Hindawi
Scientific Programming
Volume 2020, Article ID 8812754, 9 pages
https://doi.org/10.1155/2020/8812754

Research Article

Named Entity Recognition in Chinese Medical Literature Using Pretraining Models

Yu Wang,[1,2] Yining Sun ,[1,2] Zuchang Ma,[1] Lisheng Gao,[1] and Yang Xu[1]

[1]*Anhui Province Key Laboratory of Medical Physics and Technology, Institute of Intelligent Machines, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei 230031, China*
[2]*University of Science and Technology of China, Hefei 230026, China*

Correspondence should be addressed to Yining Sun; ynsun@iim.ac.cn

Received 12 July 2020; Revised 8 August 2020; Accepted 20 August 2020; Published 9 September 2020

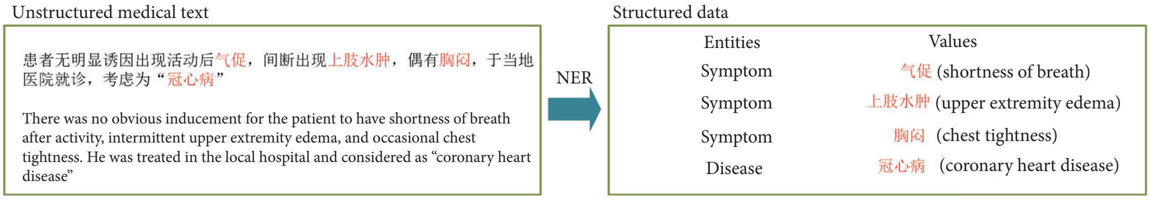
Academic Editor: David Ruano-Ordás

The medical literature contains valuable knowledge, such as the clinical symptoms, diagnosis, and treatments of a particular disease. Named Entity Recognition (NER) is the initial step in extracting this knowledge from unstructured text and presenting it as a Knowledge Graph (KG). However, the previous approaches of NER have often suffered from small-scale human-labelled training data. Furthermore, extracting knowledge from Chinese medical literature is a more complex task because there is no segmentation between Chinese characters. Recently, the pretraining models, which obtain representations with the prior semantic knowledge on large-scale unlabelled corpora, have achieved state-of-the-art results for a wide variety of Natural Language Processing (NLP) tasks. However, the capabilities of pretraining models have not been fully exploited, and applications of other pretraining models except BERT in specific domains, such as NER in Chinese medical literature, are also of interest. In this paper, we enhance the performance of NER in Chinese medical literature using pretraining models. First, we propose a method of data augmentation by replacing the words in the training set with synonyms through the Mask Language Model (MLM), which is a pretraining task. Then, we consider NER as the downstream task of the pretraining model and transfer the prior semantic knowledge obtained during pretraining to it. Finally, we conduct experiments to compare the performances of six pretraining models (BERT, BERT-WWM, BERT-WWM-EXT, ERNIE, ERNIE-tiny, and RoBERTa) in recognizing named entities from Chinese medical literature. The effects of feature extraction and fine-tuning, as well as different downstream model structures, are also explored. Experimental results demonstrate that the method of data augmentation we proposed can obtain meaningful improvements in the performance of recognition. Besides, RoBERTa-CRF achieves the highest *F*1-score compared with the previous methods and other pretraining models.

1. Introduction

In recent decades, it has been generally known that the rapid growth of information technology has resulted in huge amounts of information generated and shared in the field of medicine, where the number of published documents, such as articles, books, and technical reports, is increasing exponentially [1]. For example, PubMed houses over 380,000 publications found by just searching the keyword "Diabetes" (Jan. 2009 to Oct. 2019). The medical literature contains valuable knowledge, such as the clinical symptoms, diagnosis, and treatments of a particular disease. However, it is time-consuming and laborious for medical researchers to obtain knowledge from these documents. Thus, it is critical to extract information and knowledge from unstructured medical literature using novel information extraction techniques and present the findings in a visually intuitive Knowledge Graph which supports machine-understandable information about the medicine [2, 3].

Named Entity Recognition (NER) is the fundamental task in Natural Language Processing (NLP). It is also the initial step in extracting valuable knowledge from unstructured text and building a medical Knowledge Graph (KG). As shown in Figure 1, NER aims to recognize entities from unstructured text, and the results of NER may affect



患者无明显诱因出现活动后气促，间断出现上肢水肿，偶有胸闷，于当地医院就诊，考虑为“冠心病”

There was no obvious inducement for the patient to have shortness of breath after activity, intermittent upper extremity edema, and occasional chest tightness. He was treated in the local hospital and considered as “coronary heart disease”

Entities	Values
Symptom	气促 (shortness of breath)
Symptom	上肢水肿 (upper extremity edema)
Symptom	胸闷 (chest tightness)
Disease	冠心病 (coronary heart disease)

Figure 1: An example of NER.

subsequent knowledge extraction tasks, such as the Relation Extraction (RE). In the early years, researchers used rule-based or dictionary-based methods for NER tasks [4, 5]. However, these methods lack generalization, for they are proposed for particular types of entities. Traditional machine learning and deep learning methods emerging in recent years are also used in NER tasks [6]. Nevertheless, the performance of these methods often suffers from small-scale human-labelled training data, resulting in poor generalization capability, especially for rare words. Moreover, recognizing entities from Chinese documents is a more complex task because there is no segmentation between Chinese characters. Furthermore, in the field of Chinese medical literature, some English symbols, such as the chemical symbols Na and K, may appear in the documents, which makes the NER task more difficult. Therefore, it is of interest to know whether the prior semantic knowledge can be learned from large amounts of unlabelled corpora to improve the performance of NER.

