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

Anonymous ACL submission

Abstract

Entities lie in the heart of biomedical natural language understanding, and the biomedical entity disambiguation (ED) task remains challenging due to the fine-grained and diversiform concept names. Generative methods achieve remarkable performances in general domain ED with less memory usage while requiring expensive pre-training. Previous biomedical ED 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 ED 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 ED tasks without candidate selection which displays the effectiveness of proposed pre-training and fine-tuning strategies.

1 Introduction

Biomedical entity disambiguation (ED) 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 ED mainly used neural networks to encode mentions and each concept name into the same dense space, then disambiguated mentions 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. However, similarity-based methods requires large memory footprints to store representation for each concept which are hard to deploy.

In general domain, GENRE (Cao et al., 2021a) viewed ED 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 when inference. GENRE pre-trained on Wikipedia ED datasets to boost performances. However, directly implementing GENRE on biomedical ED cannot harvest satisfying results. The gap occurs in two aspects: (1) There are no such large-scale human-labeled biomedical ED datasets for pre-training. (2) Biomedical concepts come with synonyms, we find selecting synonyms for training is sensitive to the results, 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 ED. For **pre-training**, we construct pre-training samples from synonyms and definitions collected by KBs using sentence templates. This pre-training has the same format as seq2seq ED, 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 ED 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. When 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 ED 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.

2 Approach

Define a set of concepts $\mathcal{E}$ as target concepts (i.e. 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 $\sigma = f^{-1}$. 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 $\hat{e}_m \in \mathcal{E}$.

2.1 Seq2seq ED

Our model applies an encoder-decoder architecture following Cao et al. (2021a). The encoder input is: `[BOS]`c_{l}`[ST]`m`[ED]`c_{r}`[EOS]`, where `[ST]` and `[ED]` are the special tokens marking m. For the decoder side, unlike GENRE decoding target names directly, we use simple prefix prompts P_m =<m `is`> to strengthen the interaction between mentions and make the decoder side output resemble a natural language sentence: `[BOS]` m `is` s, where s is a target name belong to label concept e. The training objective of Seq2Seq ED 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 ith token. The inference of Seq2seq ED applies beam search (Sutskever et al., 2014) with targets constrained to the name set $\mathcal{S}$ by a prefix trie (constructed by name set $\mathcal{S}$). Unlike GENRE use provided candidates to decrease the size of the prefix tree, we use the whole name set instead.

2.2 KB-Guided Pre-training

As the training data of ED is tiny compared to the vast number of concepts in KB, thus it makes ED for some mentions zero-shot problems. We want to leverage synonyms knowledge from KB to

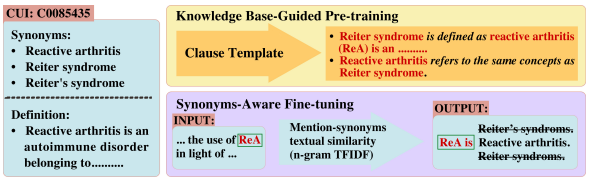

Figure 1: Overview of our approach.

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 propose a set of clause templates to splice with synonyms and definitions in KB to form synthetic language discourses. We select two synonyms s_e^a and s_e^b 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:

`[BOS]` `[ST]` s_e^a `[ED]` **is defined as** c_e `[EOS]`

`[BOS]` c_e **describes** `[ST]` s_e^a `[ED]` `[EOS]`

For the decoder: `[BOS]` s_e^a `is` s_e^b. c_e is the simulated context and s_e^b 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 ED 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

$$s = \arg\max_{s \in \{s_e^i\}} \cos(\text{TFIDF}(m), \text{TFIDF}(s)).$$

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.

Model	BC5CDR
Bhowmik et al. (2021)	84.8
Angell et al. (2021)	91.3
Varma et al. (2021)	$91.9_{\pm 0.2}$
FT (Ours)	$92.6_{\pm 0.1}$
PT + FT (Ours)	$\mathbf{93.3}_{\pm 0.2}$

Table 1: Recall@1 on BC5CDR test dataset. Sung et al. (2020); Liu et al. (2021a); Lai et al. (2021) evaluate it by splitting into two subsets which is a easier setting and we do not compare to them. FT corresponds to fine-tuning and PT corresponds to pre-training.

Model	NCBI	COMETA	AAP
Sung et al. (2020)	91.1	71.3	82.6
Liu et al. (2021a)	92.3	75.1	89.0
Lai et al. (2021)	92.4	80.1	-
FT (Ours)	$91.6_{\pm 0.1}$	$80.7_{\pm 0.2}$	$88.8_{\pm 0.1}$
PT + FT (Ours)	$\mathbf{91.9}_{\pm 0.2}$	$\mathbf{81.4}_{\pm 0.1}$	$\mathbf{89.3}_{\pm 0.1}$

Table 2: Recall@1 on NCBI-disease, COMETA and AskAPatient test datasets.

