EMBERT: A Pre-trained Language Model for Chinese Medical Text Mining

Zerui Cai[1], Taolin Zhang[2,3], Chengyu Wang[3], and Xiaofeng He[1](✉)

[1] School of Computer Science and Technology, East China Normal University, Shanghai, China
hexf@cs.ecnu.edu.cn
[2] School of Software Engineering, East China Normal University, Shanghai, China
[3] Alibaba Group, Hangzhou, China
chengyu.wcy@alibaba-inc.com

Abstract. Medical text mining aims to learn models to extract useful information from medical sources. A major challenge is obtaining large-scale labeled data in the medical domain for model training, which is highly expensive. Recent studies show that leveraging massive unlabeled corpora for pre-training language models alleviates this problem by self-supervised learning. In this paper, we propose EMBERT, an entity-level knowledge-enhanced pre-trained language model, which leverages several distinct self-supervised tasks for Chinese medical text mining. EMBERT captures fine-grained semantic relations among medical terms by three self-supervised tasks, including i) context-entity consistency prediction (whether entities are of equivalence in meanings given certain contexts), ii) entity segmentation (segmenting entities into fine-grained semantic parts) and iii) bidirectional entity masking (predicting the atomic or adjective terms of long entities). The experimental results demonstrate that our model achieves significant improvements over five strong baselines on six public Chinese medical text mining datasets.

Keywords: Pre-trained language model · Chinese medical text mining · Self-supervised learning · Deep context-aware neural network

1 Introduction

The outbreak of COVID-19 brings an urgent need for text mining techniques to discover valuable medical information automatically [14,30]. Although a lot of medical texts have been accumulated online, training models for medical text mining often require large-scale annotated data. A significant challenge arises in that labeling high-quality medical data is expensive since the data must be collected by experts with domain knowledge.

Pre-trained Language Models (PLMs) trained on unlabeled data ease the demand for annotated data by self-supervised learning [1,5]. Existing works on PLMs often focus on the general domain. For example, BERT [9], SpanBERT

L. H. U et al. (Eds.): APWeb-WAIM 2021, LNCS 12858, pp. 242–257, 2021.
https://doi.org/10.1007/978-3-030-85896-4_20

Table 1. Characteristics of Chinese medical texts, including i) diversity of synonyms, ii) nestification of entities and iii) misunderstanding of long entities in different medical text mining tasks. The blue underscore contents corresponding to English translations in brackets are shown to explain why this example belongs to the underlying category.

Task	Example	Characteristics
Question Matching	• 我父亲07年8月查出患有结核病，请问用什么药好? (My father was diagnosed with *tuberculosis* in August 2007. Which kind of medicine is suitable?) • 连续三周咳，怀疑是痨病，请问吃什么药? (Coughing for three weeks may be a sign for *consumption*. What medicine should I take?)	Diversity of Synonyms
Named Entity Recognition	• 新型冠状病毒肺炎的症状一般有发热、干咳等。 (Symptoms of *COVID-19* generally include fever, dry coughing, etc.)	Nestification of Entities
Question Answering	• Question: 糖尿病酮酸中毒怎么办 (What to do with *diabetic ketoacidosis*?) • Answer (Correct): 按酸中毒程度不同采取相应治疗措施...(According to the degree of *acidosis*, take appropriate treatment measures...) • Answer (Incorrect): 糖尿病需要综合治疗... (*Diabetes* needs to be treated comprehensively...)	Misunderstanding of Long Entities

[18], XLNet [35] and SemBERT [39] outperform previous models in various downstream NLP tasks [4,16,33,34], which are pre-trained over large-scale unstructured corpora collected from Wikipedia or BookCorpus [9]. However, applying models for the general domain directly to the closed domain usually leads to unsatisfactory result due to the differences in text characteristics [24]. To the best of our knowledge, MC-BERT [36] is the only Chinese medical pre-trained model, which merely applies the whole-word level masking with domain-specific entities and phrases to Chinese medical corpus, neglecting the internal relations of medical entities. We hypothesis that the pre-training method of MC-BERT is sub-optimal, as we observe that there exist three unique characteristics in Chinese medical texts, illustrated in Table 1.

i) **Diversity of Synonyms:** Many terms with different surface forms actually refer to the same concept. Specially, the colloquial expressions and professional terminology of a medical concept may be seemingly irreverent. For example, although "肺结核" (tuberculosis in modern medicine) and "痨病" (consumption in traditional Chinese medicine) mean the same disease, it is difficult for models to learn without medical background knowledge.

ii) **Nestification of Entities:** In the Chinese medical knowledge graph, a lot of Chinese medical entities contain multiple sub-entities. For example, in the entity "新型冠状病毒肺炎" (COVID-19), both "肺炎" (pneumonia) and “新型冠状病毒” (novel coronavirus) are also entities in the KG. In previous works, MC-BERT [36] only masks the complete entity, neglecting the fine-grained information of the sub-entities.

iii) **Misunderstanding of Long Entities:** Besides the above characteristics, there are also strong semantic relations among those sub-entities. Existing open-domain PLMs [31,36] do not consider the semantic relations between the core entities (named atomic terms) [21] and the entities other than the atomic terms in the long entities (named adjective terms). Refer to the example w.r.t. “糖尿病” (diabetes) and “糖尿病酮酸” (diabetic ketoacidosis) in Table 1.