Recently, pretraining models (e.g., BERT and ERNIE) have achieved state-of-the-art (SOTA) results on several NLP tasks. The pretraining models obtain prior semantic knowledge from large-scale unlabelled corpora through pretraining tasks and improve the performance of downstream tasks by transferring this knowledge to them. However, the capabilities of pretraining models have not been fully exploited, and most of the previous works have focused on BERT [7, 8], but applications of other pretraining models in specific domains, such as NER in Chinese medical literature, are also of interest.

In this paper, we enhance the performance of NER in Chinese medical literature using pretraining models. The dataset we used is “A Labelled Chinese Dataset for Diabetes (LCDD),” which contains authoritative Chinese medical literature in recent seven years. The main contributions of this paper can be summarized as follows:

(1) Firstly, we proposed a method of data augmentation based on the Masked Language Model (MLM). Pretraining models will predict the masked words during the procedure of MLM, which can be used for synonym replacement to augment the training set [9]. Considering that there is no segmentation between Chinese characters, we choose ERNIE to conduct this task because it has the entity-level and phrase-level masking strategies.

(2) Secondly, we consider NER as a downstream task of six kinds of pretraining models (BERT, BERT-WWM, BERT-WWM-EXT, ERNIE, ERNIE-tiny, and RoBERTa) and transfer the prior semantic knowledge obtained during pretraining to the downstream task to enhance the performance.

(3) Finally, exhaustive experiments are conducted based on the LCDD dataset. We compare the performance of the NER task on the original and augmented training set. Meanwhile, in addition to comparing the pretraining models with previous methods, we compare the six pretraining models to each other. Moreover, we also explore the performance under different downstream models and two main approaches: feature extraction and fine-tuning. Experimental results demonstrate that the method of data augmentation we proposed can obtain meaningful improvements in the performance of recognition. Besides, RoBERTa-CRF based on the augmented training set with fine-tuning obtains the SOTA result.

2. Related Work

In this section, we will introduce the related works of the Named Entity Recognition, pretraining models, and data augmentation.

2.1. Named Entity Recognition. The Named Entity Recognition aims to identify chunks of text which refer to specific entities of interest, such as drugs, symptoms, treatments, and diseases. Rule-based and dictionary-based approaches had played an important role. For example, Gerner et al. [10] used a dictionary-based approach to identify species names in biomedical literature. Fukuda et al. [11] proposed a rule-based method to extract material names such as proteins from biological documents. However, these methods lack generalization because they need hand-craft rules. Researchers also tried using machine learning methods to recognize entities from unstructured data. He et al. [12] presented a CRF-based approach to recognize drug names in biomedical texts. Wang et al. [13] compared six biomedical NER tools based on the Hidden Markov Model (HMM) and Conditional Random Field (CRF). Nevertheless, machine learning methods need to choose a set of features manually, which is time-consuming and laborious. In recent years, deep learning methods, which can improve the performance of NER without feature engineering, have received increasing attention. For example, Zhu et al. [14] proposed an end-to-end deep learning approach for biomedical NER tasks which leverages the local contexts via Convolutional Neural Network (CNN). For Recurrent Neural Network (RNN), Chen et al. [15] used a Bidirectional Long Short-Term Memory (BiLSTM) model for the NER from Chinese adverse drug event reports. Chen et al. [16] used dictionary features to help identify rare and unseen clinical named entities. However, deep learning methods still suffer from insufficient training data.

2.2. Pretraining Models. Recently, the pretraining models, which generate representations of words with prior semantic knowledge on large-scale unlabelled corpora, have achieved state-of-the-art results for a wide variety of NLP tasks [17]. Various pretraining models have emerged after Devlin et al. [18] released BERT in 2018. These models consist of multilayer bidirectional Transformer blocks [19]. The main differences among pretraining models lie in the pretraining tasks and pretraining corpora. Table 1 shows the difference in detail. We denote the number of Transformer layers as L, the hidden size as H, and the number of self-attention heads as A. During the procedure of the Next Sentence Prediction (NSP), which is a kind of pretraining task, the pretraining models are trained to predict whether two sentences have a contextual relationship, and the pretraining models can understand the relationship between the sentences in this way.

For the NER task, Devlin et al. [18] first consider NER as a downstream task of BERT for extracting named entities from the news (MSRA-NER). Pires et al. [7] realized zero-shot NER through multilingual BERT. Besides, pretraining models are also used on domain-specific NER, such as biomedicine. For example, Hakala and Pyysalo [8] applied a CRF-based baseline approach and multilingual BERT to the Spanish biomedical NER task. However, the capabilities of pretraining models have not been fully exploited. Furthermore, applications of other pretraining models except BERT in specific domains, such as NER in Chinese medical literature, are also of interest.

