Medical Name Entity Recognition Based on Lexical Enhancement and Global Pointer

Pu Zhang, Wentao Liang
School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, P. R. China

***Abstract*—Named entity recognition (NER) in biological sources, also called medical named entity recognition (MNER), attempts to identify and categorize medical terminology in electronic records. Deep neural networks have recently demonstrated substantial effectiveness in MNER. However, Chinese MNER has issues that cannot use lexical information and involve nested entities. To address these problems, we propose a model which can handle both nested and non-nested entities. The model uses a simple lexical enhancement method for merging lexical information into each character's vector representation, and then uses the Global Pointer approach for entity recognition. Furthermore, we retrain a pre-trained model with a Chinese medical corpus to incorporate medical knowledge, resulting in F1 score of 68.13% on the nested dataset CMeEE, 95.56% on the non-nested dataset CCKS2017, 85.89% on CCKS2019, and 92.08% on CCKS2020. These data demonstrate the efficacy of our proposed model.**

Keywords*—MNER; nested NER; Global Pointer; lexical enhancement

I. INTRODUCTION

Electronic medical records are a digital repository of a patient's comprehensive medical and health information that can be used for medical and healthcare services. These are vital repositories of medical knowledge and therapeutic experience, containing detailed content about patients' diagnoses and treatment histories. These enable deeper analysis to extract and consolidate valuable medical knowledge, transform implicit knowledge into explicit knowledge, and build a medical knowledge system with a clear hierarchy, well-defined concepts, rich connotations, and significant practical implications. However, due to the high percentage of unstructured data in electronic medical records, direct exploitation is difficult. MNER is a vital step in the extraction of medical information and is essential for biomedical text mining.

MNER has received a lot of attention in recent years, which has resulted in a lot of assessment contests. The China Conference on Knowledge Graph and Semantic Computing (CCKS) held NER evaluation challenges for medical texts from 2017 to 2021. Meanwhile, AliCloud and the Chinese Information Processing Society of China (CIPSC) have released the Chinese Medical Information Extraction (CMeIE) subtask of the Chinese Biomedical Language Understanding Evaluation (CBLUE).

The previous studies in Chinese MNER have mainly used methods based on English MNER to improve the model performance [1]. Although these methods produce good results, they rely on word-level annotation models. However, unlike English, Chinese is not naturally tokenized and does not segment words by spaces in sentences, leading to additional ambiguities when using word segmentation as an additional step in Chinese MNER, which could result in inaccurate entity boundary detection and class prediction due to improper word segmentation errors [2]. As a solution to this problem, character-based Chinese MNER techniques that are better at avoiding segmentation errors have been proposed. Considering that character-based approaches cannot fully exploit lexical information, Zhang et al. [3] introduced the Lattice-LSTM model, which combines lexical information into a character-based recognition model. However, the complex structure of Lattice-LSTM made it difficult to combine with other models.

On the other hand, in the Chinese MNER, the challenge of identifying nested entities has remained unsolved. Several earlier models are based on sequence models, however not all entities in electronic medical records are self-contained, and there may be nested structures between them. As shown in Fig. 1, in the sentence, where the entity "肺 (lung)" is nested within the other entity "肺内病变 (lung lesions)". Because of the complexity of the nested entity structure and the irregularity in its granularity and number of nested levels, it is difficult to rapidly and accurately gather nested entity information for semantic comprehension, a critical component of improving Chinese MNER.

Fig. 1. Example of nested entities.

To address the aforementioned challenges, our research provides a Chinese MNER model based on lexical enhancement and Global Pointer. The model first adds lexical information to each character in the electronic medical record by using the lexical enhancement method, and then uses the Global Pointer method to score the beginning and end of each character to identify the medical entity. The following are the primary contributions of our work:

1) We incorporate the word lexicons into the character representations by introducing a lexical enhancement approach. The approach matches each character with a dictionary built from a corpus to get word sets. Then the word set is compressed and combined with character representation. The experimental results demonstrate that the model with lexical enhancement can improve performance.

2) The Global Pointer is used as the entity recognition module. Extensive experiments are carried out on the nested dataset CMeEE and the non-nested datasets including CCKS2017, CCKS2018, and CCKS2020. The experimental results show that the Global Pointer model can use a unified method to deal with both nested and non-nested entity recognition problems and has better performance than some recent models.

3) The pre-trained model is retrained with Chinese medical corpus for experimentation. It has better results than the basic pre-trained model.

The final experimental results show that the proposed model can be universally applied to nested and non-nested Chinese MNER tasks, and both yield the best results on the four datasets.

The remaining sections are organized as follows: Section II presents the related work on MNER. Section III introduces our proposed model. Section IV presents experiments and results analysis on our model. Section V concludes this paper.

II. Related Work

The purpose of NER is to discover entities in a text and classify them into specified categories such as person, organization, place, and so on. NER is a critical component of information extraction that allows structured information to be extracted from unstructured text input. Effective NER models are required for a variety of downstream tasks, including entity linking, relation extraction, and event extraction. MNER focuses on recognizing and categorizing clinical terminologies such as symptoms, medicines, and therapies in medical data. The MNER task is often treated as a sequential labelling issue, with the objective of assigning a category to each word or character in the text.

