Enhancing Entity Representations with Prompt Learning for Biomedical Entity Linking

Tiantian Zhu[1,2], **Yang Qin**[1], **Qingcai Chen**[1,2,*], **Baotian Hu**[1] and **Yang Xiang**[2,*]
[1]Harbin Institute of Technology (Shenzhen), Shenzhen, China
[2]Peng Cheng Laboratory, Shenzhen, China
zhu.tiantian110@gmail.com, {csyqin, qingcai.chen, hubaotian}@hit.edu.cn, xiangy@pcl.ac.cn

Abstract

Biomedical entity linking aims to map mentions in biomedical text to standardized concepts or entities in a curated knowledge base (KB) such as Unified Medical Language System (UMLS). The latest research tends to solve this problem in a unified framework solely based on surface form matching between mentions and entities. Specifically, these methods focus on addressing the *variety* challenge of the heterogeneous naming of biomedical concepts. Yet, the *ambiguity* challenge that the same word under different contexts may refer to distinct entities is usually ignored. To address this challenge, we propose a two-stage linking algorithm to enhance the entity representations based on prompt learning. The first stage includes a coarser-grained retrieval from a representation space defined by a bi-encoder that independently embeds the mention and entity's surface forms. Unlike previous one-model-fits-all systems, each candidate is then re-ranked with a finer-grained encoder based on prompt-tuning that utilizes the contextual information. Extensive experiments show that our model achieves promising performance improvements compared with several state-of-the-art techniques on the largest biomedical public dataset MedMentions and the NCBI disease corpus. We also observe by cases that the proposed prompt-tuning strategy is effective in solving both the *variety* and *ambiguity* challenges in the linking task.

1 Introduction

With the advancements in the healthcare domain, the volume of biomedical text, such as electronic health records (EHRs) or biomedical literature, is explosively growing daily, creating an urgent need to utilize the useful information contained in these records. Entity linking is a necessary step that builds on the output mentions of named entity recognition (NER), which refers to the process of automatically linking mentions in plain text, to a standardized list of entities in a knowledge base (KB). Although there are numerous entity linking approaches in the general domain, they cannot be directly transferred to biomedical domain due to the huge differences in language characteristics and knowledge bases. Thus, it is critical to have accurate biomedical entity linking techniques, which would better assist the understanding of biomedical text. Furthermore, biomedical entity linking can be useful in many downstream applications such as medical-related decision making, predictive modeling, medical information retrieval, information extraction, and question answering.

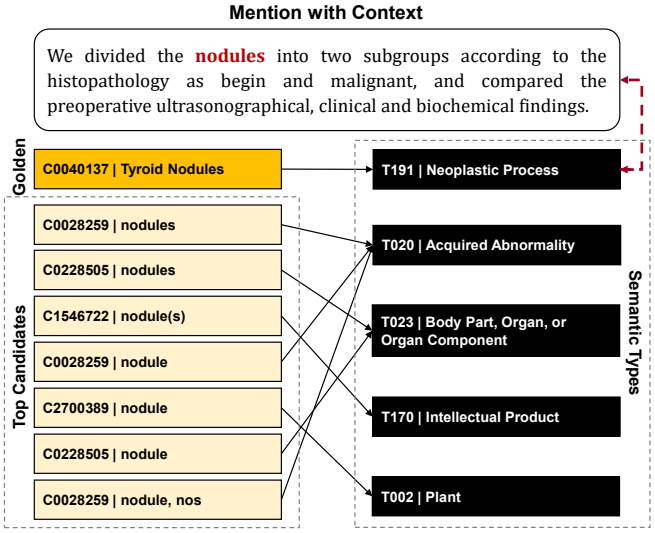

Figure 1: An example of biomedical entity linking, where the mention is in red. The orange box denotes the golden entity, which is also retrieved as the top 1 candidate by our finer-grained encoder. The yellow boxes refer to the top candidates retrieved by SAPBERT and the black boxes refer to the semantic types of these entities.

Generally, there are two main problems, namely *variety* and *ambiguity*, which challenge the accuracy of biomedical entity linking systems. The variety problem refers to a named entity may appear in different surface forms in a given text. The ambiguity problem refers to two named entities with the same surface form may have different semantic meanings. The latest research tends to address the biomedical entity linking task in a unified framework that considers only the biomedical entities themselves but without

*Corresponding authors.

their contexts, ignoring the ambiguity issue[Liu *et al.*, 2021; Sung *et al.*, 2020]. However, the ambiguous instances account for a high proportion of the overall entities and mentions, and failure in disambiguation may lead to misinterpretation of the entire context. For example, there are $79,609$ synonyms in Unified Medical Language System (UMLS) share the same surface but have distinct Concept Unique IDs (CUIs), which relies on extra information to distinguish the correct entities. Moreover, the mentions can be ambiguous and it is difficult to associate a mention to the correct entity based on surface-level features alone, which requires a model to have a good semantic understanding of the mention and its context.

