Improving Biomedical Entity Linking with Cross-Entity Interaction

Zhenran Xu, Yulin Chen, Baotian Hu[*]

Harbin Institute of Technology (Shenzhen), Shenzhen, China
xuzhenran@stu.hit.edu.cn, 200110528@stu.hit.edu.cn, hubaotian@hit.edu.cn

Abstract

Biomedical entity linking (EL) is the task of linking mentions in a biomedical document to corresponding entities in a knowledge base (KB). The challenge in biomedical EL lies in leveraging mention context to select the most appropriate entity among possible candidates. Although some EL models achieve competitive results by retrieving candidate entities and then exploiting context to re-rank them, these re-ranking models concatenate mention context with one candidate at a time. They lack fine-grained interaction among candidates, and potentially cannot handle ambiguous mentions when facing candidates both with high lexical similarity. We cope with this issue using a re-ranking model based on prompt tuning, which represents mention context and all candidates at once, letting candidates in comparison attend to each other. We also propose a KB-enhanced self-supervised pretraining strategy. Instead of large-scale pretraining on biomedical EL data in previous work, we use masked language modeling with synonyms from KB. Our method achieves state-of-the-art results on 3 biomedical EL datasets: NCBI disease, BC5CDR and COMETA, showing the effectiveness of cross-entity interaction and KB-enhanced pretraining strategy. Code is available at https://github.com/HITsz-TMG/Prompt-BioEL.

1 Introduction

Biomedical entity linking (EL) refers to linking mentions in a biomedical document to corresponding entities in a curated knowledge base (KB) such as UMLS (Bodenreider 2004) and SNOMED-CT (Donnelly et al. 2006). Considering the example in Figure 1, given the sentence "After a few days of **feeling emotions**, I will get extreme anxiety.", the mention **feeling emotions** should be linked to the entity 408453002 - *Emotional* in SNOMED-CT. Biomedical EL, as a bridge that connects mentions in unstructured text and entities in structured KBs, serves as a fundamental component for many downstream tasks, such as biomedical question answering (Lee et al. 2020), information extraction (Huang, Yang, and Peng 2020) and automatic diagnosis (Yuan and Yu 2021). Although EL systems have achieved great success in general domain, they cannot be directly implemented to solve biomedical EL due to the data scarcity (Yuan, Yuan, and Yu 2022) and KB format (Chen, Varoquaux, and Suchanek 2021), so it is critical to propose EL approaches particularly for biomedical domain.

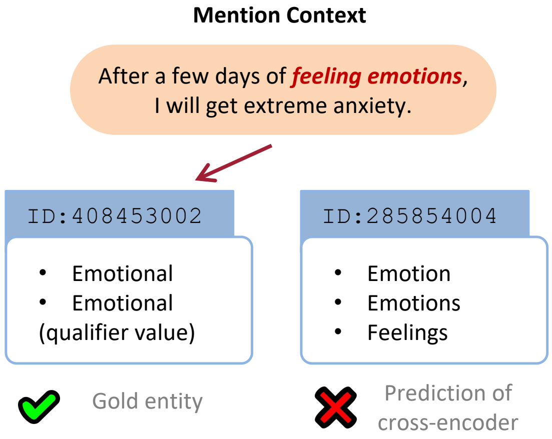

Figure 1: An error case of the cross-encoder proposed in Varma et al. (2021). The cross-encoder tends to confuse candidates both with high lexical similarity, indicating the need for cross attention among candidates in comparison.

There are two main challenges of EL: (1) *variety* - the same entity can appear as different words or phrases in different texts; (2) *ambiguity* - the same word or phrase can refer to different entities. Previous work mainly addresses the *variety* challenge. There is a common bi-encoder design for most of previous models (Sung et al. 2020; Liu et al. 2021a): the bi-encoder represents mention words and entity names independently and calculates similarity through a dot product between dense vector encodings. Although these models achieve promising performance gains, they ignore the mention context and cannot solve the *ambiguity* challenge, i.e., they link the same mention to the same entity no matter what mention context is.

To overcome the *ambiguity* bottleneck, most recent methods (Angell et al. 2021; Varma et al. 2021) present a two-stage linking algorithm: they first retrieve a small set of candidate entities with the bi-encoder above, and then re-rank the candidates according to mention context with a cross-encoder. Although the cross-encoder can capture mention-entity interactions, each entity gets encoded with mention context independently from all the other candidates, ignor-

[*]Corresponding author.

ing fine-grained entity-entity interactions. None of the previous methods attends to all candidates at once, potentially reducing the performance on ambiguous mentions when facing candidates both with high lexical similarity.

Figure 1 shows an error case of the cross-encoder. When re-ranking candidates which both have high lexical similarity with the mention, the cross-encoder tends to give both candidates high scores, even ranking wrong entities higher than gold entities (e.g., in the error case of Figure 1, based on our implementation of Varma et al. (2021), the probability for the gold entity is 0.16, but the probability for the wrong candidate is 0.84), indicating the need for entity-entity interaction among candidates in comparison. More examples will be discussed in Section 4.5. In addition, it has shown significant potential in other semantic tasks to let the model see all of its possible candidates (Barba, Pasini, and Navigli 2021). Motivated by the above example and the conclusion from other tasks, we propose a re-ranking model which attends to mention context and all candidates jointly, explicitly modeling both mention-entity interaction and entity-entity interaction. Since prompt learning can reduce the gap between pretraining objectives of language models and downstream tasks, here we use a *cloze prompt* (Liu et al. 2021b), rooted from the widely-used masked language modeling (MLM), to make choices among candidates.