The neural network models are now the recommended strategy for English NER. The Bi-LSTM-CRF [4] model is the most typical, incorporating Bi-directional Long Short-Term Memory (Bi-LSTM) for feature extraction and Conditional Random Fields (CRF) for decoding. Unlike English text, Chinese text does not possess clear boundary information. Hence, Chinese NER methods can be broadly categorized into two groups: word-based methods and character-based methods. In the word-based method, word segmentation is performed first, followed by entity recognition [5]. However, this approach may result in the propagation of errors from inaccurate word segmentation, thus leading to incorrect NER. The character-based method, on the other hand, operates on each word individually and does not have the issue of error propagation, but cannot utilize lexical information. Consequently, researchers in word-based models are striving to improve the use of word information [6]. The majority of current studies demonstrate that character-based methods often surpass word-based methods in Chinese NER, due to the issue of error propagation in word segmentation. So, we build a character-based model for Chinese MNER that incorporates lexical information to address the limitations of traditional character-based methods.

MNER in the Chinese language is primarily researched using deep learning-based approaches. Bi-LSTM-CRF model has been proposed for predicting the sequence labels in a posterior conditional random field. Gridach et al. [7] were the pioneers in using a Bi-LSTM-CRF model for NER in the biomedical domain. Dang et al. [8] fine-tuned the word vectors using linguistic information based on the Bi-LSTM-CRF model. Liu et al. [9] combined a multi-channel convolutional neural network with the Bi-LSTM-CRF model and used lexical and morphological features of words as information for entity recognition.

In addition to the Bi-LSTM-CRF model, there have been several studies that apply other deep learning models to the medical field for MNER. For instance, Qiu et al. [10] utilized a residual dilated convolution model for efficient and quick NER in the medical field. Zhang et al. [11] proposed a hybrid model of Dilated Convolutional Neural Network (DCNN) and Bi-LSTM for hierarchical encoding, taking advantage of DCNN to gain global information with fast computational speed. Du et al. [12] proposed a multi-task learning approach with multi-strategies based on MRC. NER in the medical field can be approached as a sequence labeling task or a span boundary detection task.

However, the methods mentioned above are based on sequence labeling and cannot be directly used to solve the identification problem of nested named entities, because the same lexical entry in a nested named entity may have two or more different labels at the same time.

Solving the identification problem of nested entities will also improve the accuracy of the model in extracting entities. Previously, a combination of rule-based and machine-learning-based approaches was often used to deal with nested named entities. First, the inner non-nested named entities are identified using the Hidden Markov Model (HMM). Then, the other named entities are identified using rule-based post-processing. Alex et al. [13] proposed several CRF-based models for nested named NER on the GENIA dataset. These methods apply CRFs to entity types in a specific order, so that each CRF can use the output of the previous CRFs. This cascading approach can achieve the best results for nested named NER. In 2009, Finkel and Manning [14] implemented the task of nested NER from a parsing perspective. They constructed a selection tree to map all named entities to a node in the tree. Rule based and machine learning based methods have high accuracy and can make rules according to a specific domain to extract nested named entities. However, there are some problems such as difficulty in recognizing the same type of nested named entities, high time complexity, and difficulty in scaling to large datasets with long sentences.

Recently, the research of nested NER has become a hot topic in the field of information extraction. Span-based methods have become increasingly popular due to their high

performance. For example, Xu et al. [15] used a local detection approach, where each possible entity span is classified independently. Sohrab and Miwa [16] introduced a simple deep neural model that enumerates all possible spans and classifies them using LSTM. Wang et al. [17] proposed a transition-based method that builds nested entities incrementally by performing a series of actions designed for this purpose. Tan et al. [18] extended the span-based approach by including a boundary detection task that predicts entity boundaries in addition to classifying spans. Quoc et al. [19] proposed a two-stage entity recognition method to address the limitations of span-based models. Our method is also a span-based approach, but unlike previous studies, our model predicts each character as the beginning or end of each span without enumerating each span. As a result, it is highly efficient.

III. Proposed Model

To address the problem that traditional lexical enhancement methods are complicated and cannot be easily transferred to different deep learning neural network architectures, a simpler approach is used to quickly merge lexical information into each character's vector representation. Meanwhile, the entity recognition module uses a span-based entity recognition approach as the entity recognition module. The Fig. 2 depicts the model's general architecture:

A. Lexical Enhancement

The input sentence $s = \{c_1, c_2, \cdots, c_n\} \in V_c$ is processed as a sequence of characters by the Chinese character-based NER model, where $\mathcal{V}_c$ denotes the character vocabulary. Each character c_i is represented by a word vector:

$$x_i^c = e^c(c_i) \tag{1}$$

where e^c denotes the character embedding lookup table.

Character-based NER techniques have the disadvantage of not fully utilizing the information contained in words. To address this, we use SoftLexicon, a lightweight dictionary matching approach for entity recognition in Chinese electronic medical records. SoftLexicon incorporates lexical information into character representations, addressing the issue of character-based NER models' inability to use word information while simplifying the lexical enhancement process [20]. The lexical enhancement features are constructed in three steps:

First, in order to preserve the lexical information of characters, all matching words for each character C_i are divided into four sets of "BMES," which are constructed as follows:

$$\begin{aligned} B(c_i) &= \{w_{i,k}, \forall w_{i,k} \in L, i < k \le n\} \\ M(c_i) &= \{w_{j,k}, \forall w_{j,k} \in L, 1 \le j < i < k \le n\} \\ E(c_i) &= \{w_{j,i}, \forall w_{j,i} \in L, 1 \le j < i\} \\ S(c_i) &= \{c_i, \exists c_i \in L\} \end{aligned} \tag{2}$$

Where $\boldsymbol{w}_{i,j}$ denotes the subsequence $\{c_i, c_{i+1}, \cdots, c_j\}$ and L is the lexicon we built. We constructed it from the already labelled texts of the train and dev sets and counted the number of their occurrences for the second step. The set $B(c_i)$ represents the word set with the character c_i at the beginning and the length is greater than 1, the set $M(c_i)$ represents the word set with the character c_i in the middle and the length is greater than 1, and the set $E(c_i)$ represents the character c_i at the end and the length is greater than 1, the set $S(c_i)$ represents a single character c_i. If the word set corresponding to the current character is empty, it is set to the special word "NONE". Fig. 3 is an example of the SoftLexicon vocabulary extension. In this way, word embedding can be introduced and there is no loss of information because the tag matching result can be accurately restored by the four word sets.