2.3. Data Augmentation. A common approach of data augmentation in the area of NLP is synonym replacement [24]. A previous work found synonyms with k-nearest neighbours using Word2Vec [25]. However, the MLM of pretraining models is more suitable for synonym replacement. It is not only because the word representations obtained by the pretraining models contain more abundant semantic knowledge than previous models but also because Word2Vec cannot handle polysemous words. Wu et al. [9] proposed a method of data augmentation based on BERT. However, BERT will mask the Chinese characters, not words, during the procedure of the MLM because there is no segmentation between Chinese characters. Therefore, we perform data augmentation based on ERNIE because it has entity-level and phrase-level masking strategies in the MLM process. The method of data augmentation will be presented in Section 3.1.

3. Methods

3.1. Data Augmentation Using ERNIE. As mentioned earlier, the Masked Language Model (MLM) is intensely suitable for data augmentation. During the procedure of the MLM, a certain portion (e.g., 15%) of words are replaced by a special symbol [MASK], and the pretraining model is trained to predict the masked word. Specifically, for a token sequence $x = \{x_1, \ldots, x_T\}$, the pretraining model first constructs a corrupted sequence $\widehat{x}$ by randomly setting a portion of tokens in $\boldsymbol{x}$ to a special symbol [MASK] [26]. The training objective is to reconstruct $\overline{x}$ from $\widehat{x}$:

$$\max_{\theta} \log_{p_\theta}(\overline{x} \mid \widehat{x}) = \sum_{t=1}^{T} m_t \log_{p_\theta}\left(x_t \mid \widehat{x}\right), \tag{1}$$

where $m_t = 1$ indicates that x_t is masked. The whole process is like a *Cloze task* [18]. We repeat the process of MLM using a trained pretraining model. The model is not retrained and is only used to predict masked words. Obviously, the words predicted by the model can be regarded as the synonyms of the masked words. We perform data augmentation based on ERNIE because it has entity-level and phrase-level masking strategies in the MLM process.

A visualization of the process can be seen in Figure 2. ERNIE randomly masks a portion of characters or words in the input sequence by default [21]. It is worth noting that masking the named entities is not appropriate because these entities may be proper nouns or rare words in medical literature, especially the disease and drug entities like "糖尿病 (diabetes)" and "胰岛素 (insulin)" in Figure 2. When ERNIE predicts these entities, the result may not be correct Chinese words because the information of these entities may not be obtained during pretraining. Therefore, we only randomly mask the tokens except for named entities. Furthermore, we input a single sequence that starts with a particular classification token [CLS] and ends with an ending token [SEP], because the context information of sentence pairs is not necessary, which is different from inputting sentence pairs during pretraining [18, 21]. As shown in Figure 2, one sequence input into ERNIE consists of the following four parts:

(1) Token IDs: We use the original vocabulary provided by ERNIE to get the ID number of each token.

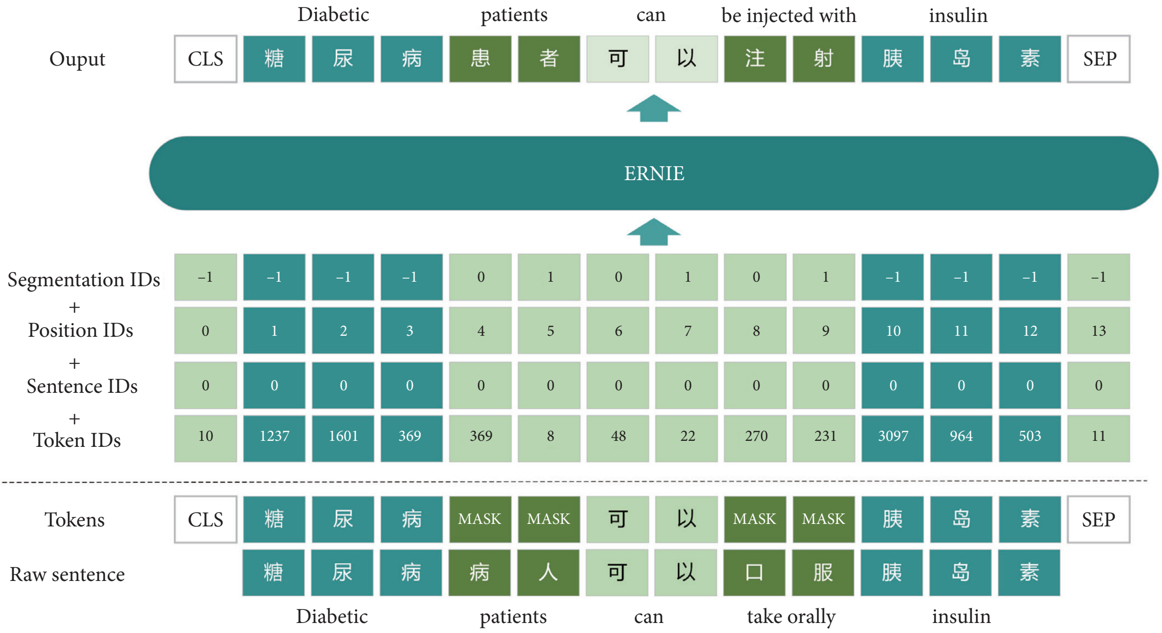
(2) Sentence IDs: ERNIE uses this mark to determine the sentences to which the token belongs. As mentioned earlier, we input the single sentence, not a sentence pair. Accordingly, all the sentence ID numbers are 0.