Consider the example in Figure 1, where *nodules* is a mention of the entity *Thyroid Nodules*, and others are the top candidates retrieved from UMLS by SAPBERT [Liu *et al.*, 2021]. In this case, although the top candidates that ranked by SAPBERT are similar to the mention in surface, they are not the golden entities. However, if the linking model has the context of mentions and the type information of entities, the golden concept may have a greater chance of being recalled, which will make the entity disambiguation more precise.

Motivated by the above observations, in this paper, we propose a two-stage biomedical entity linking algorithm to enhance the entity representations based on prompt learning. To address the variety challenge and comprehensively consider the inference speed, we utilize a coarser-grained retrieval from a representation space defined by a bi-encoder that independently embeds the mention and entity's surface forms to retrieve candidates in the first stage. Afterwards, each candidate is re-ranked with a finer-grained encoder based on prompt-tuning that utilizes the mention context and entity type information, mainly aiming to solve ambiguity and improve linking accuracy. In detail, prompt-tuning is a new paradigm that bridges the gap of objective forms between pre-training and fine-tuning. And it can further inject and stimulate the knowledge in biomedical pre-trained language models (PLMs), thus boosts the model performance. Finally, we evaluate the performance of our model using the MedMentions [Mohan and Li, 2019] and the NCBI disease [Dogan *et al.*, 2014] datasets, and against several state-of-the-art (SOTA) baselines. From the extensive experimental results, the proposed prompt-tuning strategy is demonstrated to be effective in recalling the correct entities that missed by the coarser-grained retrieval and enhancing the representations of the bio-entities. For example, in Figure 1, the entity *Thyroid Nodules* can be recalled by the proposed finer-grained encoder but missed by SAPBERT. The code is available at https://github.com/TiantianZhu110/BioPRO.

To sum up, the contributions of this paper are as follows: (1) We propose a two-stage algorithm to address both the variety and the ambiguity challenges of the entity linking task; (2) For the first time, we propose a finer-grained encoder based on prompt-tuning to re-rank the candidates output by the first stage; (3) Extensive experiments show that our model achieves promising performance improvements on two public datasets, which also demonstrates the effectiveness of prompt-tuning strategy in recalling golden entities that even failed in the first stage.

2 Related Work

2.1 Entity Linking

Entity linking focuses on mapping an input mention from biomedical text to its associated entity in a curated KB. In the biomedical domain, UMLS is often used as the KB for entities, as it is a rich conceptual ontology differentiated by CUIs and relations between them. There are many existing entity linking tools, such as TaggerOne [Leaman and Lu, 2016], QuickUMLS [Luca and Goharian, 2016], which rely on rule-based approaches. However, such approaches struggle when there is a large discrepancy between the morphologies of mentions and concepts. Recently, the field has shifted toward machine learning methods, which can be divided into classification [Niu *et al.*, 2019] and learning to rank [Sung *et al.*, 2020; Liu *et al.*, 2021; Zhu *et al.*, 2020] categories. Among them, classification approaches have the disadvantage that the output space must be the same as the number of concepts. However, ranking methods rank the similarities between input mention and candidate entities, which would reduce the output space. BIOSYN [Sung *et al.*, 2020] encodes input and target terms by both TF-IDF and BioBERT [Lee *et al.*, 2020] and uses synonym marginalization to maximize similarities between synonyms. SAPBERT [Liu *et al.*, 2021] proposes self-alignment pretraining to fine-tune BERT on synonyms extracted from the UMLS and currently holds SOTA across all major English biomedical entity linking datasets. LATTE [Zhu *et al.*, 2020] uses an attention-based mechanism to rank candidates for a given mention based on their semantic representations along with the latent-types.

In our framework, we propose a two-stage algorithm by first generating candidates and then ranking them, which takes the context of mentions as well as the type information of entities into consideration to obtain better representations.