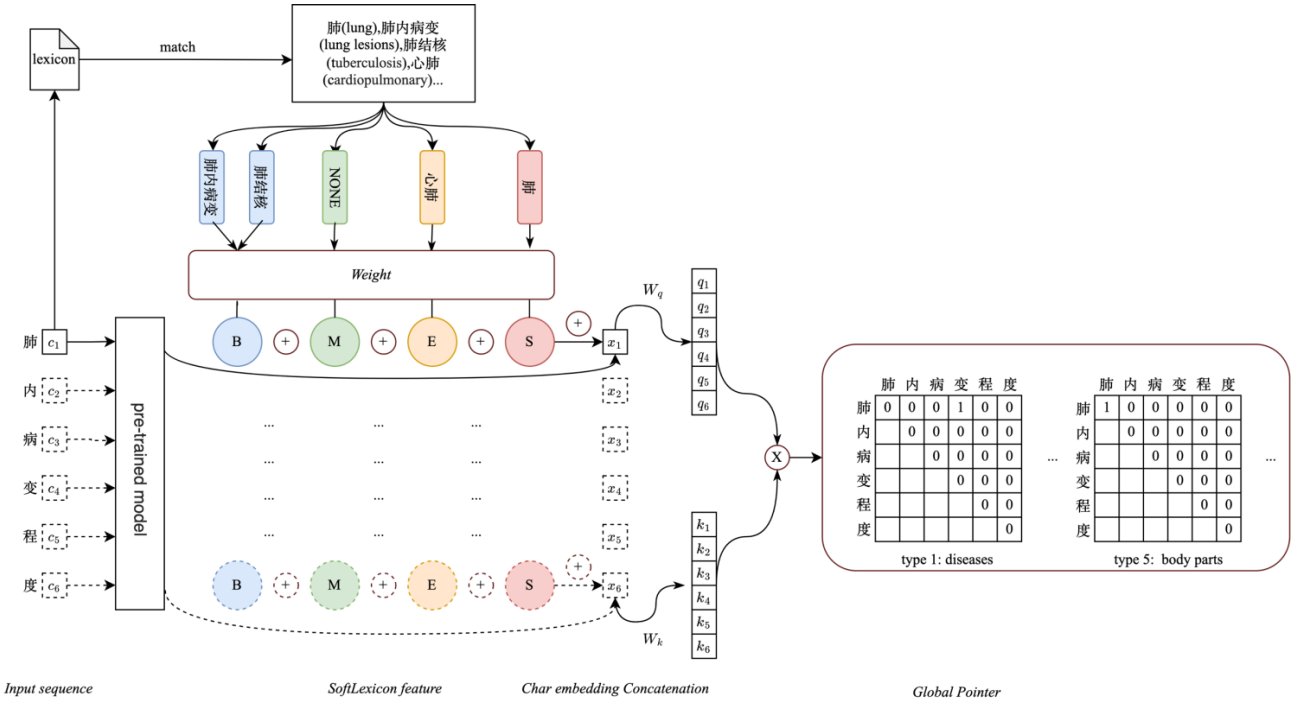

Fig. 2. Architecture of the model.

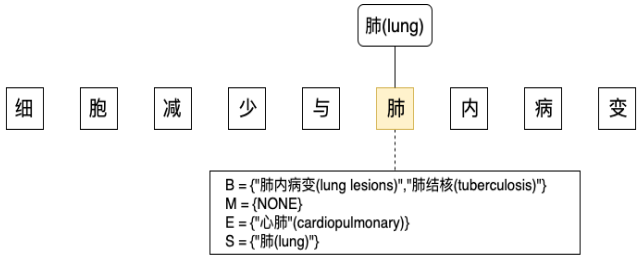

Fig. 3. SoftLexicon method.

In the second step, transform each word in the word set using a pre-trained word vector; then perform weight normalization on all words in the four word sets, using a static weighting method based on statistics, that is, the frequency of each word in the static data. This frequency can reflect the importance of the word to a certain extent, and the frequency is obtained statistically when building a dictionary. The weighted method is as follows:

$$\boldsymbol{v}^s(S) = \frac{4}{Z}\sum_{w\in S} z(w)\boldsymbol{e}^w(w) \tag{3}$$

where S is the "BMES" word sets, $z(w)$ is the frequency of word w in the dictionary in the static statistics, Z is the sum of all words in the word set, and e^w is the lexical embedding lookup table.

Finally, the representations of the four word sets are combined into a one-dimensional feature, which is then stitched onto the representation of that character vector to obtain the final input vector.

$$\boldsymbol{e}^s(\mathrm{B},\mathrm{M},\mathrm{E},\mathrm{S}) = [\boldsymbol{v}^s(\mathrm{B});\boldsymbol{v}^s(\mathrm{M});\boldsymbol{v}^s(\mathrm{E});\boldsymbol{v}^s(\mathrm{S})] \tag{4}$$

$$x^c \leftarrow [x^c;\boldsymbol{e}^s(\mathrm{B},\mathrm{M},\mathrm{E},\mathrm{S})] \tag{5}$$

where v^s is the result of the calculation in the previous step.