In this paper, we propose EMBERT, a pre-trained model for Chinese medical text mining[1]. EMBERT leverages three novel self-supervised tasks for pre-training that are utilized to model Chinese medical entities in fine grains:

- **Context-Entity Consistency Prediction**: In the medical knowledge graph[2], we leverage the relation *“SameAs”* to build a thesaurus and replace terms with their synonyms in the corpus to generate more training samples. For each sample, we promote our model to predict whether the entities for replacing are consistent with its context.
- **Entity Segmentation:** We segment long entities by our rule-based system and label the resulting sub-entities contained in the entity. Then, our model is trained to predict sub-entities with labels mentioned above as ground-truth.
- **Bidirectional Entity Masking:** For each long entity, we merge the sub-entities into an adjective term and an atomic term. Meanwhile, we propose a bidirectional masking strategy to capture internal semantic relations within the long entity. We mask the adjective term and predict it based on the representation of atomic terms and vice versa.

In the experiments, we choose BERT-base [9], BERT-wwm [6], RoBERTa [23], ERNIE [38], MC-BERT [36] as the baseline models and apply our model to six Chinese medical datasets to evaluate its performance. The results shows that EMBERT achieves significant improvement compared to strong baselines on all six datasets. In summary, our work contributions are as follows:

- We propose a novel Chinese medical PLM by modeling the distinct characteristics of medical terms, which is named EMBERT.
- Three self-supervised learning tasks are introduced to capture the semantic relations at the entity level, including context-entity consistency prediction, entity segmentation and bidirectional entity masking.
- Experimental results on six Chinese medical text mining datasets show that our model achieves significant improvement over strong baselines.

The rest of this paper is organized as follows. Section 2 summarizes the related work on PLMs. Details of our approach for Chinese medical text mining are

[1] “EMBERT” refers to Entity-rich Medical BERT.

[2] http://www.openkg.cn/.

described in Sect. 3. Implementation detailed and experimental results are presented in Sect. 4. Finally, We summarize our paper and discuss the future work in Sect. 5.

2 Related Work

We overview the related work on open-domain and domain-specific PLMs.

2.1 Pre-trained Language Models in the Open Domain

As discovered, the meaning of a word depends on the context [2,8,29]. Hence, several PLMs have been proposed to learn context-aware word distributed representations. ELMo [26] is proposed to extract context-sensitive features leveraging bidirectional long short-term memory networks (LSTMs) [13]. However, feature-based language models such as ELMo only produce token representations that serve as basic input features, rather than acting as a backbone encoder. Recently, a two-stage training paradigm, namely pre-training and fine-tuning, is proposed to train models on large-scale corpora to learn general syntactic and semantic knowledge. Next, the models are fine-tuned on downstream tasks. SA-LSTMs [7] is proposed to train auto-encoders by LSTM, achieving a more stable training process and generalizing better. OpenAI GPT [27] utilizes multiple transformer decoder layers [32], and learns contextualized token representations by unidirectional auto-regressive language model objective. BERT [9] (as well as its robustly optimized version RoBERTa [23]) is trained based on bidirectional transformer architecture by two novel self-supervised tasks, including mask language modeling (MLM) and next sentence prediction (NSP). Following BERT, a large number of PLMs have been proposed to further improve performance in various NLP tasks, leveraging the following three techniques, such as self-supervised pre-training (Baidu-ERNIE [31] and spanBERT [18]), encoder architectures (XLNet [35]) and multi-task learning (MT-DNN [22]).

2.2 Pre-trained Language Models in Medical Domain

Developing PLMs in the medical domain has been a hot topic recently. To the best of our knowledge, BioBERT [20] is the first work that preforms continuous pre-training on a biomedical domain corpora (PubMed abstracts and PMC full-text articles) based on BERT in English. BlueBERT [25] is pre-trained on PubMed abstracts and MIMIC-III clinical notes and evaluated on the Biomedical Language Understanding Evaluation (BLUE) benchmark. ClinicalBert [15] utilizes the clinical notes including lab values and medications instead of plain-text data based on BERT. Meanwhile, PubMedBERT [10] learns model weights from scratch by large-scale training corpus and argues that the key point of training domain-specific PLMs is learning from scratch, which can obtain an in-domain vocabulary, alleviating the out-of-vocabulary (OOV) problem. Yet there are very few works on Chinese medical PLM, mainly due to the limitations of data resources. The work MC-BERT [36] proposes the entity masking