When 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 σ.

3 Experiments

3.1 Datasets and KBs

Pre-training We use a subset of UMLS (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 split 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 as metric for comparison, 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 is 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 methods with ours. Our methods with KB-guided pre-training exceed 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 pre-training. Besides, our method shows competitive results on NCBI compared to SOTA, and we analysis results of NCBI in Appendix D.

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

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.

4 Discussion

Does pre-training help? In all datasets, KB-guided pre-training improves fine-tuning consistently, which is 0.7 on BC5CDR, 0.3 on NCBI, 0.7 on COMETA, and 0.5 on AAP. To better understand KB-guided pre-training, we conduct ablation studies. We compare different pre-trained models without fine-tuning in Table 3. BART fails to disambiguate due to the mismatch of the pre-training task. GENRE has been pre-trained on the large-scale BLINK dataset (Wu et al., 2020), and it obtains a decent ability to disambiguate biomedical mentions. Our pre-trained model shows improvement on BC5CDR and COMETA compared to GENRE with fewer pre-training resources (6GPU days vs. 32GPU days). We then conduct synonyms-aware fine-tuning on different pre-trained models in Table 3. Our pre-trained model outperforms BART (+0.8 on BC5CDR, +0.5 on COMETA) and GENRE (+0.4 on BC5CDR, +0.6 on COMETA) which proves the effectiveness of pre-training.

s	$\mathcal{S}$	Prompt	BC5CDR	COMETA
TF-IDF	All	$\surd$	**93.3**	**81.4**
Shortest	All	$\surd$	87.2	80.6
Sample	All	$\surd$	86.9	80.8
Shortest	Shortest	$\surd$	76.3	77.5
Sample	Sample	$\surd$	72.4	77.8
TF-IDF	All	$\times$	93.0	80.9

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.

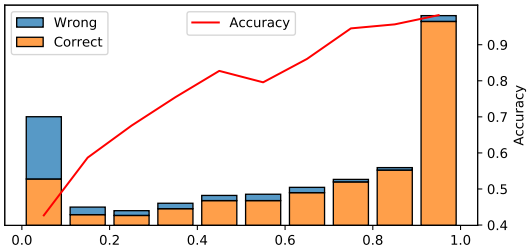

Figure 2: The accuracy of our model on COMETA dataset. The X-axis represents the TF-IDF similarities between mention m and our selected target names s.

Selection of Names We ablate how to select target names s for fine-tuning: (a) proposed TF-IDF similarity; (b) the shortest name in a concept; (c) randomly sampled name in a concept. We also compare how to construct $\mathcal{S}$ for inference: (i) use all synonyms from UMLS and target KB; (ii) use the shortest name for a concept; (iii) use randomly sampled name for a concept. We note using (i) establish an N-to-1 mapping from synonym name to concept, while (ii) and (iii) establish a 1-to-1 mapping. We conduct experiments on available combinations. From Table 4, we conclude that (1) N-to-1 mapping performs better than 1-to-1 mappings, which means synonyms can boost performances. Using one synonym like GENRE degrades performances. (2) Textual similarity criterion performs better than shortest or sampled names when training.

We also check the model accuracy on COMETA trained by using all synonyms as $\mathcal{S}$ and TF-IDF for selecting target names s. From Figure 2, we find the distribution of TF-IDF distance 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 inspires us to select textually similar names for fine-tuning.

Decoder Prompting One difference between GENRE and us is decoder prompt tokens. Prompting has shown improvement on various NLP tasks (Liu et al., 2021b). Here, prompt tokens serve as informative hints by providing additional decoder attention keys 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).

5 Related Work

Biomedical ED 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 samples in ED and augmented data using Wikipedia and PubMed, while our pre-training corpus is solely constructed by KB which can be complementation.

Generative ED Cao et al. (2021a) proposed to view ED 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 (data and GPUs) to achieve competitive performance. Cao et al. (2021b) explored dealing synonyms in multilingual generative ED by adding language identifiers that cannot be directly implemented in biomedical ED.

6 Conclusion

To the best of our knowledge, our work is the first to explore generative ED in the biomedical domain. We inject synonyms and definition knowledge into the model by KG-guided pre-training. We emphasize the synonyms 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 ED benchmarks. GENRE shows that well-selected candidate sets can improve seq2seq ED, and we believe this will further boost our performances.

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(suppl_1):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 Plekhanov, Luke Zettlemoyer, Nicola Cancedda, 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. Ncbi 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, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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. Pre-train, 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. *arXiv preprint arXiv:1907.07972*.

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

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

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models.

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.

Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical entity representations with synonym marginalization. In *ACL*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Shogo Ujiie, Hayate Iso, and Eiji Aramaki. 2021. Biomedical entity linking via contrastive context matching. *arXiv preprint arXiv:2106.07583*.