(3) Position IDs: The Transformer cannot obtain position information through self-attention heads, since it contains no recurrence and no convolution [19]. Therefore, the position ID number is injected to get information about the relative or absolute position of the tokens.

(4) Segmentation IDs: The segmentation IDs represent the segmentation information. Specifically, "0" denotes the beginning of a word, and "1" does not

Table 1: Parameters, pretraining tasks, and corpora of pretraining models.

Pretraining model	L	H	A	Pretraining task	Pretraining corpora
BERT [18]	12	768	12	Masked Language Model, NSP	Books Corpus, Wikipedia
BERT-WWM [20]	12	768	12	Whole Word Masking, NSP	Wikipedia
BERT-WWM-EXT [20]	12	768	12	Whole Word Masking, NSP	General data (Baike, News, and QA), Wikipedia
ERNIE [21]	12	768	12	Phrase-level and entity-level masking, NSP	Chinese Wikipedia, Baidu Baike, News, and Tieba
ERNIE-tiny [22]	3	1024	12	Phrase-level and entity-level masking, NSP	Chinese Wikipedia, Baidu Baike, News, and Tieba
RoBERTa [23]	12	768	12	Dynamic masking	Books Corpus, Wikipedia



Ouput	CLS	糖	尿	病	患	者	可	以	注	射	胰	岛	素	SEP
Segmentation IDs	–1	–1	–1	–1	0	1	0	1	0	1	–1	–1	–1	–1
Position IDs	0	1	2	3	4	5	6	7	8	9	10	11	12	13
Sentence IDs	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Token IDs	10	1237	1601	369	369	8	48	22	270	231	3097	964	503	11
Tokens	CLS	糖	尿	病	MASK	MASK	可	以	MASK	MASK	胰	岛	素	SEP
Raw sentence		糖	尿	病	病	人	可	以	口	服	胰	岛	素	

Figure 2: Data augmentation using ERNIE.

denote the beginning. Moreover, we assign "–1" to the corresponding position of [CLS], [SEP], and named entities. ERNIE will not mask the token where the segmentation ID equals "–1." We use THULAC (http://thulac.thunlp.org/) for word segmentation [27].

As can be seen in Figure 2, "病人 (patients)" and "口服 (take orally)" in the raw sentence are replaced by "患者 (patients)" and "注射 (be injected with)," respectively. These two groups of words are synonyms in Chinese. We perform the above operation on all samples in the training set to obtain the dataset D'. Finally, we combine the dataset D' generated by ERNIE with the original training data D to get the augmented training data D_{aug}.

3.2. Named Entity Recognition Using Pretraining Models. We consider NER in medical literature as the downstream task of the pretraining model. As the pretraining models are pretrained on large-scale unlabelled corpora, the output of pretraining models can be regarded as the representations of tokens with prior semantic knowledge. The key to using a pretraining model for NER is how to transfer the prior semantic knowledge obtained from the source domain to the target domain (e.g., Chinese medical literature NER in this paper). There are two main approaches to transfer the prior semantic knowledge to the downstream tasks: feature extraction and fine-tuning [28]. For feature extraction, the parameters of pretraining models are fixed and only the parameters in downstream models are trained through the downstream task. The pretraining models are regarded as the feature extractors and output the representations of tokens with prior semantic knowledge in the source domain. The representations, which are higher-level and more abstract features, will be input into the downstream task. On the other hand, for fine-tuning, all the parameters of pretraining models and downstream models are trained through the downstream task. The pretraining models will learn the semantic knowledge of the target domain from the training data of the downstream tasks. These two approaches are illustrated by Figure 3, where areas marked by blue squares indicate that the parameters of the corresponding models are trained through the downstream task.

For the structure of downstream model, we test the following three common modules: Full Connection (FC), LSTM, and CRF. As shown in Figure 3, the LSTM and CRF are optional. The performance of different modules will be shown in the fourth section.