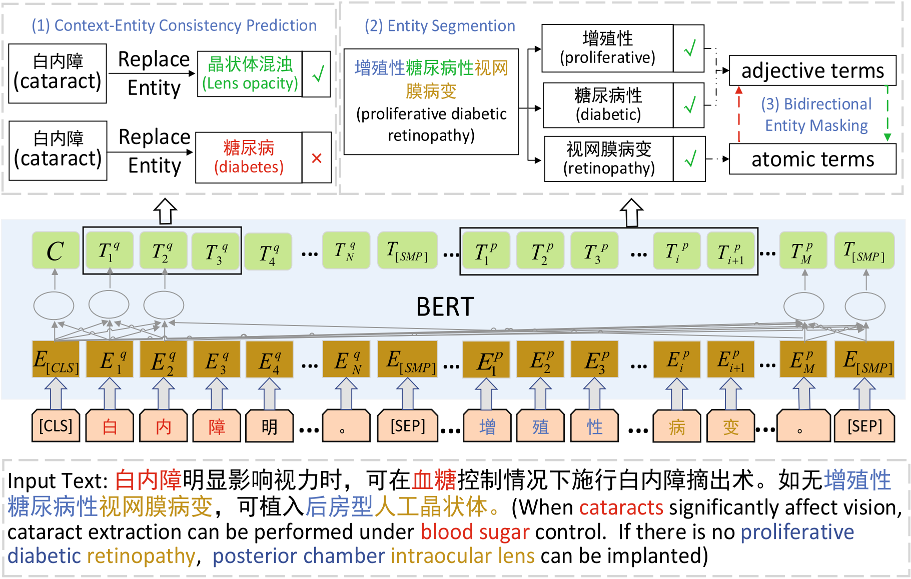

(1) Context-Entity Consistency Prediction
白内障 (cataract)
Replace Entity
晶状体混浊 (Lens opacity)
√
白内障 (cataract)
Replace Entity
糖尿病 (diabetes)
×
(2) Entity Segmention
增殖性糖尿病性视网膜病变 (proliferative diabetic retinopathy)
增殖性 (proliferative)
√
糖尿病性 (diabetic)
√
视网膜病变 (retinopathy)
√
adjective terms
(3) Bidirectional Entity Masking
atomic terms
C
T_1^q T_2^q T_3^q T_4^q ... T_N^q $T_{[SMP]}$ ... T_1^p T_2^p T_3^p ... T_i^p T_{i+1}^p ... T_M^p $T_{[SMP]}$
BERT
$E_{[CLS]}$ E_1^q E_2^q E_3^q E_4^q ... E_N^q $E_{[SMP]}$ ... E_1^p E_2^p E_3^p ... E_i^p E_{i+1}^p ... E_M^p $E_{[SMP]}$
[CLS] 白 内 障 明 ... 。 [SEP] ... 增 殖 性 ... 病 变 ... 。 [SEP]
Input Text: 白内障明显影响视力时，可在血糖控制情况下施行白内障摘出术。如无增殖性糖尿病性视网膜病变，可植入后房型人工晶状体。(When cataracts significantly affect vision, cataract extraction can be performed under blood sugar control. If there is no proliferative diabetic retinopathy, posterior chamber intraocular lens can be implanted)

Fig. 1. Model overview. EMBERT incorporates three novel self-supervised pre-training tasks: context-entity consistency prediction, entity segmentation and bidirectional entity masking. Contents in brackets refer to English translations. (Best viewed in color.) (Color figure online)

and phrase masking mechanisms in a coarse-grained aspect to learn the medical word representations from a medical corpora while neglects the internal relations of medical entities. In this paper, we propose three novel self-supervised tasks in fine-grained entity-level aspects to further enhance the understanding of Chinese medical texts.

3 The EMBERT Model

In this section, we formally present EMBERT, a self-supervised PLM for Chinese medical text mining. We first introduce three novel pre-training tasks, namely context-entity consistency prediction, entity segmentation and entity bidirectional masking. Finally, we give the whole training loss of EMBERT. Figure 1 illustrates our model architecture.

3.1 Context-Entity Consistency Prediction

In Chinese medical knowledge graphs, there is a kind of relation “SameAs”, meaning that two entities refer to the same concept, such as “白内障” (cataract) and “晶状体混浊” (phacoscotasmus). We build a thesaurus leveraging the relations mentioned above and replace a target entity with its synonym or a randomly selected entity, increasing the training samples simultaneously. In this task, we promote our model to predict whether the given entity is consistent with its context. If an entity is replaced with its synonym, the meaning would still be consistent. Formally, we denote the output token representations of the PLM by $\mathbf{x_1}, \cdots, \mathbf{x_n}$. The output tokens of the i-th replaced entity are $x_{s_i}, \cdots, x_{e_i}$, where (s_i, e_i) means the starting position and the ending position of the entity. Particularly, we predict the consistency of the entity with its context as c_i using the output context-aware encodings of its boundary token $\mathbf{x}_{s_i-1}$ and $\mathbf{x}_{e_i+1}$ as $\mathbf{y}_i$:

$$\mathbf{y}_i = f\left(\mathbf{x}_{s_i-1}, \mathbf{x}_{e_i+1}\right) \tag{1}$$

We implement the representation function $f_c(\cdot)$ as a 1-layer feed-forward network with the GeLU activate function [12] and layer normalization [3].

$$\begin{aligned} \mathbf{h}_i &= \left[\mathbf{x}_{s_i-1}; \mathbf{x}_{e_i+1}\right] \\ \mathbf{y}_i &= \text{ LayerNorm } \left(\text{GeLU}\left(\mathbf{W}_1 \mathbf{h}_i\right)\right) \end{aligned} \tag{2}$$

where $\mathbf{W}_1$ is the trainable matrix. We then utilize the vector representation $\mathbf{y}_i$ to predict c_i. The loss function of this task (denoted as $\mathcal{L}_{eccp}$) is shown as follows:

$$\begin{aligned} \mathbf{p}_\theta(c_i \mid \mathbf{y_i}) &= \text{SoftMax}(\mathbf{W}_2 \mathbf{y}_i) \\ \mathcal{L}_{eccp} &= -\frac{\sum_{i=1}^{N} m_i \log \mathbf{p}_\theta\left(c_i \mid \mathbf{y_i}\right)}{N} \end{aligned} \tag{3}$$

where $\mathbf{W}_2$ is the trainable matrix, m_i is the ground-truth label (consistent or inconsistent) and N is the total number of entities.