2.2 Prompt Learning

Recently, PLMs provide a new trend to utilize massive unlabeled corpora and have been demonstrated their effectiveness on various natural language processing (NLP) tasks [Devlin *et al.*, 2019; Raffel *et al.*, 2020]. However, a fine-tuning process is still needed for applying the rich lexical, syntactic, and factual knowledge of the PLMs to different downstream NLP tasks. Specifically, pre-training is usually formalized as a cloze-style task, yet conventional fine-tuning often adds extra classifiers on the top of PLMs. And by analyzing the objective forms of the pre-training and the fine-tuning processes, we can find a significant gap between them, which would restrict taking full advantage of knowledge in PLMs.

Since the emergence of GPT-3 [Brown *et al.*, 2020], prompt-tuning has attracted widespread attention from the NLP community. In contrast with fine-tuning, prompt-tuning is a new paradigm that leverages language prompts as contexts to adapt different tasks, which is also termed in-context learning and makes downstream tasks expressed as some objectives similar to pre-training objectives. Prompting means adding instructions and demos before input and output predictions to stimulate knowledge from PLMs. Recently, handcrafted prompts have achieved promising results in knowledge probing [Davison *et al.*, 2019], natural language infer-

ence [Schick and Schütze, 2021], and so on. Furthermore, to avoid handcrafting prompts for different tasks, a series of work focus on automatic prompt search [Jiang *et al.*, 2020; Gao *et al.*, 2021]. Moreover, unlike these discrete prompts, continuous prompts have also been used to steer PLMs [Li and Liang, 2021].

In this paper, we aim to steer PLMs with prompt-learning to capture the contextual information of both mentions and entities. We take biomedical entity linking, a crucial task in the biomedical domain, as the foothold to design the prompt-learning strategy. In our work, we mainly emphasize that using prompt-learning to perform linking task can not only improve experimental performance, but also learn deeper biomedical entity representations.

3 Method

3.1 Problem Statement

Given a biomedical document of l tokens $D = \{w_1, w_2, \ldots, w_l\}$ with a list of n mentions $M = \{m_1, m_2, \ldots, m_n\}$, and a KB of g entities $K = \{e_1, e_2, \ldots, e_g\}$, the entity linking system will output the target entity e_t in the KB for each mention m. Note that each mention has its own golden CUI and each CUI has at least one entity synonym in the KB. We assume that the type information of the entities is available, which is a common setting in biomedical KB such as UMLS.

3.2 Model Architecture

Following previous work [Wu *et al.*, 2020], we adopt a two-stage approach to perform biomedical entity linking and the overall architecture is illustrated in Figure 2. In the first stage, a coarser-grained retrieval model based on the bi-encoder is used to produce the mention and entity's vector representations and further generate the top candidates. The retrieved candidates along with contextual information are then passed to the finer-grained ranking model in the second stage for re-ranking. In the following sections, we describe in detail the coarser-grained retrieval model and the finer-grained ranking model used in this work.

3.3 Coarser-grained Retrieval Model

Considering the inference speed, we utilize a bi-encoder architecture similar to the work of [Wu *et al.*, 2020] to independently embed the mention and entity's surface forms in the first stage. We use pre-trained SAPBERT to encode representations and share the same SAPBERT parameters for both mentions and entities. Specifically, the surface representation of the mention m is calculated as follows:

$$r_m^s = \texttt{SAPBERT}(m)\texttt{[CLS]} \tag{1}$$

where $r_m^s \in \mathbb{R}^{d_h}$ and d_h denotes the hidden dimension of SAPBERT. `[CLS]` denotes the special token that BERT-style models use to compute a single vector representation of the input. The entity surface representation r_e^s is computed similarly. Then, the score of entity candidate e_i is given by the scoring function:

$$S(m, e_i) = f(r_m^s, r_{e_i}^s) \tag{2}$$

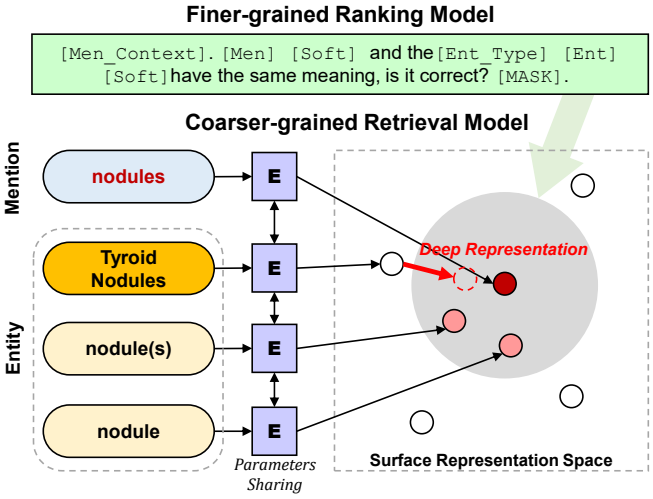

Figure 2: The overall architecture of the proposed model.

where we use cosine similarity for f. Note that for each mention, we compute $S(m, e_i)$ for all entities $e_i \in K$ and return the top k nearest candidates. In addition, the entity representations can be cached at the inference time to reduce computational cost.