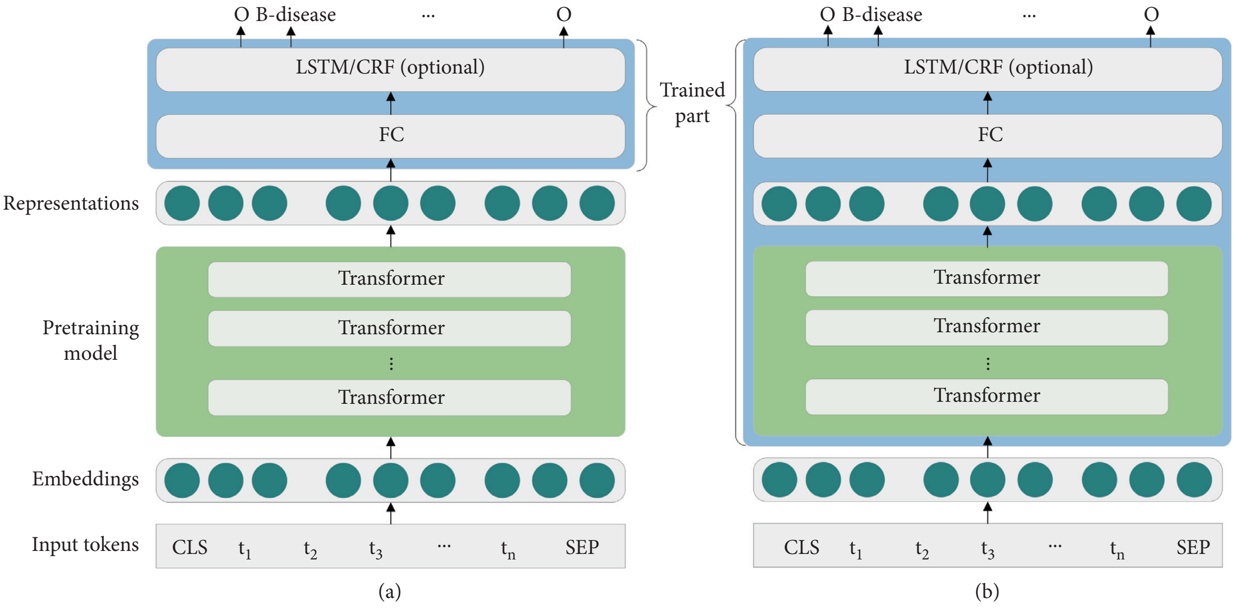


Figure 3: Two main approaches: (a) feature extraction and (b) fine-tuning.

4. Experiments and Results

In this section, we will introduce the dataset for the NER task and show the results. The experiments were performed with PaddlePaddle, which is a framework of deep learning. For hardware, we used an eight-core CPU and a V100 GPU.

4.1. Dataset. The dataset we used is “A Labelled Chinese Dataset for Diabetes,” which is provided by Alibaba Cloud [29]. This dataset comes from the authoritative Chinese diabetes journals in recent seven years, from which the literature related to basic research, clinical research, drug usage, diagnosis, and treatment methods are selected. The dataset covers the latest research hotspots on diabetes and is labelled by professionals with a medical background. We divided this dataset into training set, development set, and test set within the ratio of 6 : 2 : 2. The details of the labels are given in Table 2.

4.2. Experiment Settings. We tested the performance of NER from the following three aspects:

(1) Using the method of data augmentation we proposed

(2) Using pretraining models and common deep learning models like the BiLSTM

(3) Using downstream models with different structures

Firstly, we tested the performance using the original dataset and the augmented dataset. Then, the performance of pretraining models, including the BERT series, ERNIE, and RoBERTa, was compared with common deep learning models, such as BiLSTM. Finally, we compared the performance when the downstream model is the LSTM or CRF. For the pretraining models, the parameters were established based on the pretrained parameters provided by their authors. For the downstream models, the weights were established using Xavier initialization, while the biases were initialized as 0.

The hyperparameters are set up based on trial and error. We evaluated the performance at every 1000 steps on the development set, and the experiment would be terminated prematurely once the loss no longer drops. The final selection of the hyperparameters would be the best on the development set. All the hyperparameters involved are listed in Table 3.

For the evaluation, we introduced the precision, recall, and $F1$-score. The precision value refers to the ratio of correct entities to predicted entities. The recall value is the proportion of the entities in the test set which are correctly predicted. The $F1$-score is calculated according to the following formulation:

$$F1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (2)$$

It can be seen that the $F1$-score is the harmonic mean of the precision and recall, which can comprehensively reflect the performance of the model on NER tasks. We use P, R, and F to represent precision, recall, and $F1$-score, respectively.

4.3. Results. Firstly, we tested the effects of data augmentation method we proposed. The augmented dataset is obtained through the MLM of ERNIE as described in the third section. We used three pretraining models (BERT, ERNIE, and ERNIE-tiny) based on the original dataset and augmented dataset, respectively. The parameters of pretraining models are updated through fine-tuning. The downstream model is a single-layer FC without the CRF or LSTM. The results are shown in Table 4.

The performance of NER in Chinese medical literature can be improved when using the augmented dataset,

Table 2: Statistics of "A Labelled Chinese Dataset for Diabetes."

	Training set	Development set	Test set
Disease related			
Disease	25197	8399	8399
Reason	2849	950	950
Symptom	3166	1055	1056
Test	28819	9606	9606
Test value	6402	2134	2134
Therapy related			
Drug	9946	3315	3315
Frequency	309	103	103
Amount	871	290	290
Method	606	202	203
Treatment	896	298	299
Operation	493	164	164
Side effect	1052	351	350
Common entities			
Duration	6543	2181	2180
Anatomy	16866	5622	5622
Level	1333	446	448
Total	105348	35116	35119

Table 3: Hyperparameters.

Parameters	Values
Learning rate	$5e-5$
Batch size	32
Weight decay	0.01
Epoch	6
Optimizer	Adam optimizer

and the $F1$-score can be increased by approximately 0.14% on average. The subsequent experiments are all based on the augmented dataset.

Then, we compared the performance when using pretraining models and common deep learning models. The results are shown in Table 5. The parameters of pretraining models are also updated during fine-tuning, and the downstream model is a single-layer FC without the CRF or LSTM, too. As we can see from Table 5, using pretraining models can obtain meaningful improvements in the performance of NER. Among pretraining models, the $F1$-score of ERNIE-tiny is the lowest, at only 89.466%. In contrast, RoBERTa obtained the highest $F1$-score with 91.209%. Moreover, the performance of BERT series models (BERT, BERT-WWM, and BERT-WWM-EXT) is relatively higher than that of ERNIE.