B. Global Pointer

In our work, we use the Global Pointer [21] as an entity recognition module, which is a span-based entity recognition method. Span-based methods identify named entities by classifying sub-sequences of sentences. This method outperforms sequence annotation-based methods in preventing error propagation and is able to easily detect nested named entities as they belong to different sub-sequences.

The Global Pointer concept is similar to a simplified multi-head attention mechanism, with as many heads as there are entities. For a sequence of length n, NER is performed by the αth category. The sequence has a total of $\frac{n(n+1)}{2}$ candidate entities, containing all possible entities. The entity recognition task is to select the actual entity from these $\frac{n(n+1)}{2}$ candidate entities. The Global Pointer scores each character to determine whether it is the beginning or the end of an entity. Based on this idea, the lexically enhanced vector token x^{ci} can be transformed into $\boldsymbol{q}_{i,\alpha}$ and $\boldsymbol{k}_{i,\alpha}$:

$$\boldsymbol{q}_{i,\alpha} = \boldsymbol{W}_{q,\alpha}\boldsymbol{x}^{ci} + \boldsymbol{b}_{q,\alpha} \tag{6}$$

$$\boldsymbol{k}_{i,\alpha} = \boldsymbol{W}_{k,\alpha}\boldsymbol{x}^{ci} + \boldsymbol{b}_{k,\alpha} \tag{7}$$

where $\boldsymbol{W}$ is the weight matrix and $\boldsymbol{b}$ is the bias. The $\boldsymbol{q}_{i,\alpha}$ and $\boldsymbol{k}_{i,\alpha}$ are the vector representations of the token which used to identify the entity of type α. Specifically, for span s $[i{:}j]$ of type α, the start and end positions are represented by $\boldsymbol{q}_{i,\alpha}$ and $\boldsymbol{k}_{i,\alpha}$. The score of an entity of type α can then be calculated as follows:

$$s_\alpha(i,j) = \boldsymbol{q}_{i,\alpha}^{\top}\boldsymbol{k}_{j,\alpha} \tag{8}$$

In the inference step, the segments for which the condition $s_\alpha(i,j) > 0$ is satisfied are considered the output entities of type α.

The use of Global Pointers alone is insufficient to accurately identify different types of entities as it does not take into account the length and span of the entities. For example, in the sentence "细胞减少与肺内病变程度密切相关 (cytopenia is closely related to the degree of lung lesions)", "细胞减少与肺 (cytopenia and lung)"may be mistaken as a single entity, as "细(fine)" is the start of the entity "细胞减少 (cytopenia)" while "肺(lung)" is the end of the entity "肺 (lung)". The model treats this combination of start and end positions as one entity. To overcome this issue, it is critical to incorporate relative position information into the Global Pointer method, which is more sensitive to the length and span of entities. Global Pointer uses rotational position encoding to encode relative position information. This encoding is based on a transformation matrix $\boldsymbol{R}_i$ with the property $R_i^{\top}R_j = R_{j-i}$, which can be applied to q and k respectively. The dot product between q and k is then transformed by the relative position information, resulting in a more accurate representation of the entity.

$$\begin{aligned} s_\alpha(i,j) &= \left(R_i\boldsymbol{q}_{i,\alpha}\right)^{\top}\left(R_j\boldsymbol{k}_{j,\alpha}\right) \\ &= \boldsymbol{q}_{i,\alpha}^{\top}R_i^{\top}R_j\boldsymbol{k}_{j,\alpha} \\ &= \boldsymbol{q}_{i,\alpha}^{\top}R_{j-i}\boldsymbol{k}_{j,\alpha} \end{aligned} \tag{9}$$

C. Loss Function

Entity recognition is a multi-label classification problem, so we use a multi-label classification loss function for category imbalance to perform multi-label classification for each label category in the $\frac{n(n+1)}{2}$category after obtaining scores for all segments of each label category. Because the number of entities present in each input text phrase is frequently fewer than $\frac{n(n+1)}{2}$, direct bi-classification causes serious category imbalance problems. Therefore, the method treats the problem as a two-by-two comparison of target and non-target category scores and uses cross-entropy to compute self-balancing weights to avoid the label category imbalance problem, which can be formulated as follows:

$$\mathcal{L} = \log\left(1 + \sum_{(i,j)\in P_\alpha} e^{-s_\alpha(i,j)}\right) + \log\left(1 + \sum_{(i,j)\in Q_\alpha} e^{s_\alpha(i,j)}\right) \tag{10}$$

where i, j represent the start and end indexes of a span, P_α refers to a set of spans that have entity type α, while Q_α refers

to a set of spans that either do not have an entity type or have a different entity type from α. The function $s_\alpha(i,j)$ calculates the score for a span $s[i:j]$ to be an entity of type α.

IV. EXPERIMENTS

A. Experimental Settings and Evaluation Metrics

The PyTorch deep learning framework is used to create the experimental model. For our experiments, we used pre-trained models including BERT-Base-Chinese [22] and RoBERTa-large [23]. To incorporate more medical information, we also used Bertcner [24], which was retrained on a medical corpus based on BERT-Base-Chinese.