3.2 Entity Segmentation

As long entities usually have complicated semantic meanings, in this task, we promote the model to segment the entities into semantic parts. The ground-truth labels are given by our rule-based system depicted in the Appendix.

In practice, the model is asked whether the given position t is the end of a sub-entity or not, and the prediction from the model is marked as s_t. Formally, given the i-th nested entity $x_{s_i}, \cdots, x_{e_i}$, we further split it into j-th fine-grained sub-entities $x_{s_{ij}}, \cdots, x_{e_{ij}}$. As the last tokens of sub-entities are the split positions, we label $x_{e_{ij}}$ as positive, with the rest of the tokens in the entities labeled as negative. Tokens in the non-entity part of the sentence are ignored. We implement the token representation $\mathbf{y_t}$ and the loss of this task similar to token-level masked language modeling while we only have two categories for the model to predict:

$$\begin{aligned} \mathbf{y_t} &= \tanh\left(\mathbf{W}_3 \text{LayerNorm}\left(\text{GeLU}\left(\mathbf{W}_4 \mathbf{x_t}\right)\right) + \mathbf{b}\right) \\ \mathbf{p}_\theta(s_t \mid \mathbf{y_i}) &= \text{SoftMax}(\mathbf{W}_5 \mathbf{y}_t) \end{aligned} \tag{4}$$

where the matrices $\mathbf{W}_3, \mathbf{W}_4$ and $\mathbf{W}_5$ are initialized randomly. The loss function of entity segmentation $\mathcal{L}_{est}$ is as follows:

$$\mathcal{L}_{est} = -\sum_{i=1}^{N} m_t \log \mathbf{p}_\theta\left(s_t \mid \mathbf{y_t}\right) \tag{5}$$

where m_t is the ground-truth label and N is the length of the sentence.

3.3 Bidirectional Entity Masking

We observe that long entities can be further divided into two parts: adjective terms and atomic terms. In this task, we mask one of the components and predict it based on the other and vice versa. Hence, the bidirectional masking strategy can model the relationship between semantic units in long entities.

Formally, we denote the adjective term as $x_{s_{adj_i}}, \cdots, x_{e_{adj_i}}$ and the atomic term as $x_{s_{ato_i}}, \cdots, x_{e_{ato_i}}$. We take the case of masking the atomic term as an example. We represent the token in the atomic term utilizing the output hidden state vector $\mathbf{x}_{s_{adj_i}}, \mathbf{x}_{e_{adj_i}}$, as well as the relative position embedding $\mathbf{p}_{j-s_{ato_i}}$ of the target token:

$$\mathbf{y}_j = f_b\left(\mathbf{x}_{s_{adj_i}}, \mathbf{x}_{e_{adj_i}}, \mathbf{p}_{j-s_{ato_i}}\right) \tag{6}$$

The representation function $f_b(\cdot)$ is 2-layer feed-forward network with GeLU and layer normalization similar to spanBERT [18]:

$$\begin{aligned} \mathbf{h}_j^0 &= \left[\mathbf{x}_{s_{adj_i}}; \mathbf{x}_{e_{adj_i}}; \mathbf{p}_{j-s_{adj_i}}\right] \\ \mathbf{h}_j^1 &= \text{LayerNorm } \left(\text{GeLU}\left(\mathbf{W}_6 \mathbf{h}_j^0\right)\right) \\ \mathbf{y}_j &= \text{LayerNorm } \left(\text{GeLU}\left(\mathbf{W}_7 \mathbf{h}_j^1\right)\right) \end{aligned} \tag{7}$$

We use the vector representation $\mathbf{y}_j$ to predict the token x_j and compute the cross-entropy loss $\mathcal{L}_{bem}^{ato_i}$ for the i-th entity similar to MLM:

$$\begin{aligned} \mathbf{p}_\theta(x_j \mid \mathbf{y}_j) &= \frac{\exp(\mathbf{y}_j \cdot \mathbf{w}_j)}{\sum_{k=1}^{K} \exp(\mathbf{y}_j \cdot \mathbf{w}_k)} \\ \mathcal{L}_{bem}^{ato_i} &= -\sum_{j=s_{ato_i}}^{e_{ato_i}} m_j \log \mathbf{p}_\theta\left(x_j \mid \mathbf{y}_j\right) \end{aligned} \tag{8}$$

where K is the size of the vocabulary and $\mathbf{w}_j$ is the representation of the embedding layer of the true token in the position j in the original sentence.

Those $\mathcal{L}_{bem}^{ato_i}$ sum to $\mathcal{L}_{bem}^{ato}$. Similarly, We acquire $\mathcal{L}_{bem}^{adj}$ for predicating the adjective term and the loss of this task $\mathcal{L}_{bem}$ is the sum of the two mentioned loss functions, i.e.,

$$\mathcal{L}_{bem} = \mathcal{L}_{bem}^{ato} + \mathcal{L}_{bem}^{adj} \tag{9}$$

3.4 Overll Loss Function

In summary, the total loss of EMBERT is the sum of four losses:

$$\mathcal{L}_{total} = \mathcal{L}_{ex} + \lambda_1 \mathcal{L}_{cecp} + \lambda_2 \mathcal{L}_{est} + \lambda_3 \mathcal{L}_{bem} \tag{10}$$

where $\mathcal{L}_{ex}$ is the existing loss function used in BERT [9]. λ_1, λ_2 and λ_3 are the hyper-parameters in this model.