As a knowledge-intensive task, the progress of biomedical EL is hindered by incomplete structural resources of KB and the scarcity of annotated training data (Varma et al. 2021; Yuan, Yuan, and Yu 2022). Previous researches integrate data from other domains (e.g., Wikipedia) or generate data by filling manual templates with entity descriptions. With more EL data, they perform large-scale pretraining and boost the results of their models. However, the above approaches are not suitable for our setting: following Chen, Varoquaux, and Suchanek (2021), for each entity in KB, except for a list of names, we do not assume the availability of any other information (e.g., entity types or descriptions) or any external resources (e.g., entity co-occurrence). Our formulation is general and suitable for a wide range of real-world settings. In our setting, we propose a KB-enhanced self-supervised pretraining strategy. Since we use a *cloze prompt* to do prompt tuning, we use MLM with entity synonyms for pretraining to keep the consistency between pretraining objective and downstream finetuning task.

We perform experiments on 3 biomedical EL datasets: NCBI disease (Doğan, Leaman, and Lu 2014), BC5CDR (Li et al. 2016) and COMETA (Basaldella et al. 2020). We find that, without pretraining, our model achieves the best results on BC5CDR and COMETA compared with all previous results with finetuning only, showing the effectiveness of cross-entity interaction. With pretraining, our model sets new state-of-the-art results on all the datasets above. The performance gain through our self-supervised pretraining is greater than the previous supervised pretraining on NCBI disease, showing the efficiency of our KB-enhanced pretraining strategy.

Our contributions are summarized as follows:

- We propose a re-ranking model based on prompt tuning, which attends to mention context and all candidate entities together, capturing both mention-entity interaction and entity-entity interaction.
- We propose a KB-enhanced self-supervised pretraining strategy, using masked language modeling (MLM) with entity synonyms in KB, no need for large-scale pretraining with more EL data from external resources.
- We achieve state-of-the-art results on 3 biomedical EL datasets: NCBI disease, BC5CDR and COMETA, showing the effectiveness of cross-entity interaction and KB-enhanced pretraining strategy.

2 Related Work

2.1 Entity Linking

Recent work in entity linking (EL) of the general domain follows a "retrieve and re-rank" two-stage approach. For candidate retrieval, recent years have seen dense embeddings from bi-encoders working accurately and efficiently (Gillick et al. 2019; Botha, Shan, and Gillick 2020). In the re-ranking stage, Transformer-based models are proposed and achieve promising performance gains: a BERT-based cross-encoder that concatenates the context and entity description is frequently used (Logeswaran et al. 2019; Wu et al. 2020). The cross-encoder outputs whether or not the mention in context refers to the concatenated entity. Besides formulating re-ranking as a classification problem, Barba, Procopio, and Navigli (2022) formulate ED as a text extraction problem, and De Cao et al. (2021) use BART (Lewis et al. 2020) to generate corresponding entity name in an autoregressive manner.

Despite huge progress of EL in general domain, the above methods cannot be transferred directly into biomedical domain due to the scarcity of labeled data (Yuan, Yuan, and Yu 2022) and the difference of KB format (Varma et al. 2021). It is a common practice to pretrain models with Wikipedia hyperlinks in general domain, but the labeled EL datasets in biomedical domain is rare, making data-hungry generative models (e.g., De Cao et al. (2021)) hard to implement. Besides lack of data, biomedical KBs often contain incomplete structural resources. Wikidata in the general domain has a highly-organized entity hierarchy and comprehensive entity metadata, but in the biomedical UMLS KB, only 7% of entities have associated descriptions, i.e., most entities only have a list of names. In addition, as Chen, Varoquaux, and Suchanek (2021) suggest, biomedical EL cannot rely on external resources such as alias tables or entity co-occurrence, which are often used in EL systems of general domain. Therefore, it is necessary to propose EL methods specifically for biomedical domain.

Recent biomedical EL approaches (Bhowmik, Stratos, and de Melo 2021) use the bi-encoder architecture. They encode mention words and entity names into the same vector space, and disambiguate mentions by nearest neighbors. Built upon the bi-encoder, Angell et al. (2021) and Varma et al. (2021) add a re-ranking model to boost performance, enabling fine-grained mention-entity interaction. However, none of the above models focus on entity-entity interaction when re-ranking candidates in comparison. Moreover, previous pretraining methods (Yuan, Yuan, and Yu 2022; Varma et al. 2021) do not work in our setting (i.e., the only information for each entity in KB is a list of names). In this work, we focus on the these shortcomings, and propose our model and its accompanied pretraining strategy.

2.2 Prompt Learning

As a series of pretrained language models (PLMs) like GPT (Radford et al. 2019), BERT (Devlin et al. 2019) and BART (Lewis et al. 2020) have been proposed, the "pretrain, finetune" paradigm has demonstrated its effectiveness on various important NLP tasks, such as dialogue (Zhang et al. 2020), summarization (Liu and Lapata 2019) and question answering (Lewis et al. 2021). Despite the success of this paradigm, the huge objective gap between pretraining and finetuning limits the full use of PLM's knowledge for downstream tasks (Liu et al. 2021b). To this end, a new paradigm called "pretrain, prompt and predict" has come into being.

Instead of adapting PLMs to downstream tasks through "pretrain, finetune", downstream tasks are reformulated to look more like pretraining tasks with the help of a textual prompt in the new "pretrain, prompt and predict" paradigm. Prompting means adding instructions or examples before input and output predictions to stimulate knowledge in PLMs. There are mainly two types of prompts: (1) *cloze prompt* (Cui et al. 2021): filling in the blanks of a textual string, suitable for PLMs with MLM pretraining task. (2) *prefix prompt* (Li and Liang 2021): continuing a string prefix, suitable for autoregressive language models. Since the emergence of GPT-3 (Brown et al. 2020), which uses handcrafted prompts and achieves impressive zero-shot and few-shot performance, hand-crafted prompts are trending in various knowledge-intensive tasks to elicit the knowledge in PLMs, such as knowledge probing (Petroni et al. 2019) and entity typing (Ding et al. 2021). In our work, we explore prompt learning in another knowledge-intensive task, i.e., biomedical EL, by stimulating PLMs to capture the contextual information of both mentions and all candidate entities.