Table I lists the hyper-parameter settings used in the model:

TABLE I. THE HYPER-PARAMETER SETTINGS

Parameters	Value
learning rate	2e-5
batch size	16
epochs	10
max sequence length	256
hidden size	768

We calculate the classification metrics including True Positive (TP), False Positive (FP), and False Negative (FN), and evaluate the performance of the model using the metrics of precision (P), recall (R) and Micro-F1($F1$):

$$\text{Πρεχισιον} = \frac{\sum_{i=1}^{n} \text{TP}_i}{\sum_{i=1}^{n} \text{TP}_i + \sum_{i=1}^{n} \text{FP}_i} \quad (11)$$

$$\text{Recall} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FN_i} \quad (12)$$

$$\text{F1} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (13)$$

B. Introduction of Nested Dataset

The Tianchi Chinese Biomedical Language Understanding Evaluation Benchmark, jointly provided by Peking University, Zhengzhou University, Pengcheng Laboratory, and Harbin Institute of Technology (Shenzhen), published the CMeEE dataset, with a total of 938 files annotated with 47,194 sentences divided into nine major medical entity categories: diseases (dis), clinical (sym), and so on. We conducted experiments and analyzed them on this dataset to validate the effectiveness of our model. The data sources of the dataset include clinical trials, electronic medical records, medical books and search logs from real-world search engines. As biomedical data may contain privacy information such as patients' names, ages and genders, all collected data are anonymous and reviewed by ethics committees to protect privacy.

The CMeEE dataset differs from traditional NER in that there is a nested relationship between entities, which is a common phenomenon in medical texts and makes the model processing more complex. The nested entity instances accounted for about 11% of the training set. The dataset contains 15,000 sentences in the training set, 5,000 sentences in the validation set, and 3,000 sentences in the test set. To protect privacy, all data were anonymized and reviewed by an ethics committee.

C. Baselines of Nested Dataset

The comparison models used are the benchmark models officially released by CBLUE, and these methods are implemented based on the pre-trained model, including:

1) BERT-Base-Chinese: The basic model used had 12 layers, 768 hidden layers, 12 headers, and 110 million parameters.

2) RoBERTa-large: RoBERTa removed the next sentence prediction target and dynamically changed the masking pattern applied to the training data.

3) Bertcner: A medical pre-training model was obtained by crawling a 1.05G clinical text consisting of different medical domain corpora obtained on the web to train the BERT model again.

4) ZEN [25]: A BERT-based Chinese text encoder was enhanced by n-gram representation, considering different character combinations in training.

5) Mac-BERT-Base/large [26]: The Mac-BERT was an upgraded BERT that included a new Masked Language Model (MLM) as a correction pre-training assignment, which reduced the differences between pre-training and fine-tuning.

6) PCL-MedBERT: A pre-trained medical language model was proposed by the Intelligent Medicine Research Group in the PengCheng Lab, which excelled in medical question matching and NER.

In addition to the officially provided baselines, the following experiments are used for comparison:

7) TPLinker: Wang et al. [27] developed a single-stage joint extraction approach for addressing entity relationship extraction designs that could find overlapping relationships between one or two entities while being unaffected by exposure bias. The approach utilized to identify entity pieces was chosen as a comparison method in this paper.

8) Muti-head: Li et al. [28] developed a training strategy based on fragment annotation to solve the lack of entity data annotation. The essential idea was to employ negative sampling, which prevented NER models from being trained on unlabeled items.

9) Biaffine: Yu et al. [29] introduced a new NER approach that treats NER as a problem of dependent syntactic analysis, using graph neural networks to model the global information of the input sequence.

D. Results and Analysis of Nested Dataset

As shown in Table II, the F1 scores for entity recognition using only pre-trained models including BERT-Base-Chinese, RoBERTa-large, Bertcner, ZEN, Mac-BERT-Base/large, and PCL-MedBERT range from 60.7% to 62.8%. Compared to the human score of 67%, there is a gap of at least 4.2%. One

important reason for this gap is the presence of many nested entities in this dataset, which cannot be recognized by these basic models based on sequence modeling. Therefore, our model uses a span-based model called Global Pointer as the entity recognition module. This model can calculate whether each character in the input text can be the beginning or end of an entity. It can recognize small entities nested within larger ones without a limit on the nesting level. We conducted three sets of experiments to validate this approach. Specifically, GP+BERT-Base-Chinese improved performance by 4.66% as compared to utilizing only BERT-Base-Chinese. When compared to the model utilizing only Bertcner, using GP+Bertcner improved the score by 4.58%, while using Global Pointer with RoBERTa-large, the GP+RoBERTa-large model improved the score by 5.99%. Additionally, our model GP+ BERT-Base-Chinese outperformed two other span-based models, Muti-head and Biaffine, by 1.93% and 3.91%, respectively. The above results demonstrate the effectiveness of Global Pointer. This also demonstrates that if nested entities can be recognized, it can significantly improve the performance of model and make it more suitable for MNER.

TABLE II. RESULTS OF CMEEE DATASET

Methods	F1(%)
BERT-Base-Chinese	62.1
RoBERTa-large	62.1
Bertcner	62.8
ZEN	61.0
Mac-BERT-base	60.7
Mac-BERT-large	62.4
PCL-MedBERT	60.6
TPLinker	64.32
Muti-head	64.83
Biaffine	62.85
GP[a]+ BERT-Base-Chinese	66.76
GP[a]+BERT-Base-Chinese+ SoftLexicon	67.43
GP[a]+Bertcner	67.38
GP[a]+Bertcner+SoftLexicon	67.71
GP[a]+RoBERTa-large	68.09
GP[a]+RoBERTa-large+SoftLexicon	**68.13**
Human	67.0

[a] "GP" means Global Pointer

To verify that a character-based model with added lexical information would perform better, we conducted experiments using SoftLexicon as a vocabulary enhancement method. Compared to experiments using only Global Pointer, we found a generally improved performance, with an increase of approximately 0.1% to 0.67%. This demonstrates that, under the same experimental conditions, adding lexical information into word embeddings has a certain gain effect on entity recognition.