Furthermore, we also compared the two main approaches transferring prior semantic knowledge to the NER task: feature extraction and fine-tuning. For feature extraction, we fixed the parameters of pretraining models. On the contrary, the parameters of pretraining models were trainable and can be updated during fine-tuning based on the training set. The downstream model structure is also a single-layer FC without the CRF or LSTM. The results shown in Table 6 indicate that the $F1$-score can be slightly increased through fine-tuning.

Table 4: Recognition results of original dataset and augmented dataset.

	P (%)	R (%)	F (%)
Models			
BERT	90.778	90.674	90.726
ERNIE	90.488	90.354	90.421
ERNIE-tiny	89.361	89.162	89.261
Models (augmented dataset)			
BERT	90.968	91.048	91.008
ERNIE	90.659	90.527	90.593
ERNIE-tiny	89.466	89.348	89.407

Table 5: Recognition results of pretraining models and deep learning models.

	P (%)	R (%)	F (%)
Deep learning models			
BiGRU	89.431	81.842	85.443
BiGRU-CRF	88.463	84.332	86.341
BiLSTM	89.511	82.251	85.700
BiLSTM-CRF	89.113	84.983	86.992
Pretraining models			
BERT	90.968	91.048	91.008
BERT-WWM	91.023	91.108	91.065
BERT-WWM-EXT	91.059	90.996	91.027
ERNIE	90.659	90.527	90.593
ERNIE-tiny	89.466	89.348	89.407
RoBERTa	91.164	91.254	91.209

Finally, we also tested the performance of different downstream model structures. RoBERTa was used as the pretraining model in this test. For the downstream model, we tested the FC, CRF, LSTM-CRF, and BiLSTM-CRF, respectively. For LSTM-CRF and BiLSTM-CRF, the dimension of the hidden layer was 128. It can be found from Table 7 that the performance of recognition reduced when a fairly complex model was used as the downstream model.

5. Discussion

In this section, we will discuss the experimental results in detail.

5.1. Data Augmentation. Results also show that the augmentation method we proposed can increase the $F1$-score by approximately 0.14% on average. Although the improvement is not significant, the result is meaningful for it demonstrates that the data augmentation using ERNIE is feasible. As mentioned in Section 2.3, BERT will mask the Chinese characters, not words, during the procedure of the MLM because there is no segmentation between Chinese characters, and the results may not be grammatically correct Chinese sentences. However, the MLM of ERNIE can replace a portion of Chinese phrases or words with synonyms. The semantics of the new Chinese sentences generated by ERNIE are similar to those of the original sentences, and they are combined as the augmented dataset. We do not mask the named entities in light of these entities which may

Table 6: Recognition results of feature extraction and fine-tuning.

Pretraining models	P (%)	R (%)	F (%)
BERT (feature extraction)	90.881	90.983	90.932
BERT (fine-tuning)	90.968	91.048	91.008
ERNIE (feature extraction)	90.519	90.639	90.579
ERNIE (fine-tuning)	90.659	90.527	90.593
RoBERTa (feature extraction)	91.109	**91.275**	91.192
RoBERTa (fine-tuning)	**91.164**	91.254	**91.209**

Values in bold represent the maximum values.

Table 7: Recognition results of different downstream model structures.

Pretraining models	P (%)	R (%)	F (%)
RoBERTa-FC	91.164	91.254	91.209
RoBERTa-CRF	91.187	91.358	91.270
RoBERTa-LSTM-CRF	90.615	90.784	90.697
RoBERTa-BiLSTM-CRF	90.820	90.911	90.650

be proper nouns or rare words in the field of medical literature. The results also demonstrate that the augmentation method we proposed is meaningful and feasible.

5.2. Comparison of Pretraining Models with Common Deep Learning Methods. Obviously, using pretraining models can obtain meaningful improvements in the performance of NER. The pretraining models have learned abundant prior semantic knowledge from the pretraining corpora (e.g., Chinese Wikipedia and Baidu News) [20, 21]. Pretraining corpora can also be regarded as the "source domain." When conducting the NER task, the prior semantic knowledge will be transferred to the downstream task, which can also be known as the "target domain." The whole process can be regarded as transfer learning. Task-specific semantic knowledge contained in the target domain will be obtained during fine-tuning.

On the contrary, the common deep learning models can only learn knowledge from the training set, also known as the target domain. The training process is done from scratch on the target domain, whether it is the baseline model (BiLSTM-CRF) or other deep learning models. Therefore, these models can only learn the knowledge in the target domain from the training set. The experimental results also indicated that using pretraining models can get a meaningful increase in the $F1$-score by at least 3%.