3 Method

We formulate the task of biomedical entity linking (EL) as follows: given an **entity mention** m in a biomedical text and a knowledge base (KB) $\mathcal{E}$ consisting of N **entities**, i.e., $\mathcal{E} = \{e_1, e_2, ..., e_N\}$, the task is to find the entity $e_i \in \mathcal{E}$ that m refers to. Following Lai, Ji, and Zhai (2021), We assume that each entity only has a set of names (i.e., synonyms) in KB. For each entity e_i, we use $\mathcal{N}_{e_i}$ to denote the set of all n_{e_i} synonyms of e_i: $\mathcal{N}_{e_i} = \{s^j_{e_i} | j \in \{1, 2, ..., n_{e_i}\}\}$.

Following Wu et al. (2020), we use a "retrieve and rerank" approach to perform biomedical EL. We use a bi-encoder to retrieve K candidates, with unified entity representations and hard negative mining (Section 3.1), and then use a prompt-based model with cross-entity interaction to re-rank the K candidates (Section 3.2). We further propose a self-supervised KB-enhanced pretraining strategy, suitable for our cloze prompt in the re-ranking stage (Section 3.3).

3.1 Candidate Retrieval

Similar to Liu et al. (2021a), we use a bi-encoder initialized from SapBERT to jointly learn representations of mentions and entities. The mention and its surrounding context get encoded in the same dense space where all entity representations lie. Instead of creating n_e entity representations for entity e's n_e synonyms (Sung et al. 2020), we create a unified view for every entity.

Given a mention m with surrounding context and an entity e, the bi-encoder computes:

$$\boldsymbol{y_m} = \mathrm{red}(T_1(\tau_m)) \tag{1}$$

$$\boldsymbol{y_e} = \mathrm{red}(T_2(\tau_e)) \tag{2}$$

where τ_m and τ_e are input representations of mention context and entity, T_1 and T_2 are mention encoder and entity encoder respectively, sharing the same parameters. $\mathrm{red}(.)$ is a function which reduces sequence of vectors into one vector. We choose $\mathrm{red}(.)$ to be the last layer of the output of `[CLS]` token following Humeau et al. (2020).

The input representation of mention τ_m is the word-pieces of context surrounding the mention and the mention itself:

$$\texttt{[CLS]}\ \mathrm{ctxt}_l\ \texttt{[START]}\ m\ \texttt{[END]}\ \mathrm{ctxt}_r\ \texttt{[SEP]}$$

where ctxt_l and ctxt_r are context before and after the mention m respectively. `[START]` and `[END]` are special tokens to tag the mention.

The input representation of entity τ_e is the word-pieces of the concatenation of n_e synonyms with special token `[OR]`:

$$\texttt{[CLS]}\ s^1_e\ \texttt{[OR]}\, s^2_e\ \texttt{[OR]}\ ...\ \texttt{[OR]}\ s^{n_e}_e\ \texttt{[SEP]}$$

The score of (m, e) pair is given by the dot-product:

$$s(m, e) = \boldsymbol{y_m} \cdot \boldsymbol{y_e} \tag{3}$$

Optimization. For each training pair (m, e) (i.e., entity e is the corresponding entity of mention m), the loss is computed as:

$$\begin{aligned}\mathcal{L}(m, e) = &-s(m, e)\\ &+ \log(\exp(s(m, e)) + \sum_{e^{'} \in N(e)} \exp(s(m, e^{'})))\end{aligned} \tag{4}$$

where $N(e) \subset \mathcal{E} \backslash \{e\}$ is a set of negatives that excludes gold entity e. As hard negative mining has shown great potential in noise contrast estimation (Zhang and Stratos 2021), we obtain 90% of $N(e)$ by random sampling from $\mathcal{E} \setminus \{e\}$ and 10% by hard negative mining (i.e. highest-scoring incorrect entities) before every epoch.

Inference. We pre-compute and store entity embedding $\boldsymbol{y_e}$ for every $e \in \mathcal{E}$, and use Faiss (Johnson, Douze, and Jégou 2019) to perform nearest neighbor search for fast top-K retrieval.

3.2 Cross-Entity Re-ranking

Previous cross-encoders concatenate the mention context and one candidate at a time, enabling cross attention only between mention and an entity. We propose a re-ranking model based on prompt learning, which attends to mention context and all candidates together, and thus enables both mention-entity interaction and entity-entity interaction.