We also conducted comparative experiments using Bertcner, which was trained on medical corpora based on the BERT-base model. When using pre-trained models directly for experiments, Bertcner outperformed BERT-Base-Chinese by 0.7%. With the addition of Global Pointer, GP+Bertcner outperformed GP+BERT-Base-Chinese by 0.62%. When SoftLexicon was added, GP+Bertcner+SoftLexicon still resulted in a 0.28% improvement over GP+BERT-Base-Chinese+ SoftLexicon. These results demonstrate that using BERT models trained on medical corpora can further improve the recognition of medical entities.

E. Introduction of Non-nested Datasets

Three non-nested datasets provided by the CCKS competition were used in the experiments. the CCKS-2017 dataset has 300 electronic clinical records and 29,866 labeled entities, categorized into five entity types: treatments, signs and symptoms, diseases and diagnoses, examinations and tests, and body parts. the CCKS-2019 dataset has 23,401 labeled entities, annotated into six entity types: diseases and diagnoses, examinations, tests, procedures, medications, and anatomical sites. The CCKS-2020 dataset has 32,120 labeled entities, with the same entity type classification as CCKS2019.

F. Baselines of Non-nested Datasets

The comparative models are presented below:

1) RoBERTa-Bi-LSTM-CRF[30]: The model contains three layers, a character embedding layer, a Bi-LSTM layer, and a CRF layer, relying on character-based word representations learned from a supervised corpus. The Bi-LSTM-CRF model can improve medical named entity recognition by capturing contextual information using bi-directional LSTM, and by modeling dependencies between tags using CRF.

2) RoBERTa-Bi-GRU-CRF[31]: The neural network model integrated Bi-GRU and CRF for sequence labeling tasks. Bi-GRU is a gated recurrent unit, an improved RNN that can solve the gradient vanishing and long-term dependency problems. By feeding the Bi-GRU output into CRF, the Bi-GRU-CRF network may use both bi-directional context information and label constraints for sequence tagging at the same time.

3) Bertcner[24]: A medical pre-training model was obtained by crawling a 1.05G clinical text consisting of different medical domain corpora obtained on the web to train the BERT model again.

4) Ra-RC [32]: The model used RoBERTa as an encoder to capture contextual information and adds part-level features on top of it to enhance the understanding of Chinese language.

5) AR-CCNER [33]: The model used part-level characteristics to augment character semantic information and a self-attention technique to record character interdependence.

6) ACNN [34]: This method effectively learned global context information using an attention mechanism and multi-layer CNNs and captured both short-term and long-term contextual information.

7) BE-Bi-CRF-JN [35]: This method combined the original text in NER tasks with its medical encyclopedia knowledge by establishing connections and interactions to enhance the ability of entity recognition.

8) RGT-CRF [36]: This model used two sets of features, word-based and word-based features, to make full use of the characteristics of Chinese language. The model also used a rule generator to automatically construct rules to improve the generalization ability of the model.

G. Results and Analysis of Non-Nested Datasets

As shown in Tables III, IV and V, we conduct experiments on the CCKS2017, CCKS2019, and CCKS2020 datasets to evaluate our model's performance on non-nested datasets. The experimental results indicate that using Global Pointer as the entity recognition module has a significant improvement compared to using only pre-trained models. Compared with BERT-Base-Chinese, RoBERTa-large, and Bertcner, using the Global Pointer method shows an improvement of 2.37% to 5.35% on the three datasets.

TABLE III. RESULTS OF CCKS2017 DATASET

Methods	P(%)	R(%)	F1(%)
BERT-Base-Chinese	90.92	90.09	90.50
RoBERTa-large	91.99	93.01	92.49
Bertcner	91.10	90.44	90.76
RoBERTa-Bi-GRU-CRF	92.56	93.09	92.82
RoBERTa-Bi-LSTM-CRF	92.41	94.11	93.25
Ra-RC	94.14	92.39	93.26
ACNN	90.19	90.78	90.49
AR-CCNER	92.27	93.73	93.00
RGT-CRF	95.47	95.76	95.61
GP[b]+BERT-Base-Chinese	93.68	95.77	94.71
GP[b]+RoBERTa-large	94.34	95.40	94.86
GP[b]+Bertcner	94.07	95.62	94.83
GP[b]+BERT-Base-Chinese+SoftLexicon	94.51	95.64	95.06
GP[b]+RoBERTa-large+SoftLexicon	95.05	95.05	**95.56**
GP[b]+Bertcner+SoftLexicon	94.93	95.71	95.32