3.4 Finer-grained Ranking Model

The input of the finer-grained ranking model is the mention with context and a set of top k candidate entities $C = \{c_1, c_2, \ldots, c_k\}$ with type information, the model computes a relevance score for each entity in C. In this stage, we apply a prompt learning method based on PubMedBERT to perform the task, which transforms the original task into a masked language modeling problem. By incorporating the mention context and entity type information into the prompt template, the proposed finer-grained ranking model can reduce the semantic drift problem and achieve better performance.

Generally, a prompt consists of a *template* and a *verbalizer*. Among them, the template $T(\cdot)$ wraps the input sequence x into $T(x)$ and there is at least one `[MASK]` in $T(x)$ for the language model $\mathcal{M}$ to fill the label words. Moreover, for each label $y \in \mathcal{Y}$, we define a corresponding label word set $\mathcal{V}_y$, which is a subset of the vocabulary $\mathcal{V}$ of the $\mathcal{M}$. And the verbalizer will map the label word in $\mathcal{V}_y$ to the label in $\mathcal{Y}$. Thus, in prompt learning, the task can be transformed into a masked language modeling problem:

$$p(y \in \mathcal{Y}|x) = p(\texttt{[MASK]} = w \in \mathcal{V}_y|T(x)). \tag{3}$$

Templates

In this work, we construct two mention-entity prompts for the entity linking task. Specifically, one is a hard-encoding template and the other is a mixed template with both hard- and soft-encoding.

In hard-encoding template, we fuse the labels into answers via question answering. Given the mention with context and the entity with semantic type, the template is constructed as: T_{hard} =`[Men_Context]. [Men]` and the `[Ent_Type] [Ent]` have the same meaning, is it correct? `[MASK].`

Among them, [Men] and [Ent] denote the mention and entity names, respectively. [Men_Context] denotes the context of the mention and here we take the shortest sentence containing the mention. [Ent_Type] is the semantic type of the entity and we use *"unknown type"* for unknown entity types. The predicted words in the position [MASK] will be mapped to the final label. Here, we set *"yes"* as the positive label word and *"no"* as the negative label word. Thus, the probability of the label word w can fill the masked position, which is calculated as follows:

$$p(\texttt{[MASK]} = w \in \mathcal{V}_y | T(x)) = \frac{\exp\left(\boldsymbol{v}_w \cdot \boldsymbol{h}_{\texttt{[MASK]}}\right)}{\sum_{v_i \in \mathcal{V}} \exp\left(\boldsymbol{v}_i \cdot \boldsymbol{h}_{\texttt{[MASK]}}\right)} \quad (4)$$

where $\boldsymbol{v}_w$ is the embedding of the token w in the PLM $\mathcal{M}$ and $\boldsymbol{h}_{\texttt{[MASK]}}$ is the hidden vector of [MASK].

In order to prevent the long-distance forgetting problem, we also add the trainable soft tokens to enhance the influence of the mention and the entity on the [MASK] position in the mixed template. The mixed template is constructed as follows: T_{mixed} =[Men_Context]. [Men] [Soft] and the [Ent_Type] [Ent] [Soft] have the same meaning, is it correct? [MASK]. Among them, [Soft] denotes the soft token, which is randomly initialized as a word in the PLM dictionary. And the soft token embeddings are free parameters which are initialized as a trainable matrix P_θ (parametrized by θ). Thus, the input embeddings of PLM can be calculated as follows:

$$\text{TokenEmbedding}(i) = \begin{cases} P_\theta[i], & \text{if } i \in \text{Soft}_{idx} \\ \text{PLM}(token_i), & \text{otherwise} \end{cases} \quad (5)$$

where Soft_{idx} denotes soft token indices and $P_\theta \in \mathbb{R}^{|\text{Soft}_{idx}| \times d_{\tilde{h}}}$. $|\text{Soft}_{idx}|$ denotes the length of the soft tokens and $d_{\tilde{h}}$ denotes the hidden dimension of PLM.