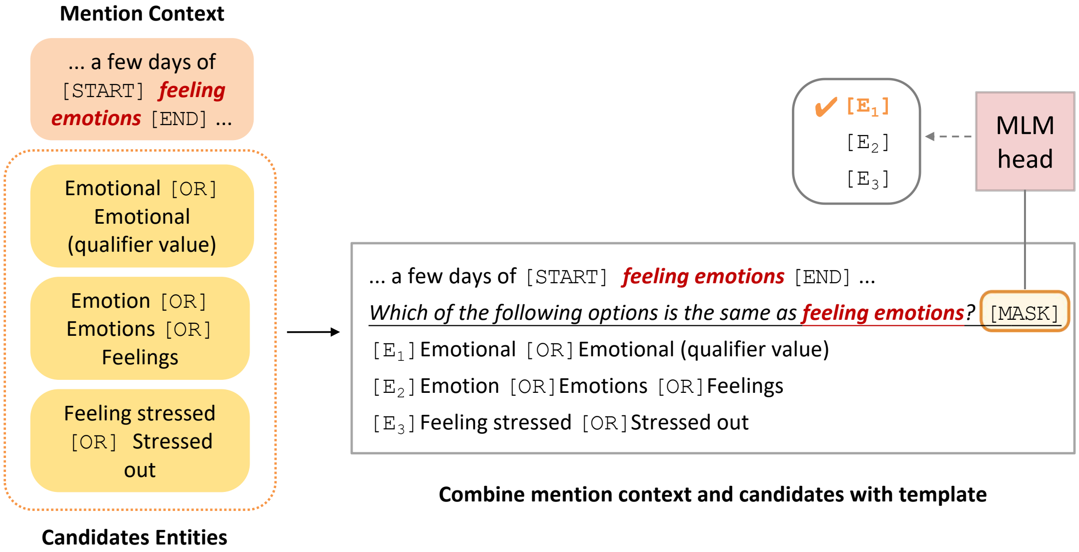
Mention Context
... a few days of [START] feeling emotions [END] ...
Emotional [OR] Emotional (qualifier value)
Emotion [OR] Emotions [OR] Feelings
Feeling stressed [OR] Stressed out
Candidates Entities
... a few days of [START] feeling emotions [END] ...
Which of the following options is the same as feeling emotions? [MASK]
[E1] Emotional [OR] Emotional (qualifier value)
[E2] Emotion [OR] Emotions [OR] Feelings
[E3] Feeling stressed [OR] Stressed out
Combine mention context and candidates with template
✔ [E1]
[E2]
[E3]
MLM head

Figure 2: Illustration of our re-ranker with an example. Our re-ranker combines mention context, multiple entity candidates and a cloze prompt as input. Entity e_i is finally chosen because of the highest probability of filling in [MASK] with the [E$_i$] token.

Synonym ranking. Suppose we have retrieved K candidates (i.e., e_1, e_2, ..., e_K) for the mention m. As PLMs support inputs only up to a fixed maximum length, the length of input tokens for every candidate is limited, so some entity input (i.e., τ_e) might be truncated, leaving out some synonyms. As shown in Yuan, Yuan, and Yu (2022), the surface form similarity (e.g., TF-IDF) displays promising ability in synonym selection. We use the length of the longest common subsequence (LCS) to do a preliminary synonym ranking, putting synonyms with higher lexical similarity at the beginning of the entity input. We rank the synonyms $\mathcal{N}_e$ in a descending order based on the length of LCS between mention m and the synonym s_e^j. Suppose the synonyms of e_i after ranking are listed as $s_e^{k_1}$, $s_e^{k_2}$, ..., $s_e^{k_n}$, the entity e is represented by $\tau_e^{'}$, i.e., word-pieces of the concatenation of the above synonym list with special token [OR]:

$$s_e^{k_1}\ \texttt{[OR]}\, s_e^{k_2}\ \texttt{[OR]}\ ...\ \texttt{[OR]}\ s_e^{k_n}$$

Prompt tuning with entity-entity interaction. We apply a cloze prompt to transform the candidate re-ranking task into masked language modeling (MLM). The input representation $\tau_{m,e}$ of our re-ranking model concatenates mention context, all K candidates with a hand-crafted template, i.e. $\tau_{m,e}$ is defined as "[CLS] ctxt$_l$ [START] m [END] ctxt$_r$ Which of the following options is the same as m? [MASK] [SEP] [E$_1$] $\tau_{e_1}^{'}$ [SEP] [E$_2$] $\tau_{e_2}^{'}$ [SEP] ... [E$_K$] $\tau_{e_K}^{'}$ [SEP]". Figure 2 shows an example of the re-ranking process, solving the choice of the corresponding entity as a MLM task: if the [E$_i$] token has the highest probability of filling in [MASK], we choose entity e_i as the linking result.

Our re-ranking model takes the above input and computes mention-entity embedding $\boldsymbol{y_{m,e}}$, denoted as:

$$\boldsymbol{y_{m,e}} = \text{red}^{'}(T_{cross}(\tau_{m,e})) \qquad (5)$$

where the re-ranking model T_{cross} is a SapBERT, and the function red$^{'}$(.) is the last layer of the output of [MASK] token. In this way, our re-ranking model can see all the possible output choices, letting candidates in comparison attending to each other.

For a given mention m, the score of each candidate e_i is computed by the MLM head's probability of [E$_i$] $\in$ {[E$_1$], [E$_2$], ..., [E$_K$]}, denoted as below:

$$s_{cross}(m, e_i) = \sigma(\boldsymbol{y_{m,e}}\boldsymbol{W}\boldsymbol{h_{e_i}}) \qquad (6)$$

where σ represents the sigmoid function, $\boldsymbol{W}$ is the MLM head which transforms $\boldsymbol{y_{m,e}}$ from the hidden size to the word vocabulary size $|\mathbb{V}|$, and $\boldsymbol{h_{e_i}} \in \mathbb{R}^{|\mathbb{V}|\times 1}$ is the one-hot encoding of the token [E$_i$].

Optimization. For every mention m, we use the gold entity e as the positive example, and use $K-1$ retrieved entities (e is not included) as negative examples. We optimize the re-ranker with a binary cross entropy loss, as below:

$$\begin{aligned}\mathcal{L}(m, e_i) = &-y(m, e_i)\log(s_{cross}(m, e_i))) \\ &+ (1 - y(m, e_i))\log(1 - s_{cross}(m, e_i)))\end{aligned} \qquad (7)$$

where $y(m, e_i) = 1$ for positive examples, $y(m, e_i) = 0$ for negative examples. Note that the concatenation order of $\tau_{e_i}^{'}$ is random during training.

