grad         | 0.1       | 0.1    | 0.1    | 0.1         | 0.1      |
| Label Smoothing       | 0.1       | 0.1    | 0.1    | 0.1         | 0.1      |