Table 2. The statistical data and metrics of the six datasets.

Dataset	Train	Dev	Test	Task	Metric
cNNER *	12000	3000	5000	Nested-NER	F1
cMedQANER [36]	1,673	175	215	NER	F1
cMedQQ [36]	16,071	1,793	1,935	PI	F1
cMedQNLI [36]	80,950	9,065	9,969	NLI	F1
cMedQA [37]	186,771	46,600	46,600	QA	ACC@1
WebMedQA [11]	252,850	31,605	31,655	QA	ACC@1

* cNNER is the Chinese Nested Named Entity Recognition task released in CHIP 2020. (http://cips-chip.org.cn/2020/eval1)

4 Experiments

In this section, we conduct extensive experiments to evaluate the performance of EMBERT over multiple datasets. We also compare EMBERT with strong baselines to show its superiority for Chinese medical text mining.

4.1 Experimental Settings

The pre-training data used in EMBERT is collected from the DXY community medical question answering data[3] and the DXY BBS data[4]. The total amount of data is 5 GB. The pre-processing of the pre-training corpus is similar to that of BERT [9]. For one document in the corpus, we use punctuation as the split symbol to generate text segments. We aggregate these text segments into raw training samples that are no longer than 512 tokens. Details on processing Chinese medical entities are further described in the Appendix.

The model configurations of all BERT-based models are the same as BERT-base[5]. We use a linear warmup schedule with a peak value of 5e-5 and the warmup proportion is 10% of the total training steps. The AdamW optimizer [19] is used with default parameters ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e{-}8$) and a decoupled weight decay of 0.01. Our implementation uses a batch size of 256 with a maximum length of 512. For the bidirectional entity masking task, we use 200-dimension positional embeddings. The hyper-parameters $\lambda_1, \lambda_2, \lambda_3$ of the pre-training loss are $2, 2, 0.5$[6] respectively. The pre-training process is run on a single RTX-Titan GPU and takes nearly ten days to complete.

[3] https://portal.dxy.cn/.

[4] https://www.dxy.cn/bbs/newweb/pc/home.

[5] https://huggingface.co/bert-base-chinese.

[6] In the experiment, we try several groups of hyper-parameters and find that the setting [2, 2, 0.5] performs well.

Table 3. Performance of the five baseline models and EMBERT on six datasets. ♣ and ♠ indicate EMBERT initialized by BERT-base and MC-BERT, respectively.

Model	Dataset					
	cMedQQ	cMedNLI	cMedQANER	cNNER	cMedQA	WebMedQA
MC-BERT	87.16	96.36	83.99	66.61	74.46	80.54
ERNIE-THU	87.03	96.04	84.43	66.87	74.13	79.96
BERT-wwm	86.82	96.08	83.12	66.59	72.96	79.68
RoBERTa	86.97	96.11	83.29	66.72	73.18	79.57
BERT-base	86.72	96.06	83.07	66.46	73.82	79.72
EMBERT♣	87.59	96.50	84.49	67.07	75.10	80.51
EMBERT♠	**88.06**	**96.59**	**85.02**	**67.22**	**75.32**	**80.63**

4.2 Baseline Models and Downstream Task Datasets

In our experiments, we choose five strong PLMs as our baselines. I) BERT-base [9] is the PLM trained by two self-supervised tasks, including MLM (Masked Language Model) and NSP (Next Sentence Prediction). II) BERT-wwm [6] explicitly forces the model to recover the whole word in pre-training tasks, which is much more challenging. III) RoBERTa [23] changes the model hyper-parameters, training strategies, and the corpus, with the BERT model re-trained. IV) MC-BERT [36] proposes the entity masking and phrase masking self-supervised tasks in the Chinese medical domain. V) ERNIE [36] infuses knowledge graph embedding generated by TransE algorithm into BERT layer, thus equipping BERT with structural knowledge of KG. Note that we generate TransE embedding with our KG and pre-train ERNIE with the same setting of EMBERT.

Our experimental datasets include the following six datasets, involving two Named Entity Recognition (NER) tasks, two Question Answering (QA) tasks, one Natural Language Inference (NLI) task and one Paraphrase Identification (PI) task. The dataset statistical results are shown in Table 2. Note that cMedQANER, cMedQQ, cMedQNLI and cMedQA are from ChineseBLUE[7]. The cMedQANER dataset is labeled from the Chinese community question answering dataset. Each cMedQQ sample contains a question pair with the task of predicting whether the two sentences are similar. Both cMedQNLI and cMedQA consist of question-answer pairs from BBS. The cNNER dataset is from the Chinese medical NER evaluation of CHIP 2020. WebMedQA is a real-world Chinese medical question answering dataset collected from online health consultancy websites [11].

[7] https://github.com/alibaba-research/ChineseBLUE. We do not include other datasets in our experiments due to their small sizes.