Inference. we use the output score, namely $s_{cross}(m, e_i)$, to choose the best candidate. The concatenation order of $\tau_{e_i}^{'}$ is the ascending order of the distance between mention representation $\boldsymbol{y_m}$ and entity representation $\boldsymbol{y_{e_i}}$ computed by Faiss in the retrieval stage.

3.3 KB-Enhanced Pretraining

We propose to enhance the re-ranker with entity synonyms in KB, because some synonyms are ignored in entity representations. When representing candidate entities with $\tau_e^{'}$ in Section 3.2, we rank the synonyms based on the length of

LCS between them and the mention. However, this strategy may result in truncation of synonyms which have low lexical similarity but high semantic similarity. For example, given the sentence "Patients with generalized atrophic benign epidermolysis bullosa, a usually nonlethal form of **junctional epidermolysis bullosa** ..." and the entity $D016109$ in the MEDIC KB (Davis et al. 2012), the entity's synonym "epidermolysis bullosa junctional herlitz type" is clearly ranked first because of high lexical similarity, and its another synonym "epidermolysis bullosa generalized atrophic benign" has low lexical similarity and will be left out, but "atrophic benign" in the synonym has appeared in the context and could be useful for re-ranking. To mitigate the negative impact of leaving out synonyms, we need to incorporate such synonym information into the PLM.

We use masked language modeling (MLM) to pretrain, suitable for the cloze prompt used in re-ranking. Our pretraining task is aimed at using other synonyms to predict the masked token. For every entity e with more than one name (i.e., $n_e > 1$), we randomly mask one token for every synonym, and then concatenate all synonyms with `[OR]`. Take the above entity $D016109$ for example. It has 3 synonyms, namely "epidermolysis bullosa junctional herlitz type", "epidermolysis bullosa generalized atrophic benign" and "epidermolysis bullosa letali", then the input is "epidermolysis bullosa junctional `[MASK]` type `[OR]` epidermolysis `[MASK]` generalized atrophic benign `[OR]` epidermolysis bullosa `[MASK]`". The MLM head needs to predict 3 `[MASK]` tokens with original tokens , i.e. "herlitz", "bullosa" and "letali" respectively. The backbone of the reranker is a SapBERT, and during pretraining, it is optimized with a cross entropy loss over the token vocabulary.

4 Experiment

4.1 Datasets and Evaluation Metrics

We experiment across three datasets: NCBI disease (Doğan, Leaman, and Lu 2014), BC5CDR (Li et al. 2016) and COMETA (Basaldella et al. 2020). We pre-process the datasets by the following two steps: (1) expand the abbreviations in texts using AB3P (Sohn et al. 2008); (2) lowercase the texts, and mark the beginning and ending of a mention with two special tokens `[START]` and `[END]`. Table 1 shows the basic statistics of datasets and the number of entities and synonyms in their corresponding KBs.

NCBI Disease Corpus (Doğan, Leaman, and Lu 2014) contains manually annotated disease mentions in 793 PubMed abstracts, with CUIs (Concept Unique IDs) mapped into the MEDIC ontology (Davis et al. 2012). In our work, we use the processed data and the target KB provided by Liu et al. (2021a), and follow their evaluation protocol[1].

BC5CDR (Li et al. 2016) is originally designed as a challenge for chemical-induced disease (CID) relation extraction. The dataset consists of 1,500 PubMed article abstracts annotated with disease and chemical entities. All annotations are mapped to MeSH ontology, which comprises a subset of UMLS. Following most recent work (Angell et al. 2021; Varma et al. 2021), we use MeSH contained in UMLS 2017 AA release as the target KB. In our work, we use the data splits and the target KB provided by Yuan, Yuan, and Yu (2022), and follow their evaluation protocol[2].

	NCBI disease	**BC5CDR**	**COMETA**
Ontology	MEDIC	MeSH	SNOMED-CT
Entities	14,967	268,162	350,830
Synonyms	108,071	809,929	904,798
Train samples	5,784	9,285	13,489
Dev samples	787	9,515	2,176
Test samples	960	9,654	4,350

Table 1: Statistics of NCBI disease, BC5CDR and COMETA datasets and the number of entities and synonyms in their corresponding KBs.

COMETA (Basaldella et al. 2020) is a large-scale biomedical EL dataset that specifically focuses on the social media domain, containing 20K medical mentions extracted from the Reddit forum. All mentions are expert-annotated and mapped to SNOMED-CT. We use the "stratified (general)" split and follow the evaluation protocol of the original paper.

Following previous work (Sung et al. 2020; Liu et al. 2021a), we use the top k accuracy as the evaluation metric. We define Acc@k as 1 if a correct entity ID is included in the top k predictions, otherwise 0. For the convenience of comparison, we report Acc@1 and Acc@5.

4.2 Implementation Details

Our retriever and re-ranker are implemented with PyTorch 1.10.0 (Paszke et al. 2019). Both are initialized with SapBERT parameters. Note that the mention encoder and the entity encoder in the retriever share the same parameters. The number of parameters for retriever and re-ranker are roughly 109M and 133M respectively.