Training

For training the ranking model, golden entities are selected as the positive samples, which may have multiple semantic types. We construct multiple positive samples by looping on different types of entities. Negative samples come from unmatched candidates generated by the SAPBERT model as well as negative instances randomly sampled from the full dataset. The loss function of the finer-grained ranking model is defined as follows:

$$\begin{aligned} L = \sum_{i=1}^{N} \omega [& y_i \cdot \log\left(p(\texttt{[MASK]} = w^+ | T(x_i))\right) \\ & + (1 - y_i) \cdot \log\left(p(\texttt{[MASK]} = w^- | T(x_i))\right)] \end{aligned} \quad (6)$$

where ω is the label weight to balance the positive and the negative instances, w^+ denotes the label word *"yes"* and w^- denotes *"no"*.

Prompt for Candidates Retrieval

Since the vector representations produced by the coarser-grained model does not consider the extra information of mentions and entities, we use the trained finer-grained ranking model to generate deep representations. Furthermore, the candidates is retrieved by scoring both surface and deep representations in the first stage of our experiments. When generating deep representations, in order to maintain consistency between training and inference, we utilize the similar data format as prompt-tuning as the input. Specifically, when generating the entity deep representations, we replace the [Men_Context] and [Men] with [Ent_Type] and [Ent] in the mixed template, respectively. Then we extract the average representations of the corresponding entity tokens before the second soft token as the entity deep representations. Similarly, we can replace the [Ent_Type] and [Ent] with *"unknown type"* and [Men] in the template and get the mention deep representations in the same way.

4 Experiments

4.1 Datasets

We used two datasets to evaluate our proposed model: MedMentions [Mohan and Li, 2019] and NCBI Disease [Dogan *et al.*, 2014] datasets. MedMentions is the largest biomedical entity linking dataset consisting of $4,392$ abstracts from PubMed, with over $350,000$ mentions linked to the 2017AA full version of UMLS concepts. The size of the concept set in UMLS is $3,415,665$, and the size of all surface forms / synonyms set is $14,815,318$. And the concepts in UMLS belongs to 127 semantic types. NCBI Disease corpus contains 793 manually annotated PubMed abstracts and $6,881$ mentions with each CUI mapped to the July 6, 2012 version of MEDIC dictionary [Davis *et al.*, 2012]. The size of the concept set in MEDIC dictionary is $11,915$, and the size of all surface forms / synonyms set is $71,923$. Descriptions of the datasets and their statistics are provided in Table 1.

4.2 Evaluation Metrics

Since picking the correct target entity among candidates is a ranking problem, we use the top *k* accuracy as an evaluation metric following the previous work [Sung *et al.*, 2020; Liu *et al.*, 2021]. For all datasets, we report Acc@1 and Acc@5 for evaluating performance of the proposed model against the baseline models, where Acc@*k* represents the accuracy when only *k* entities are retrieved.

4.3 Implementation Details

Table 2 lists the hyper-parameter search space for obtaining the set of used numbers. Note that all parameters are selected based on the best performance on the development set.

Dataset	Num. of	Train	Dev	Test
MedMentions	Documents	2,635	878	879
	Mentions	211,029	71,062	70,405
	Entities	25,640	12,586	12,402
NCBI Disease	Documents	592	100	100
	Mentions	5,134	787	960
	Entities	668	176	203

Table 1: Statistics of different sets in MedMentions and NCBI Disease datasets.

Hyper-parameters	Search Space
Optimizer	$AdamW$
Learning Rate	$1e-4, 1e-5, 2e-6, 1e-6*$
Batch Size	$128, 64, 32*$
Label Weight	$[0.5, 0.5], [0.1, 0.9], [0.4, 0.6]*$
Epoch	15
Max Sequence Length	256
Soft Token Embedding Size	768
Weight Decay	0.01

Table 2: The search space for hyper-parameters used in prompt-tuning. "$*$" denotes the used ones for reporting results.

Model	#Error	#Ambiguous	Proportion
SAPBERT	$33,037$	$11,636$	35.2%
Ours	$23,560$	$2,037$	$8.6\%^{\downarrow 26.6\%}$

Table 3: Statistics of the ambiguous instances in mispredictions on MedMentions test set. Note that the "#Error" refers to the number of top 1 candidates predicted incorrectly. "#Ambiguous" refers to the number of ambiguous instances in top 1 candidates predicted incorrectly.