4.3 Overall Model Results

Table 3 shows the performance of EMBERT and the baselines on each dataset. Compared with general-domain PLMs, MC-BERT and EMBERT achieve much larger improvement, which demonstrates that it is essential to perform close-domain pre-training for obtaining promising results. Also, it can be seen that EMBERT achieves notable improvement pre-trained from both BERT-base and MC-BERT, demonstrating the effectiveness of our model and suggests EMBERT boosts performance in a different way from MC-BERT. Besides, although EMBERT uses a much smaller training corpus, our model outperforms MC-BERT significantly on most datasets. The only exception is WebMedQA (EMBERT 80.51% vs MC-BERT 80.54% on ACC@1). We suggest that is because the texts of the dataset are more similar to the corpus used by MC-BERT.

4.4 Ablation Studies

To evaluate the effects of three important components in our model, we remove them and test our model on four datasets, respectively. Since MC-BERT uses BERT-base as the starting point of pre-training, we also evaluate EMBERT from BERT-base rather than MC-BERT to avoid the influence of MC-BERT. The experimental results are summarized in Table 4.

As we can see, the performance of EMBERT drops greatly when we remove any components from our model. This phenomenon suggests that our pre-training methods are beneficial across a variety of tasks. On the two NER datasets, the effect of *Entity Segmentation* plays a major role in performance improvement, and we conjecture that it is because this mechanism injects entity boundary information into the model, which is critical for the NER task.

For cMedQQ, most of the performance degradation is caused by removing the *Context-Entity Consistent Prediction* task (–0.37% on F1). Note that the questions in this task are relatively short, and the contexts of words are relatively incomplete. Therefore, the ability to match keywords with possibly different surface forms between question pairs is highly important. On the other hand, since our model learns lots of synonyms with the aforementioned mechanism, it is reasonable to achieve a notable improvement.

The *Bidirectional Entity Masking* mechanism improves the performance of the cMedQA task performance significantly. We manually check 100 samples where EMBERT predicts the right answers and EMBERT without bidirectional entity masking does not. According to our observation, the EMBERT without the mechanism often makes mistakes because parts of the long entities are often treated as normal words, while the complete EMBERT does not as it can understand the long entities better as we expect.

4.5 Analysis of Attention Weight Distributions

In our EMBERT model, we propose three pre-training mechanisms to capture semantic relations between and inside Chinese medical entities. To further verify

Table 4. Ablation studies on four datasets.

Model	Dataset			
	cMedQQ	cMedQANER	cMedQA	cNNER
EMBERT$^{\spadesuit}$	**88.06**	**85.02**	**75.32**	**67.22**
EMBERT$^{\clubsuit}$	87.59	84.49	75.10	67.07
MC-BERT	87.16	83.99	74.46	66.61
w/o entity consistency	**87.22** | **−0.37** ↓	84.22	74.68	66.84
w/o bidirectional mask	87.43	84.28	**74.25**| **−0.85**↓	67.03
w/o entity segmentation	87.54	**84.14**| **−0.35**↓	74.89	**66.71**| **−0.36**↓
w/o all above	87.02	83.76	74.03	66.48

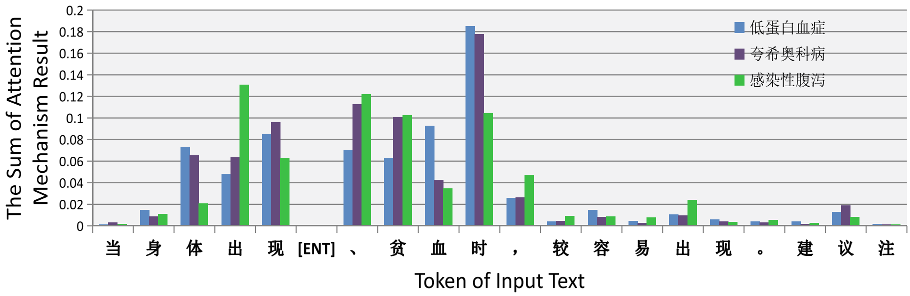
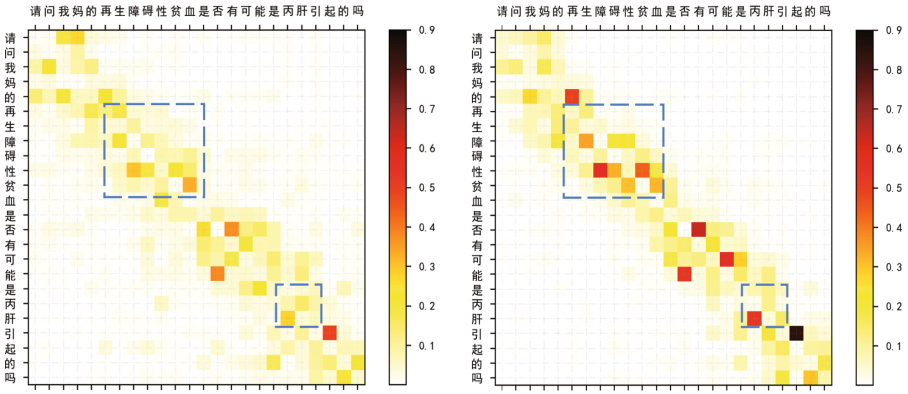


Fig. 2. The sum of attention weights of each token to “[ENT]” from the intermediate layer of EMBERT. “[ENT]” refers to three types of entity. “低蛋白血症” (hypoproteinemia) and “夸希奥科病” (kwashiorkor) are the synonymous entities. “感染性腹泻” (gastroenteritis) is an entity randomly selected from the Chinese medical knowledge graph. Due to the limitation of the page width, we only show 20 tokens in the sentence.

the effect of our mechanisms, we perform two additional intrinsic experiments as described below.