The models are trained on a single NVIDIA A100 GPU. We use Adam optimizer (Kingma and Ba 2015) with weight decay set to 0.01 for all experiments. For **retriever finetuning**, we set the learning rate to 2e-6 and batch size to 2 for all datasets. The number of negatives in $N(e)$ in Equation 4 is set to 15. For mention encoder, the maximum length of the input is 256 tokens. For entity encoder, the maximum length of the input is 128 tokens for BC5CDR, and 64 for NCBI disease and COMETA. For all datasets, we finetune for a total of 30 epochs and choose the best checkpoint based on the development set. Finetuning retriever takes 8, 45 and 66 minutes per epoch on NCBI disease, BC5CDR and COMETA respectively. For all datasets, the number of retrieved entities for further re-ranking K is set to 6. For **re-ranker pretraining**, as MEDIC dictionary is a subset of MeSH (Yuan, Yuan, and Yu 2022), for NCBI disease and BC5CDR, we use synonyms of MeSH KB to pretrain. For COMETA, we use synonyms of SNOMED-CT KB to pretrain. We set learning rate to 5e-6, batch size to 64, max-

[1] The processed NCBI disease, COMETA and their KBs are at https://github.com/cambridgeltl/sapbert.

[2] The processed BC5CDR and its KB are at https://github.com/Yuanhy1997/GenBioEL.

	Model	NCBI disease		BC5CDR		COMETA	
		@1	@5	@1	@5	@1	@5
Finetune only	BioSyn (Sung et al. 2020)	91.1	93.9	-	-	71.3	77.8
	SapBERT (Liu et al. 2021a)	92.3	95.5	-	-	75.1	85.5
	ResCNN (Lai, Ji, and Zhai 2021)	**92.4**	-	-	-	80.1	-
	Clustering-based (Angell et al. 2021)	-	-	91.3	-	-	-
	Cross-domain (FT) (Varma et al. 2021)	-	-	89.3	-	-	-
	Generative (FT) (Yuan, Yuan, and Yu 2022)	91.6	95.6	92.6	95.3	80.7	88.7
	Ours (FT)	91.5	**95.7**	**93.0**	**96.7**	**83.3**	**92.4**
Pretrain+Finetune	Cross-domain (PT+FT) (Varma et al. 2021)	-	-	91.5	-	-	-
	Cross-domain (Varma et al. 2021)	-	-	91.9	-	-	-
	Generative (PT+FT) (Yuan, Yuan, and Yu 2022)	91.9	**96.3**	93.3	95.8	81.4	88.2
	Ours (PT+FT)	**92.6**	95.8	**93.7**	**96.6**	**83.7**	**92.3**

Table 2: Results (Acc@1 and Acc@5) of our model compared with previous state-of-the-art methods in NCBI disease, BC5CDR and COMETA when finetuning on train splits only (top) and when pretraining and finetuning (bottom). FT means finetuning and PT means pretraining. Bold denotes the best results. “-” means not reported in the cited paper.

imum input length to 256 tokens, and the number of pretraining epochs to 15 for MeSH and 10 for SNOMED-CT. Pretraining one epoch takes 51 and 69 minutes on MeSH and SNOMED-CT respectively. For **re-ranker finetuning**, we set batch size to 16, maximum mention context length to 256 tokens, maximum candidate entity length to 32 tokens. We search learning rate among [5e-6,1e-5,5e-5] based on the development set. Best-performing learning rate is 5e-5 for all datasets. We finetune for a total of 40 epochs, and choose the best checkpoint based on the development set. Finetuning re-ranker takes 3, 4, 5 minutes per epoch on NCBI disease, BC5CDR and COMETA respectively.

4.3 Baselines

To evaluate the performance of our proposed model, we compare with the following 6 state-of-the-art biomedical EL systems that represent a diverse array of approaches.

- **BioSyn** (Sung et al. 2020) utilizes the synonym marginalization technique and the iterative candidate retrieval for learning biomedical entity representations.
- **SapBERT** (Liu et al. 2021a) designs a metric learning framework that learns to self-align representations for synonymous biomedical entities.
- **ResCNN** (Lai, Ji, and Zhai 2021) proposes an convolutional neural network model with residual connections to compute biomedical entity representations.

Note that the models above are solely based on the synonyms of entities. The approaches listed below use additional entity information (e.g., types and descriptions) or external resources from general domain.

- **Clustering-based** (Angell et al. 2021) introduces cross-encoders to group multiple mentions together via clustering and jointly making linking predictions.
- **Cross-domain** (Varma et al. 2021) combines the bi-encoder from Sung et al. (2020) with the cross-encoder from Angell et al. (2021). Main contribution lies in enriching UMLS with Wikidata and constructing EL datasets with 4.3M mentions for large-scale pretraining.
- **Generative** (Yuan, Yuan, and Yu 2022) injects synonyms and descriptions into the generative language model by creating more training data with manual templates.

4.4 Results

Main results. We compare our approach with 6 previous state-of-the-art models in Section 4.3, and list the performance in Table 2. When only finetuning on train split of each dataset, our approach outperforms all other models on BC5CDR and COMETA, improving 0.4 and 2.6 Acc@1 points over previous finetune-only results. Combined with pretraining strategy, our approach sets new state-of-the-art results on all datasets, outperforming previous best reported results by 0.2, 0.4, 2.3 Acc@1 points. The improvement shows the effectiveness of our overall approach.

Ablation Study. For the ablation of **pretraining**, our pretraining strategy boosts the Acc@1 on NCBI disease by 1.1 points (91.5 → 92.6). The performance increase is larger than the Generative method (0.3 points, 91.6 → 91.9). On BC5CDR, pretraining on our model increases 0.7 points of Acc@1 (93.0 → 93.7), on a par with the Generative method (92.6 → 93.3). On COMETA, pretraining only boosts the Acc@1 by 0.4 points (83.3 → 83.7). The performance increase is smaller on COMETA, and we hypothesize the reason is that the other two datasets are taken from biomedical literature but COMETA is taken from the online forum, which is not in the same domain with the biomedical KB used for our pretraining. Note that the Generative method constructs large-scale EL data partly with entity descriptions, which we do not have in our setting. Compared with pretraining 1 day on 6 A100 GPUs (reported in the paper of Generative method), our pretraining strategy is more efficient, approximately one hour for an epoch on 1 A100 GPU.