4.4 Baselines

For the evaluation of the proposed model, we use the following SOTA baseline methods for comparison.

- **Sieve-based** [D'Souza and Ng, 2015] is a multi-pass sieve approach to the under-studied task of normalizing disorder mentions in the biomedical domain.
- **TaggerOne** [Leaman and Lu, 2016] is a machine learning-based system that jointly performs disease NER and normalization by utilizing semi-Markov models.
- **NormCo** [Wright *et al.*, 2019] is a deep coherence model which considers the semantics of an entity mention, as well as the topical coherence of the mentions within a single document.
- **BNE** [Phan *et al.*, 2019] proposes a new framework for learning robust representations of biomedical terms by learning to encode names of the same concepts into similar representations, while preserving their conceptual and contextual meanings.
- **BERTRANK** [Ji *et al.*, 2020] applies and evaluates the pre-trained BERT / BioBERT / ClinicalBERT models for candidate concept ranking by transforming the ranking task as a sentence-pair classification task, which is a point-wise learning to rank method.
- **BIOSYN** [Sung *et al.*, 2020] utilizes the synonym marginalization technique and the iterative candidate retrieval for learning biomedical entity representations.
- **SAPBERT** [Liu *et al.*, 2021] is a self-alignment pre-training schema for learning biomedical entity representations.

Model	NCBI Disease		MedMentions	
	@1	@5	@1	@5
Sieve-based [D'Souza and Ng, 2015]	84.7	-	-	-
TaggerOne [Leaman and Lu, 2016]	87.7	-	OOM	OOM
NormCo [Wright *et al.*, 2019]	87.8	-	-	-
BNE [Phan *et al.*, 2019]	87.7	-	-	-
BERTRANK [Ji *et al.*, 2020]	89.1	-	-	-
BIOSYN [Sung *et al.*, 2020]	91.1	93.9	OOM	OOM
SAPBERT [Liu *et al.*, 2021]	$92.2^{\dagger}$	$95.9^{\dagger}$	$53.1^{\dagger}$	$73.5^{\dagger}$
Ours(w/o deep representations,hard)	92.6	**95.9**	65.9	74.6
Ours(w/o deep representations,mixed)	94.4	95.5	66.3	74.7
Ours	**94.5**	95.6	**66.5**	**74.7**

Table 4: Performance on the NCBI Disease and the MedMentions datasets in comparison with the SOTA methods. "w/o deep representations" means that only the SAPBERT representations are used in the first stage. Bold denotes the best result in the column. "†" denotes results produced using official released code. "-" denotes results not reported in the cited paper. "OOM" means out-of-memory.

4.5 Experimental Results

Main Results

To evaluate our approach, we compared the proposed model with several SOTA models on the two datasets, and listed the performance of each model in Table 4. As shown in Table 4, our model outperformed all other models with a large margin, which improved 2.3 and 13.4 Acc@1 points over the baselines on NCBI Disease and MedMentions test sets, respectively. The improvement of Acc@1 is significantly greater than that of Acc@5, which demonstrates the effectiveness of prompt-tuning based strategy in improving the precision of the entity linking task. In addition, we list the performance comparison of the proposed model and its variants. As shown in Table 4, mixed template is better than hard template in prompt-tuning and the model that integrates the deep representations has the best performance. Note that [Mohan and Li, 2019] first reported that training TaggerOne on a subset of MedMentions requires $>$ 900 GB of RAM due to the large search space. In comparison, our method can avoid the out-of-memory problem on the full-set of MedMentions.

Resolution of the Ambiguity Problem

In order to study the resolution of different models on the ambiguity problem, we compared the top 1 candidates prediction's performance of our proposed model with SAPBERT on the MedMentions dataset. As mentioned before, the ambiguous instances account for a high proportion of the overall entities and mentions, increasing the challenge of the task. Table 3 lists the number of wrongly predicted top 1 candidates, the number of ambiguous examples in wrongly predicted top 1 candidates and the proportion of ambiguous instances in the overall top 1 errors. It is shown that our model can greatly reduce the proportion of ambiguous examples ($\downarrow$ 26.6%) among the total misclassified examples, which demonstrates the effectiveness of the proposed prompt-tuning strategy in addressing the ambiguity challenge and understanding deep representations of mentions and entities.