Maya Varma, Laurel Orr, Sen Wu, Megan Leszczynski, Xiao Ling, and Christopher Ré. 2021. Cross-domain data integration for named entity disambiguation in biomedical text. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4566–4575, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ronald J Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Zero-shot entity linking with dense entity retrieval. In *EMNLP*.

Sheng Yu, Katherine P Liao, Stanley Y Shaw, Vivian S Gainer, Susanne E Churchill, Peter Szolovits, Shawn N Murphy, Isaac S Kohane, and Tianxi Cai. 2015. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. *Journal of the American Medical Informatics Association*, 22(5):993–1000.

Hongyi Yuan and Sheng Yu. 2021. Efficient symptom inquiring and diagnosis via adaptive alignment of reinforcement learning and classification. *arXiv preprint arXiv:2112.00733*.

Zheng Yuan, Zhengyun Zhao, Haixia Sun, Jiao Li, Fei Wang, and Sheng Yu. 2022. Coder: Knowledge-infused cross-lingual medical term embedding for term normalization. *Journal of Biomedical Informatics*, page 103983.

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 `[ST]` and `[ED]`; (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 MEDIC ontology (Davis et al., 2012). MEDIC 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 SNOMED 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 SNOMED-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 5.

B License and Availability of Resources

BC5CDR, NCBI, COMETA, and AskAPatients are all publicly available datasets on the Internet. Their target KBs MeSH, MEDIC, and SNOMED CT 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.

	NCBI	BC5CDR	COMETA	AskAPatient
Target names $\|\mathcal{S}\|$	108,071	809,929	904,798	3,381
Target concepts$\|\mathcal{E}\|$	14,967	268,162	350,830	1,036
Train samples	5784	9,285	13,489	16,826
Dev samples	787	9,515	2,176	1,663
Test samples	960	9,654	4,350	1,712

Table 5: Basic statistics of NCBI-disease, BC5CDR, COMETA, AskAPatient datasets and the corresponding target knowledge bases.

Dataset	NCBI	BC5CDR	COMETA	AAP
FT	$95.6_{\pm 0.1}$	$95.3_{\pm 0.1}$	$88.7_{\pm 0.2}$	$95.6_{\pm 0.1}$
PT+FT	$96.3_{\pm 0.1}$	$95.8_{\pm 0.2}$	$88.2_{\pm 0.1}$	$96.0_{\pm 0.1}$

Table 6: The Recall@5 results of our proposed method on different benchmarks.

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 6.

C.2 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. We split test samples into different sub-populations following Varma et al. (2021) in Table 7.

We list our model's performance on different sub-populations of BC5CDR in Table 8. 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** subset. One possible explanation for the phenomenon is that our generative method is good at learning the linkage between mentions and contextual information, thus gaining ED performance on the zero-shot samples. Our method is influenced more by the textual form of mentions.

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.

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 $\mathcal{S}$, we lowercase them and remove the symbols (e.g., dash line - or comma ,).

Subset	Definition
Overall	Full set of the data
Single Word Mentions	Mentions that have one sole word.
Multi-Word Mentions	Mentions that have multiple words(Separated by blank spaces).
Unseen Mentions	Mentions not existing in fine-tuning set.
Unseen Concepts	Concepts not existing in fine-tuning set.
Not Direct Match	Mentions that are not a synonym of the target concept in KB.
Top 100	Mentions that mapped to the top 100 concepts in existing frequency in fine-tuning set.

Table 7: 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	FT	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 8: 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.

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 9. For those concepts containing only 2 synonyms, s_e^a and s_e^b are the two synonyms respectively and c_e is the same as s_e^b. For those concepts containing only 1 sole synonym, s_e^a, s_e^b and c_e are the same.

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 10. 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.

Concepts		Templates	
Definition	>2 synonyms	Encoder Side	Decoder Side
$\surd$	$\surd$ / $\times$	s_e^a <is defined as> c_e. s_e^a <is described as> c_e. c_e <are the definitions of> s_e^a. c_e <describe> s_e^a. c_e <define> s_e^a.	s_e^a is s_e^b.
$\times$	$\surd$	c_e <are the synonyms of> s_e^a. c_e <indicate the same concept as> s_e^a. s_e^a <has synonyms such as> c_e. s_e^a <refers to the same concepts as> c_e.	s_e^a is s_e^b.
$\times$	$\times$	c_e <is> s_e^a. c_e <is the same as> s_e^a. s_e^a <is> c_e. s_e^a <is the same as> c_e.	s_e^a is s_e^b.

Table 9: The templates used for constructing pre-training samples for different kinds of concepts. s_e^a is the selected synonym as the input mention, s_e^b 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.

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 ϵ	1e-8	1e-8	1e-8	1e-8	1e-8
Adam β	(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

Table 10: Hyper-parameters for KB-guided pre-training and synonyms-aware fine-tuning.