For the ablation of our **re-ranker**, we use the same candidate set from our retriever, and implement the cross-encoder proposed in Varma et al. (2021). Compared with our imple-

Mention Context	**Cross-encoder**	**Our re-ranker**

The doctors cut most of her medication and she **vomits** the ones that are left sometimes because of the water.	Vomit;
	Vomitus;
	Vomitus (substance)
		Vomiting;
		Vomiting (disorder);
		Emesis

Invasive dental treatment only becomes an issue after the **valves** have been damaged from some other cause.	Valve; Valve (physical object)	**Valve of vein;** **Structure of valve of vein**
Please go to a doctor to get those **sores** swabbed.	Sore; Sore sensation quality	**Soreness;** **Sore pain**
Radiotherapy is not commonly used but has a role in some cases.	Radiotherapy; Radiation oncology	**Radiation therapy care;** **Radiation therapy management**

I rarely ever **feel hungry** or even peckish.	**Hungry;**
	Hungry (finding);
	Hunger
		Appetite;
		Food appetite;
		Desire for food

Table 3: Examples of top 1 candidate predicted by the cross-encoder in Varma et al. (2021) and our re-ranker. Bold in mention contexts denotes mentions. The synonyms of gold entities are in bold.

	NCBI disease	**BC5CDR**	**COMETA**
Our retriever + cross-encoder	90.9	91.8	83.1
Ours (FT)	91.5	93.0	83.3
Ours (PT+FT)	92.6	93.7	83.7
- words in prompt	92.0	93.1	83.7
- question in prompt	92.0	93.4	83.4

Table 4: Results (Acc@1) of ablations of our model when finetuning on train splits only (top) and when pretraining and finetuning (bottom).

mentation of the cross-encoder, our re-ranker outperforms it by 0.6, 1.2 and 0.2 points of Acc@1, showing the effectiveness of cross attention among candidates introduced in our re-ranker. Case study of predictions by our re-ranker and the cross-encoder are shown in Section 4.5.

For the **prompt** in our re-ranker, we consider 2 variants of $\tau_{m,e}$ in Section 3.2: (1) remove the natural language words in the template, i.e., "$\texttt{[CLS]}$ ctxt_l $\texttt{[START]}$ m $\texttt{[END]}$ ctxt_r m $\texttt{[MASK]}$ $\texttt{[SEP]}$ $[\mathrm{E}_1]$ τ'_{e_1} $\texttt{[SEP]}$ $[\mathrm{E}_2]$ τ'_{e_2} $\texttt{[SEP]}$ ... $[\mathrm{E}_K]$ τ'_{e_K} $\texttt{[SEP]}$" (2) remove the question altogether, i.e., remove m based on (1). Removing the entire question sentence causes a performance drop for all datasets, showing that the template helps the choice of candidates. For NCBI disease and BC5CDR, the removal of natural language words causes 0.6 points of Acc@1 drop, indicating that the words in the template are useful for stimulating the knowledge in the PLM. For COMETA, while removing words does not affect the performance, further removing the mention m causes 0.3 points drop, suggesting that for COMETA, the mention m is more important than natural language words in our hand-crafted template.

4.5 Case Study

To show the effectiveness of our re-ranker, we list the top 1 candidate predicted by our re-ranker and the cross-encoder proposed in Varma et al. (2021) in Table 3. From the first four examples, We can infer that, when facing candidates which both have high lexical similarity with the mention, the cross-encoder is easily influenced by the entity which has a synonym with the same surface form as the mention. However, by letting candidates in comparison attending to each other, our re-ranker can make comprehensive judgements with contextual information and all candidates, and thus disambiguate better on such ambiguous cases.

The last example presents a failure case of our re-ranker. The gold entity's names include "hungry", which has high lexical similarity with the mention, but our re-ranker incorrectly infers "appetite", which is semantically similar to the gold entity. This shows that our re-ranker may benefit from striking a balance between learning lexical similarity and learning semantic similarity.

5 Conclusion

In this work, we focus on the ambiguity in biomedical entity linking. To disambiguate better among entities with high lexical similarity, we propose a prompt-based re-ranking model, which attends to mention context and all candidate entities together, enabling entity-entity interaction through cross attention among candidates. We also propose a KB-enhanced self-supervised pretraining strategy, using masked language modeling with synonyms in KB, no need for large-scale supervised pretraining with extra EL data. Experiments show that we achieve state-of-the-art performance on NCBI disease, BC5CDR and COMETA, showing the effectiveness of cross-entity interaction and the efficiency of our pretraining strategy. Future work may make the prediction not only based on the mention and its candidates, but also based on the candidates of its surrounding mentions.

Acknowledgments

We thank the valuable feedback of Zifei Shan and Senbao Shi and the insightful suggestions of anonymous reviewers. This work is jointly supported by grants: Natural Science Foundation of China (No.62006061 and No.82171475), Stable Support Program for Higher Education Institutions of Shenzhen (No.GXWD20201230155427003-202008241550 11001) and Strategic Emerging Industry Development Special Funds of Shenzhen (No.JCYJ20200109113403826).

References

Angell, R.; Monath, N.; Mohan, S.; Yadav, N.; and McCallum, A. 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*, 2598–2608.

Barba, E.; Pasini, T.; and Navigli, R. 2021. ESC: Redesigning WSD with extractive sense comprehension. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4661–4672.

Barba, E.; Procopio, L.; and Navigli, R. 2022. ExtEnD: Extractive Entity Disambiguation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2478–2488. Dublin, Ireland: Association for Computational Linguistics.