Mention with Context	SAPBERT	Prompt
Lumbar intrathecal fentanyl was used to attenuate the central projection of μ-opioid receptor-sensitive locomotor muscle **afferents** during a 5 km cycling time trial.	afferent *Spatial Concept*	**Afferent Neurons** *Cell*
These **pathways** activate one another mutually leading to oxidative stress, increasing expression of transforming growth factor beta-1 (TGF-β 1) and release of interleukins and adhesion molecules.	pathways *Conceptual Entity*	**pathways signaling** *Molecular Function*
Thereafter, the rats were divided into DN group, DN group receiving Telmisartan or Sildenafil or Telmisartan Sildenafil **combination**	combination *Physical Object*	**drug combination** *Pharmacologic Substance*
Therefore, it is important to prevent this condition by **identifying** women at risk, allowing the clinician to implement preventive strategies, including the use of GnRH antagonist cycles with agonist trigger.	identification *Qualitative Concept*	**Identifying patient** *Health Care Activity*
In broad causes of death classification, **injuries** have been found to be the second most cause of death next to communicable diseases.	injuries Qualitative Concept	**WOUNDS INJURIES** Injury or Poisoning
We ascribed the effects of aspirin to AMP-activated protein kinase (AMPK) activation, **mTORC1** inhibition, and autophagy induction.	mtorc1 *Gene or Genome*	**mTORC1** *Amino Acid, Peptide, or Protein*
Garlic is an allelopathic crop that can alleviate the obstacles to continuous cropping of **vegetable crops.**	vegetable (substance) *Food*	**Crops** *Plant*
Cell growth was assessed by **counting** and MTS assay.	counting number *Quantitative Concept*	**cell counting** *Laboratory Procedure*

Table 5: Examples of top 1 candidates retrieved by SAPBERT and the prompt-tuning based finer-grained encoder. Bold denotes mentions. Underline denotes the predicted top 1 candidates and candidates in bold refers to the golden entities. Italic denotes the semantic types.

Method	MedMentions	NCBI Disease
SAPBERT	77.3	96.6
SAPBERT+Prompt Representations	**81.3**	**97.6**

Table 6: Recall of different candidates retrieval strategies on the MedMentions and NCBI Disaese datasets. "+Prompt Representations" denotes incorporating the representations of the finer-grained encoder.

Representations	Euclidean Distance	Cosine Similarity
SAPBERT vs. PubMedBERT	1.617	98.9
Prompt vs. PubMedBERT	1.062	99.5

Table 7: Differences in vector space representation between tuned models and PubMedBERT.

Performance of the Coarser-grained Model

Although we measure the overall performance of our two-stage entity linking algorithm in Table 4, the ranking results of the second step are based on the recall candidates output by the first step. Since the recall of the coarser-grained model will limit the performance of the finer-grained encoder, we further report the recall value of different candidate retrieval methods, namely SAPBERT and SAPBERT fusion prompt representations, in Table 6. As can be seen from Table 6, after incorporating the prompt representations in the first step, the recall rate has improved greatly. This may be due to the fact that the SAPBERT model tends to recall entities with similar surface forms, and misses instances of dissimilar forms but semantically matching. In comparison, prompt-tuning based representations preserve the contextual information well and are complementary to the SAPBERT representations, which can promote the recall improvement of the coarser-grained model.

Representations Analysis

From Table 6, we observe that the prompt-tuning based representations can improve the recall value of the coarser-grained model, and assume that the prompt representations may be complementary to the SAPBERT representations. To further study the differences between the prompt representations and the SAPBERT representations, we reported their Euclidean distance and cosine similarity to PubMedBERT's vector space representations in Table 7, respectively. Both SAPBERT and our finer-grained encoder utilize PubMedBERT as a base model for further training, yet their tuning strategies differ. As shown in Table 7, our prompt-tuning based finer-grained encoder is more similar to PubMedBERT in terms of vector space representation compared to fine-tuning based SAPBERT, which may better preserve the original biomedical knowledge of pre-trained language model and is suitable for modeling deeper contextual information.

Case Study

To illustrate the effectiveness of the prompt-tuning based finer-grained encoder, we listed the top 1 candidates retrieved by SAPBERT and our prompt encoder in Table 5, respectively. We can infer from Table 5 that SAPBERT prefers entities with similar surface form as the mentions, but may miss real matching entities. However, it clearly indicates that the proposed prompt encoder can make up for this disadvantage and make comprehensive judgments by leveraging the contextual information of entities and mentions, which will further improve the accuracy of the coarser-grained model.