5.3. Comparison between Pretraining Models. We also compared the performances of the six most common pretraining models for NER in Chinese medical literature: BERT, BERT-WWM, BERT-WWM-EXT, ERNIE, ERNIE-tiny, and RoBERTa. First of all, it is shown that the deeper the layer, the better the performance for the pretraining models with similar pretraining tasks and the same pretraining corpus, such as ERNIE and ERNIE-tiny. ERNIE has twelve Transformer layers, but ERNIE-tiny only has three Transformer layers. Although ERNIE-tiny increases the number of hidden units and optimizes the pretraining task with continual pretraining [30], three Transformer layers cannot extract semantic knowledge well. The $F1$-score of ERNIE-tiny is the lowest among all the pretraining models.

Secondly, for pretraining models with the same model structure, RoBERTa obtains the highest $F1$-score. From the perspective of the pretraining task, RoBERTa removes the sentence-level pretraining task because Liu et al. [23] found that removing the NSP loss in BERT can slightly improve the performance of downstream tasks. For the NER in Chinese medical literature, the pretraining models do not need to learn sentence-level semantic knowledge during pretraining, because the inputs are all individual sentences, not sentence pairs. The NSP and Dialogue Language Model (DLM) of BERT and ERNIE are designed to improve the performance of specific downstream tasks, such as SQuAD 1.1, which requires reasoning about the relationship between sentence pairs. Moreover, as mentioned before, RoBERTa can acquire richer semantic representations with a dynamic masking strategy [23]. In contrast, BERT and ERNIE use static masking strategy in every pretraining epoch. Therefore, their performance is slightly lower than that of RoBERTa.

Finally, different pretraining corpora will affect the performance of NER in Chinese medical literature for pretraining models with the same pretraining tasks and the same model structures, such as BERT-WWM and BERT-WWM-EXT. The pretraining corpus of BERT-WWM is the Chinese Wikipedia, while the pretraining corpus of BERT-WWM-EXT includes not only the Chinese Wikipedia but also News and Q&A [20]. The training dataset we used contains formal scientific literature, and the pretraining corpus of BERT-WWM is closer to it. The results in Table 5 demonstrate that the $F1$-score of BERT-WWM is slightly higher than that of BERT-WWM-EXT.

5.4. Comparison of Feature Extraction and Fine-Tuning Approaches. As shown in Table 6, the $F1$-score can be slightly increased through fine-tuning. This phenomenon may indicate that the pretraining models can obtain semantic knowledge from the target domain during fine-tuning. In other words, the representations outputs from the pretraining models are not adapted to the specific NER task well when the pretraining models are only used as a feature extractor, because the task-specific representations cannot be obtained in this case. Thus, general-purpose representations can be obtained through fine-tuning. However, considering that the improvement is not significant and the feature extraction is computationally cheaper than fine-tuning, the transfer method should be selected in light of specific conditions in practice.

5.5. Comparison of Different Downstream Model Structures. According to the results in Table 7, RoBERTa-CRF obtained the SOTA results. For the NER task, there are strong dependencies across labels. For example, the I-Drug label must follow the B-Drug label. As a probability model, the CRF can output the predicted sequence according to the above rules.

Therefore, the performance of RoBERTa-CRF is better than that of RoBERTa-FC with only one FC layer.

The experimental results in Table 7 also demonstrate that adding the LSTM after RoBERTa does not improve the performance of recognition. The reason is that, on the one hand, the multiheaded self-attention network in the pretraining model has extracted the abstract representations of input tokens well. Therefore, it is not necessary to add the LSTM to extract more abstract representations. On the other hand, a more complex network structure may cause overfitting, which will reduce the performance of recognition.

6. Conclusion

In this paper, we utilize the pretraining models to recognize the named entity in Chinese medical literature, which is the key step in building the medical Knowledge Graph. First of all, we propose a method of data augmentation based on the MLM of ERNIE. A portion of characters and phrases are replaced by synonyms except for the named entities in light of the fact that the named entities may be proper nouns or rare words in the field of medicine. Moreover, we consider NER as a downstream task of the pretraining models and transfer the prior semantic knowledge obtained during pretraining to it.

The results of experiments demonstrate that not only can the data augmentation method we proposed improve the performance of recognition, but also using pretraining models can obtain a meaningful improvement compared with the common deep learning models. Furthermore, for NER in Chinese medical literature, the $F1$-score can be slightly increased through fine-tuning, and using a more complex downstream model will reduce the performance of recognition. For the future work, we will attempt to carry out experiments with a dataset labelled by ourselves and conduct Relation Extraction based on the entities recognized in Chinese medical literature.

Data Availability

The dataset we used is "A Labelled Chinese Dataset for Diabetes," which can be downloaded from the Tianchi network (https://tianchi.aliyun.com/dataset/dataDetail?dataId=22288).

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors would like to thank the AIStudio platform for providing free computing resources. This work was partially supported by the major special project of Anhui Science and Technology Department under Grant 18030801133 and Science and Technology Service Network Initiative under Grant KFJ-STS-ZDTP-079.

References

[1] D. Campos, S. Matos, and J. L. Oliveira, "Biomedical named entity recognition: a survey of machine-learning tools," *Theory and Applications for Advanced Text Mining*, vol. 11, pp. 175–195, 2012.

[2] J. Du and X. Li, "A knowledge graph of combined drug therapies using semantic predications from biomedical literature: algorithm development," *JMIR Medical Informatics*, vol. 8, no. 4, 2020.