[b.] "GP" means Global Pointer

TABLE IV. RESULTS OF CCKS2019 DATASET

Methods	P(%)	R(%)	F1(%)
BERT-Base-Chinese	79.94	74.63	77.19
RoBERTa-large	79.85	79.85	79.85
Bertcner	81.29	79.43	80.35
RoBERTa-Bi-GRU-CRF	72.95	79.78	76.21
RoBERTa-Bi-LSTM-CRF	79.7	80.75	80.22
Ra-RC	83.31	82.44	82.87
ACNN	83.07	87.29	85.13
BE-Bi-CRF-JN	83.16	86.67	84.88
RGT-CRF	85.36	84.99	85.17
GP[c]+BERT-Base-Chinese	81.87	81.44	81.49
GP[c]+RoBERTa-large	83.04	87.61	85.15
GP[c]+Bertcner	84.04	87.64	85.70
GP[c]+BERT-Base-Chinese +SoftLexicon	83.35	87.72	85.40
GP[c]+RoBERTa-large+SoftLexicon	84.72	87.33	**85.89**
GP[c]+Bertcner+SoftLexicon	82.60	88.83	85.52

[c.] "GP" means Global Pointer

TABLE V. RESULTS OF CCKS2020 DATASET

Methods	P(%)	R(%)	F1(%)
BERT-Base-Chinese	87.78	88.24	88.01
RoBERTa-large	86.68	88.2	87.43
Bertcner	88.57	88.69	88.63
RoBERTa-Bi-GRU-CRF	76.74	88.09	82.02
RoBERTa-Bi-LSTM-CRF	87.05	87.67	87.35
BE-Bi-CRF-JN	82.52	85.05	83.76
RGT-CRF	90.85	91.57	91.2
GP[d]+BERT-base-Chinese	90.20	92.31	91.15
GP[d]+ RoBERTa-large	90.76	91.71	91.09
GP[d]+Bertcner	90.13	92.61	91.27
GP[d]+BERT-Base-Chinese+SoftLexicon	90.30	92.89	91.44
GP[d]+RoBERTa-large+SoftLexicon	90.76	93.61	**92.08**
GP[d]+Bertcner+SoftLexicon	89.74	93.72	91.60

[d.] "GP" means Global Pointer

Compared with the mainstream model using BiLSTM-CRF, our GP+RoBERTa-large model has gained 1.61%, 4.93%, and 3.74% F1 performance improvement over RoBerta-Bi-LSTM-CRF on the CCKS2017, CCKS2019, and CCKS2020 datasets.

In addition, we have compared our model with some other methods, such as the Ra-RC model that uses bi-directional long short-term memory networks to learn radical features of Chinese characters, the AR-CCNER model that uses convolutional neural networks to extract aggressive features while using self-attention mechanism to capture dependencies between characters, and the ACNN model that uses a multi-layer CNN structure to capture short-term and long-term contextual relationships for experiments, our model still outperformed these models in terms of F1 performance.

To verify the effect after adding vocabulary information, SoftLexicon was added separately to introduce vocabulary information for experimentation. Compared with the GP model without using vocabulary information, there is an improvement of about 1%. We also compared our model with other models combining lattice method with Chinese character information such as RGT-CRF. The performance is comparable on CCKS2017 dataset and improved by 0.72% on CCKS2019 and by 0.88% on CCKS2020 in terms of F1.

Finally, we experimented incorporating medical information into MNER model by using Bertcner trained on medical corpus data to obtain semantic features. Compared with basic pre-training model, GP+Bertcner+SoftLexicon has an improvement of 0.1%-0.5% improvement on the F1 over GP+BERT-Base-Chinese+SoftLexicon. When comparing it with BE-Bi-CRF-JN, there are improvements of about 1% and 8% on CCKS2018 and CCKS2020 datasets respectively.

H. Limitation Analysis

Compared with other methods, our model achieved the best results. However, there are still many entities that have not been recognized. In order to promote future work on MNER, this section analyzes the reasons for identification errors. The causes of errors are broadly classified into the following two cases:

1) Ambiguity: Some entities may have different meanings or belong to different categories in different contexts, resulting in difficulties and inaccuracies in entity recognition. To solve this problem, contextual information should be used to determine the true meaning or category of the entity.

2) Inadequate medical knowledge: Because the medical field contains a large number of terms and complex concepts, identifying and classifying these entities can be difficult for non-professionals. For example, in order to correctly classify entities such as diseases, drugs, symptoms, and treatment plans, it is necessary to have a thorough understanding of these concepts. Furthermore, medical knowledge evolves quickly, with new research findings and treatment methods constantly emerging. In this case, relying solely on pre-trained models may not meet real-time demands; collaboration with domain experts can ensure timely model updates, allowing for the resolution of new challenges.

V. Conclusion

In this paper, we propose a model which uses the Global Pointer with a lexical enhancement method and demonstrate its effectiveness for Chinese MNER on nested and non-nested datasets. By using lexical enhancement to incorporate word lexicons into the character representations, our model can perform Chinese NER at the character level and avoid the word segmentation errors. By using Global Pointer, our model can recognize both nested and non-nested entities by enabling a global view that takes the beginning and end locations into account. Experiment results, conducted on the CMeEE, CCKS2017, CCKS2019, and CCKS2020 datasets, show that the proposed model has excellent performance on these four data sets. Our results establish a new benchmark for Chinese MNER and open avenues for further research and exploration. In future work, we will investigate the way to leverage unlabeled data and extend our work to more datasets.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 62136002 and 61876027.

References

[1] Z. Jin, X. He, X. Wu and X. Zhao, “A hybrid transformer approach for Chinese ner with features augmentation,” Expert Systems with Applications, vol. 209, p. 118385, 2022.

[2] R. Ding, P. Xie, X. Zhang, W. Lu and L. Li, “A neural multi-digraph model for Chinese NER with gazetteers,” in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy, 2019.

[3] Y. Zhang and J. Yang, “Chinese NER using Lattice LSTM,” in the 56th Annual Meeting of the Association for Computational Linguistics (ACL), Melbourne, Australia, 2018.