Basaldella, M.; Liu, F.; Shareghi, E.; and Collier, N. 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)*, 3122–3137.

Bhowmik, R.; Stratos, K.; and de Melo, G. 2021. Fast and Effective Biomedical Entity Linking Using a Dual Encoder. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, 28–37. online: Association for Computational Linguistics.

Bodenreider, O. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32 Database issue: D267–70.

Botha, J. A.; Shan, Z.; and Gillick, D. 2020. Entity Linking in 100 Languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7833–7845. Online: Association for Computational Linguistics.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chen, L.; Varoquaux, G.; and Suchanek, F. M. 2021. A Lightweight Neural Model for Biomedical Entity Linking. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14): 12657–12665.

Cui, L.; Wu, Y.; Liu, J.; Yang, S.; and Zhang, Y. 2021. Template-Based Named Entity Recognition Using BART. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 1835–1845. Online: Association for Computational Linguistics.

Davis, A. P.; Wiegers, T. C.; Rosenstein, M. C.; and Mattingly, C. J. 2012. MEDIC: a practical disease vocabulary used at the Comparative Toxicogenomics Database. *Database*, 2012.

De Cao, N.; Izacard, G.; Riedel, S.; and Petroni, F. 2021. Autoregressive Entity Retrieval. In *International Conference on Learning Representations*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Ding, N.; Chen, Y.; Han, X.; Xu, G.; Xie, P.; Zheng, H.; Liu, Z.; Li, J.-Z.; and Kim, H.-G. 2021. Prompt-Learning for Fine-Grained Entity Typing. *ArXiv*, abs/2108.10604.

Doğan, R. I.; Leaman, R.; and Lu, Z. 2014. NCBI disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47: 1–10.

Donnelly, K.; et al. 2006. SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics*, 121: 279.

Gillick, D.; Kulkarni, S.; Lansing, L.; Presta, A.; Baldridge, J.; Ie, E.; and Garcia-Olano, D. 2019. Learning Dense Representations for Entity Retrieval. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 528–537. Hong Kong, China: Association for Computational Linguistics.

Huang, K.-H.; Yang, M.; and Peng, N. 2020. Biomedical Event Extraction with Hierarchical Knowledge Graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1277–1285.

Humeau, S.; Shuster, K.; Lachaux, M.-A.; and Weston, J. 2020. Poly-encoders: Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring. In *ICLR*.

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

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*. San Diego, CA.

Lai, T.; Ji, H.; and Zhai, C. 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*, 1631–1639. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Lee, J.; Sean, S. Y.; Jeong, M.; Sung, M.; Yoon, W.; Choi, Y.; Ko, M.; and Kang, J. 2020. Answering Questions on COVID-19 in Real-Time. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 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*, 7871–7880. Online: Association for Computational Linguistics.

Lewis, P.; Wu, Y.; Liu, L.; Minervini, P.; Küttler, H.; Piktus, A.; Stenetorp, P.; and Riedel, S. 2021. Paq: 65 million probably-asked questions and what you can do with them. *Transactions of the Association for Computational Linguistics*, 9: 1098–1115.

Li, J.; Sun, Y.; Johnson, R. J.; Sciaky, D.; Wei, C.-H.; Leaman, R.; Davis, A. P.; Mattingly, C. J.; Wiegers, T. C.; and Lu, Z. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database: The Journal of Biological Databases and Curation*, 2016.

Li, X. L.; and Liang, P. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 4582–4597. Online: Association for Computational Linguistics.

Liu, F.; Shareghi, E.; Meng, Z.; Basaldella, M.; and Collier, N. 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*, 4228–4238.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2021b. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Liu, Y.; and Lapata, M. 2019. Text Summarization with Pretrained Encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3730–3740.

Logeswaran, L.; Chang, M.-W.; Lee, K.; Toutanova, K.; Devlin, J.; and Lee, H. 2019. Zero-Shot Entity Linking by Reading Entity Descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3449–3460. Florence, Italy: Association for Computational Linguistics.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Petroni, F.; Rocktäschel, T.; Riedel, S.; Lewis, P.; Bakhtin, A.; Wu, Y.; and Miller, A. 2019. Language Models as Knowledge Bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2463–2473.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Sohn, S.; Comeau, D. C.; Kim, W.; and Wilbur, W. J. 2008. Abbreviation definition identification based on automatic precision estimates. *BMC bioinformatics*, 9(1): 1–10.

Sung, M.; Jeon, H.; Lee, J.; and Kang, J. 2020. Biomedical Entity Representations with Synonym Marginalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3641–3650.

Varma, M.; Orr, L.; Wu, S.; Leszczynski, M.; Ling, X.; and Ré, C. 2021. Cross-Domain Data Integration for Named Entity Disambiguation in Biomedical Text. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 4566–4575.

Wu, L.; Petroni, F.; Josifoski, M.; Riedel, S.; and Zettlemoyer, L. 2020. Scalable Zero-shot Entity Linking with Dense Entity Retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6397–6407. Online: Association for Computational Linguistics.

Yuan, H.; and Yu, S. 2021. Efficient symptom inquiring and diagnosis via adaptive alignment of reinforcement learning and classification. *arXiv preprint arXiv:2112.00733*.

Yuan, H.; Yuan, Z.; and Yu, S. 2022. Generative Biomedical Entity Linking via Knowledge Base-Guided Pre-training and Synonyms-Aware Fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4038–4048. Seattle, United States: Association for Computational Linguistics.

Zhang, W.; and Stratos, K. 2021. Understanding Hard Negatives in Noise Contrastive Estimation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1090–1101. Online: Association for Computational Linguistics.

Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.-C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, W. B. 2020. DIALOGPT: Large-Scale Generative Pre-training for Conversational Response Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 270–278.