[3] N. Boudjellal, H. Zhang, A. Khan, and A. Ahmad, "Biomedical relation extraction using distant supervision," *Scientific Programming*, vol. 2020, no. 9, Article ID 8893749, 2020.

[4] C. Friedman, P. O. Alderson, J. H. M. Austin, J. J. Cimino, and S. B. Johnson, "A general natural-language text processor for clinical radiology," *Journal of the American Medical Informatics Association*, vol. 1, no. 2, pp. 161–174, 1994.

[5] R. Gaizauskas, G. Demetriou, and K. Humphreys, "Term recognition and classification in biological science journal articles," in *Proceedings of the Computational Terminology for Medical and Biological Applications Workshop of the 2nd International Conference on NLP*, 2000.

[6] G. Zhou and J. Su, "Named entity recognition using an HMM-based chunk tagger," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*, pp. 473–480, Philadelphia, PA, USA, 2002.

[7] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual BERT?" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4996–5001, Florence, Italy, 2019.

[8] K. Hakala and S. Pyysalo, "Biomedical named entity recognition with multilingual BERT," in *Proceedings of the 5th Work-shop on BioNLP Open Shared Tasks*, pp. 56–61, Hong Kong, China, 2019.

[9] X. Wu, S. Lv, L. Zang, J. Han, and S. Hu, "Conditional BERT contextual augmentation," *Lecture Notes in Computer Science*, vol. 11539, pp. 84–95, 2019.

[10] M. Gerner, G. Nenadic, and C. M. Bergman, "Linnaeus: a species name identification system for biomedical literature," *BMC Bioinformatics*, vol. 11, no. 1, 2010.

[11] K. Fukuda, T. Tsunoda, A. Tamura, T. Takagi et al., "Toward information extraction: identifying protein names from biological papers," in *Proceedings of the Pacific Symposium on Biocomputing*, pp. 707–718, Maui, HI, USA, 1998.

[12] L. He, Z. Yang, H. Lin, and Y. Li, "Drug name recognition in biomedical texts: a machine-learning-based method," *Drug Discovery Today*, vol. 19, no. 5, pp. 610–617, 2014.

[13] X. Wang, C. Yang, and R. Guan, "A comparative study for biomedical named entity recognition," *International Journal of Machine Learning and Cybernetics*, vol. 9, no. 3, pp. 373–382, 2018.

[14] Q. Zhu, X. Li, A. Conesa, and C. Pereira, "Gram-CNN: a deep learning approach with local context for named entity recognition in biomedical text," *Bioinformatics*, vol. 34, no. 9, pp. 1547–1554, 2018.

[15] Y. Chen, C. Zhou, T. Li et al., "Named entity recognition from Chinese adverse drug event reports with lexical feature based BiLSTM-CRF and tri-training," *Journal of Biomedical Informatics*, vol. 96, 2019.

[16] X. Chen, C. Ouyang, Y. Liu, and Y. Bu, "Improving the named entity recognition of Chinese electronic medical records by combining domain dictionary and rules," *International Journal of Environmental Research and Public Health*, vol. 17, no. 8, 2020.

[17] M. Zaib, Q. Z. Sheng, and W. Emma Zhang, "A short survey of pre-trained language models for conversational AI-a new age in NLP," in *Proceedings of the Australasian Computer Science Week Multiconference*, pp. 1–4, Melbourne, Australia, 2020.

[18] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "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 (NAACL)*, pp. 4171–4186, Minneapolis, MN, USA, 2019.

[19] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," in *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, pp. 5998–6008, Long Beach, CA, USA, 2017.

[20] "Chinese-BERT-WWM," https://github.com/ymcui/Chinese-BERT-wwm.

[21] Z. Zhang, X. Han, Z. Liu et al., "ERNIE: enhanced language representation with informative entities," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1441–1451, Florence, Italy, 2019.

[22] "ERNIE-tiny," https://github.com/PaddlePaddle/ERNIE.

[23] "RoBERTa," https://github.com/pytorch/fairseq.

[24] J. Wei and K. Zou, "EDA: easy data augmentation techniques for boosting performance on text classification tasks," 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)*, pp. 6382–6388, Hong Kong, China, 2019.

[25] W. Y. Wang and D. Yang, "That's so annoying!!!: a lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviours using # petpeeve tweets," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2557–2563, Lisbon, Portugal, 2015.

[26] Z. Yang, Z. Dai, Y. Yang et al., "Generalized autoregressive pre-training for language understanding," in *Proceedings of the 2019 Advances in Neural Information Processing Systems (NIPS)*, pp. 5754–5764, 2019.

[27] Z. Li and M. Sun, "Punctuation as implicit annotations for Chinese word segmentation," *Computational Linguistics*, vol. 35, no. 4, pp. 505–512, 2009.

[28] M. E. Peters, S. Ruder, and N. A. Smith, "To tune or not to tune? Adapting pre-trained representations to diverse tasks," in *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pp. 7–14, Florence, Italy, 2019.

[29] Alibaba cloud labelled Chinese dataset for diabetes, https://tianchi.aliyun.com/dataset/dataDetail?dataId=22288.

[30] Y. Sun, S. Wang, Y. Li et al., "ERNIE 2.0: a continual pre-training framework for language understanding," *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*, vol. 34, no. 5, pp. 8968–8975, 2020.