5 Conclusion

In this paper, we focus on both the variety and the ambiguity challenges in biomedical entity linking. We propose a two-stage algorithm, including a coarser-grained retrieval and a finer-grained encoder based on prompt-tuning. Empirical results demonstrate that the prompt-tuning strategy of our model could considerably improve the recall of the coarser-grained model and meanwhile gain high performance.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 62006061, 61872113, 62106115), Stable Support Program for Higher Education Institutions of Shenzhen (No.GXWD20201230155427003-20200824155011001), Strategic Emerging Industry Development Special Funds of Shenzhen (No. XMHT20190108009), and Fundamental Research Fund of Shenzhen (No. JCYJ20190806112210067).

References

[Brown *et al.*, 2020] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proc. of NeurIPS*, pages 1877–1901, 2020.

[Davis *et al.*, 2012] Allan Peter Davis, Thomas C. Wiegers, Michael C. Rosenstein, and Carolyn J. Mattingly. MEDIC: a practical disease vocabulary used at the comparative toxicogenomics database. *Database J. Biol. Databases Curation*, 2012, 2012.

[Davison *et al.*, 2019] Joe Davison, Joshua Feldman, and Alexander M. Rush. Commonsense knowledge mining from pretrained models. In *Proc. of EMNLP-IJCNLP*, pages 1173–1178, 2019.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL-HLT*, pages 4171–4186, 2019.

[Dogan *et al.*, 2014] Rezarta Islamaj Dogan, Robert Leaman, and Zhiyong Lu. NCBI disease corpus: A resource for disease name recognition and concept normalization. *J. Biomed. Informatics*, 47:1–10, 2014.

[D'Souza and Ng, 2015] Jennifer D'Souza and Vincent Ng. Sieve-based entity linking for the biomedical domain. In *Proc. of ACL*, pages 297–302, 2015.

[Gao *et al.*, 2021] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In *Proc. of ACL/IJCNLP*, pages 3816–3830, 2021.

[Ji *et al.*, 2020] Zongcheng Ji, Qiang Wei, and Hua Xu. Bert-based ranking for biomedical entity normalization. *AMIA Jt Summits Transl Sci Proc.*, 2020:269–277, 2020.

[Jiang *et al.*, 2020] Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. How can we know what language models know. *Trans. Assoc. Comput. Linguistics*, 8:423–438, 2020.

[Leaman and Lu, 2016] Robert Leaman and Zhiyong Lu. Taggerone: joint named entity recognition and normalization with semi-markov models. *Bioinform.*, 32(18):2839–2846, 2016.

[Lee *et al.*, 2020] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.*, 36(4):1234–1240, 2020.

[Li and Liang, 2021] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proc. of ACL/IJCNLP*, pages 4582–4597, 2021.

[Liu *et al.*, 2021] Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. Self-alignment pretraining for biomedical entity representations. In *Proc. of NAACL*, pages 4228–4238, 2021.

[Luca and Goharian, 2016] Soldaini Luca and Nazli Goharian. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, sigir*, pages 1–4, 2016.

[Mohan and Li, 2019] Sunil Mohan and Donghui Li. Medmentions: A large biomedical corpus annotated with UMLS concepts. In *Proc. of AKBC*, 2019.

[Niu *et al.*, 2019] Jinghao Niu, Yehui Yang, Siheng Zhang, Zhengya Sun, and Wensheng Zhang. Multi-task character-level attentional networks for medical concept normalization. *Neural Process. Lett.*, 49(3):1239–1256, 2019.

[Phan *et al.*, 2019] Minh C. Phan, Aixin Sun, and Yi Tay. Robust representation learning of biomedical names. In *Proc. of ACL*, pages 3275–3285, 2019.

[Raffel *et al.*, 2020] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020.

[Schick and Schütze, 2021] Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proc. of EACL*, pages 255–269, 2021.

[Sung *et al.*, 2020] Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. Biomedical entity representations with synonym marginalization. In *Proc. of ACL*, pages 3641–3650, 2020.

[Wright *et al.*, 2019] Dustin Wright, Yannis Katsis, Raghav Mehta, and Chun-Nan Hsu. Normco: Deep disease normalization for biomedical knowledge base construction. In *Proc. of AKBC*, 2019.

[Wu *et al.*, 2020] Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. Scalable zero-shot entity linking with dense entity retrieval. In *Proc. of EMNLP*, pages 6397–6407, 2020.

[Zhu *et al.*, 2020] Ming Zhu, Busra Celikkaya, Parminder Bhatia, and Chandan K. Reddy. LATTE: latent type modeling for biomedical entity linking. In *Proc. of AAAI*, pages 9757–9764, 2020.