Attention Weight Similarity. We hypothesize that an entity should have a similar influence on the attention weight to the context with its synonym than that of a randomly selected entity. Therefore, we take a sentence from our pre-training corpus an entity “低蛋白血症” (hypoproteinemia), and then replace the entity with its synonymous term and a randomly selected entity respectively as an example in Fig. 2. The resulting three sentences are the same anywhere except on the position of the replaced entity. We feed the three sentences into EMBERT separately and take out the attention matrix at layer 11. The attention from the positions of the replaced entity to other positions in the sentences is summed for each sentence. We can easily tell that the attention weights of the entity “低蛋白血症” (hypoproteinemia) are much more similar to the attention weights of the synonym “夸希奥克病” (hypoproteinemia), which has a very different surface form than that of the randomly selected entity “感染性腹泻” (gastroenteritis).

Attention Weights Heat Map. In addition, we analyze the effect of mechanisms on long entities in our model. Figure 3 illustrates the self-attention token weights in BERT-base and EMBERT. Each row values are attention weights from the corresponding token to all tokens, which sum to one. The darker the colors of the squares in the figure, the greater the similarity between the tokens learned by the model. In this example, “再生障碍性贫血” (aplastic anemia) is a long entity and “丙肝” (HCV) is a short entity in Chinese. It can be seen from Fig. 3 that EMBERT pays the ***most*** attention to other tokens within the same entity while the attention weights of BERT are *scattered* over all tokens (see Fig. 3 blue line of dashes). Hence, EMBERT represents entities in sentences much better than BERT. Meanwhile, we also find that tokens in adjective terms attend to atomic terms in EMBERT much more than those in BERT, such as “贫血 ” (anemia) in “再生障碍性贫血” (aplastic anemia).



(a) Chinese medical long entity attention result based on BERT-base model

(b) Chinese medical long entity attention result based on EMBERT model

Fig. 3. Chinese long entity attention weights sum based on different pre-trained language models from all heads at layer 10. The blue line of dashes illustrate the self-attention weights of some important Chinese entities in the sentence. (Color figure online)

Table 5. Results on Chinese medical tasks with different corruption rates.

Corruption rate	cMedQQ	cMedQANER	cMedQA	cNNER
7.5%	87.49	84.80	74.89	66.93
15%	**88.06**	85.02	**75.32**	67.22
25%	87.80	**85.28**	75.11	**67.28**

4.6 Varying the Corruption Rate

In this section, we discuss the impact of different corruption rates. We pre-train EMBERT with different corruption rates and test them on four datasets across three tasks. The results are summarized in Table 5. In general, we find that the corruption rates have a limited effect, which is consistent with previous work [28]. The only exception is that NER tasks have maintained a steady improvement with the corruption rate increases, and we hypothesize that it is because the model has more chances to learn boundary information on a large-scale corpus. However, a larger corruption rate does degrade performances on other tasks. In order to keep a balance of the performances of EMBERT on a variety of tasks, we use a corruption rate of 15% as default.

5 Conclusion and Future Work

In this paper, we propose a large-scale PLM for Chinese medical text mining, namely EMBERT. Specifically, EMBERT captures internal and external entity-level semantic relations by three self-supervised tasks. As a result, our model achieves significant improvement over five strong baselines on six Chinese medical text mining datasets. Note that it is possible to further expand our method for other languages while overcoming the differences in language characteristics and the lack of resources. In the future, we will try to gather texts from different sources for evaluating the effect of EMBERT in more fine-grained domains.

Acknowledgements. This work is supported by the National Key Research and Development Program of China under Grant No. 2016YFB1000904. We thank the anonymous reviewers for their careful reading and insightful comments on our manuscript.

Appendix

Entity Segmentation. To segment long entities, we first build an entity vocabulary with a cut-point set. Similar to MC-BERT, we leverage AutoPhrase [17] to harvest a set of high-quality entities in the medical domain from the training corpus. Those entities combined with entities in the KG form the final entities vocabulary. Meanwhile, we utilize the segmentation model generated by AutoPhrase to create primitive segmentation results of entities. Next, we calculate the frequency of the characters at the start or the end of each segment. The characters with the top-100 frequency are manually checked and can be used as hints of segmentation, which are selected to form the cut-point set.

Long and Short Entity Detection. Initially, we choose entities in the vocabulary that are longer than three characters as long entity candidates. For the other entities, we regard them as short entities and use them as user-defined dictionary for Jieba[8], a popular Chinese word segmentation tool. For each long entity candidate, we first split at the positions of characters in the cut-point set, then feed each split part into Jieba, and combine all the return values to get intermediate segmentation results for the long entity candidate. Additionally, for long entities candidates being cut into too many single characters, we treat them as errors, and use the segmentation model from AutoPhrase to correct the segmentation results of those long entities candidates. Finally, if a long entities candidate can not be segmented into any smaller pieces, we regard it as a short entity. The remaining long entity candidates are treated as true long entities.

References

1. Alyafeai, Z., AlShaibani, M.S., Ahmad, I.: A survey on transfer learning in natural language processing. CoRR arXiv:2007.04239 (2020)
2. Apresjan, J.D.: Regular polysemy. Linguistics **12**(142), 5–32 (1974)
3. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv:1607.06450 (2016)
4. Baumann, A.: Multilingual language models for named entity recognition in German and English. In: RANLP, pp. 21–27 (2019)
5. Chronopoulou, A., Baziotis, C., Potamianos, A.: An embarrassingly simple approach for transfer learning from pretrained language models. In: NAACL, pp. 2089–2095 (2019)
6. Cui, Y., et al.: Pre-training with whole word masking for chinese BERT. CoRR arXiv:1906.08101 (2019)
7. Dai, A.M., Le, Q.V.: Semi-supervised sequence learning. In: NIPS, pp. 3079–3087 (2015)
8. Deane, P.D.: Polysemy and cognition. Lingua **75**(4), 325–361 (1988)
9. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT, pp. 4171–4186 (2019)
10. Gu, Y., et al.: Domain-specific language model pretraining for biomedical natural language processing. CoRR arXiv:2007.15779 (2020)
11. He, J., Fu, M., Tu, M.: Applying deep matching networks to Chinese medical question answering: A study and a dataset. BMC Med. Inf. Decis. Making **19**(2), 91–100 (2019)
12. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv:1606.08415 (2016)
13. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
14. Hosseini, P., Hosseini, P., Broniatowski, D.A.: Content analysis of Persian/Farsi tweets during COVID-19 pandemic in Iran using NLP. CoRR arXiv:2005.08400 (2020)
15. Huang, K., Altosaar, J., Ranganath, R.: Clinicalbert: modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:1904.05342 (2019)
16. Jiang, Z., El-Jaroudi, A., Hartmann, W., Karakos, D.G., Zhao, L.: Cross-lingual information retrieval with BERT. In: CLSSTS@LREC, pp. 26–31 (2020)

[8] https://github.com/fxsjy/jieba.

17. Shang, J., Liu, J., Jiang, M., Ren, X., Voss, C.R., Han, J.: Automated phrase mining from massive text corpora. IEEE Trans. Knowl. Data Eng. **30**(10), 1825–1837 (2018)
18. Joshi, M., Chen, D., Liu, Y., Weld, D.S., Zettlemoyer, L., Levy, O.: Spanbert: improving pre-training by representing and predicting spans. Comput. Linguist. **8**, 64–77 (2020)
19. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv:1412.6980 (2014)
20. Lee, J., et al.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)
21. Liu, Q., Wu, L., Yang, Z., Liu, Y.: Domain phrase identification using atomic word formation in Chinese text. Knowl. Based Syst. **24**(8), 1254–1260 (2011)
22. Liu, X., He, P., Chen, W., Gao, J.: Multi-task deep neural networks for natural language understanding. In: ACL, pp. 4487–4496 (2019)
23. Liu, Y., et al.: Roberta: a robustly optimized BERT pretraining approach. CoRR arXiv:1907.11692 (2019)
24. Liu, Z., Huang, D., Huang, K., Li, Z., Zhao, J.: Finbert: a pre-trained financial language representation model for financial text mining. In: IJCAI, pp. 4513–4519 (2020)
25. Peng, Y., Yan, S., Lu, Z.: Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. arXiv:1906.05474 (2019)
26. Peters, M.E., et al.: Deep contextualized word representations. In: NAACL, pp. 2227–2237 (2018)
27. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI Blog **1**(8), 9 (2019)
28. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. CoRR arXiv:1910.10683 (2019)
29. Ravin, Y., Leacock, C.: Polysemy: An Overview. Polysemy: Theoretical and Computational Approaches, pp. 1–29 (2000)
30. Sarker, A., Lakamana, S., Hogg-Bremer, W., Xie, A., Al-Garadi, M.A., Yang, Y.: Self-reported COVID-19 symptoms on Twitter: an analysis and a research resource. J. Am. Med. Inf. Assoc. **27**(8), 1310–1315 (2020)
31. Sun, Y., et al.: ERNIE: enhanced representation through knowledge integration. CoRR arXiv:1904.09223 (2019)
32. Vaswani, A., et al.: Attention is all you need. In: NIPS, pp. 5998–6008 (2017)
33. Xu, H., Liu, B., Shu, L., Yu, P.S.: BERT post-training for review reading comprehension and aspect-based sentiment analysis. In: NAACL, pp. 2324–2335 (2019)
34. Xu, S., Shen, X., Fukumoto, F., Li, J., Suzuki, Y., Nishizaki, H.: Paraphrase identification with lexical, syntactic and sentential encodings. Appl. Sci. **10**(12), 4144 (2020)
35. Yang, Z., Dai, Z., Yang, Y., Carbonell, J.G., Salakhutdinov, R., Le, Q.V.: Xlnet: generalized autoregressive pretraining for language understanding. In: NeurIPS, pp. 5754–5764 (2019)
36. Zhang, N., Jia, Q., Yin, K., Dong, L., Gao, F., Hua, N.: Conceptualized representation learning for Chinese biomedical text mining. In: WSDM 2020 HealthDay (2020)

37. Zhang, S., Zhang, X., Wang, H., Cheng, J., Li, P., Ding, Z.: Chinese medical question answer matching using end-to-end character-level multi-scale CNNs. Appl. Sci. **7**(8), 767 (2017)
38. Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., Liu, Q.: ERNIE: enhanced language representation with informative entities. In: ACL, pp. 1441–1451 (2019)
39. Zhang, Z., et al.: Semantics-aware BERT for language understanding. In: AAAI, pp. 9628–9635 (2020)