[4] R. Chalapathy, E. Z. Borzeshi and M. Piccardi, “Bidirectional LSTM-CRF for clinical concept extraction,” in Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP), Osaka, Japan, 2016.

[5] H. N. Zhang, D. Y. Wu, Y. Liu and X. Cheng, “Chinese named entity recognition based on deep neural network,” Journal of Chinese Information Processing, vol. 31, no. 4, pp. 28-35, 2017.

[6] Y. Li, L. Zou, W. Liu, W. Liu, and X. Wang, "Research on chinese clinical named entity recognition: lattice lstm with contextualized character representations," JMIR Med Inform, 2020, 8(9): e19848.

[7] M. Gridach, "Character-level neural network for biomedical named entity recognition," Journal of biomedical informatics, 2017, 70: 85-91.

[8] T. H. Dang, H. Q. Le, T. M. Nguyen, and S. T. Vu, "D3NER: biomedical named entity recognition using CRF-biLSTM improved with fine-tuned embeddings of various linguistic information," Bioinformatics, 2018, 34(20): 3539-3546.

[9] J. Liu, S. Chen, Z. He, and H. Chen, "Learning BLSTM-CRF with multi-channel attribute embedding for medical information extraction," Natural Language Processing and Chinese Computing: 7th CCF International Conference, NLPCC 2018, Hohhot, China, August 26–30, 2018, Proceedings, Part I 7. Springer International Publishing, 2018: 196-208.

[10] J. Qiu, Q. Wang, Y. Zhou, T. Ruan, and J. Gao, "Fast and accurate recognition of Chinese clinical named entities with residual dilated convolutions," 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2018: 935-942.

[11] R. Zhang, P. Zhao, W. Guo, R. Wang, and W. Lu, "Medical named entity recognition based on dilated convolutional neural network," Cognitive Robotics, 2022, 2: 13-20.

[12] X. Du, Y. Jia, and H. Zan, "MRC-based medical NER with multi-task learning and multi-strategies," China National Conference on Chinese Computational Linguistics. Springer, Cham, 2022: 149-162.

[13] B. Alex, B. Haddow, and C. Grover, "Recognising nested named entities in biomedical text," Biological, translational, and clinical language processing. 2007: 65-72.

[14] J. R. Finkel, and C. D. Manning, "Nested named entity recognition," Proceedings of the 2009 conference on empirical methods in natural language processing. 2009: 141-150.

[15] M. Xu, H. Jiang, and S. Watcharawittayakul, "A local detection approach for named entity recognition and mention detection," Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017: 1237-1247.

[16] M. G. Sohrab, and M. Miwa, "Deep exhaustive model for nested named entity recognition," Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 2843-2849.

[17] B. Wang, W. Lu, Y. Wang, and H. Jin, "A neural transition-based model for nested mention recognition," Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 1011-1017.

[18] C. Tan, W. Qiu, M. Chen, R. Wang, and F. Huang, "Boundary enhanced neural span classification for nested named entity recognition," Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(05): 9016-9023.

[19] Q. C. Quoc, and V. N. Van, "NER-VLSP 2021: two stage model for nested named entity recognition," VNU Journal of Science: Computer Science and Communication Engineering, 2022, 38(1).

[20] R. Ma , M. Peng , Q. Zhang , and X. Huang, " Simplify the usage of lexicon in chinese NER," Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 5951-5960.

[21] J. Su, A. Murtadha, S. Pan, J. Hou, J. Sun, and W. Huang, "Global Pointer: novel efficient span-based approach for named entity recognition," arXiv preprint arXiv:2208.03054, 2022.

[22] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[23] Y. Liu, M. Ott, J. Du, M. Joshi, D. Chen, and V, Stoyanov, "Roberta: A robustly optimized bert pretraining approach," arXiv preprint arXiv:1907.11692, 2019.

[24] X Li, H Zhang, and X H Zhou, "Chinese clinical named entity recognition with variant neural structures based on BERT methods," Journal of biomedical informatics, 2020, 107: 103422.

[25] S. Diao, J. Bai, Y. Song, T. Zhang, and Y Wang, "ZEN: pre-training Chinese text encoder enhanced by n-gram representations," Findings of the Association for Computational Linguistics: EMNLP 2020. 2020: 4729-4740.

[26] Y. Cui, W. Che, T. Liu, B. Qin, S. Wang, and G. Hu, "Revisiting pre-trained models for Chinese natural language processing," Findings of the Association for Computational Linguistics: EMNLP 2020. 2020: 657-668.

[27] Y. Wang, B. Yu, Y. Zhang, T. Liu, H. Zhu, and L. Sun, "TPLinker: single-stage joint extraction of entities and relations through token pair linking," Proceedings of the 28th International Conference on Computational Linguistics. 2020: 1572-1582.

[28] Y. Li, L. Liu, and S. Shi, "Empirical analysis of unlabeled entity problem in named entity recognition," arXiv preprint arXiv:2012.05426, 2020.

[29] J. Yu, B. Bohnet, and M. Poesio, "Named entity recognition as dependency parsing," Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 6470-6476.

[30] K. Xu, Z. Zhou, T Hao, and W. Liu, "A bidirectional LSTM and conditional random fields approach to medical named entity recognition," Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2017. Springer International Publishing, 2018: 355-365.

[31] Q. Qin, S. Zhao, C. Liu, "A BERT-BiGRU-CRF model for entity recognition of Chinese electronic medical records," Complexity, 2021, 2021: 1-11.

[32] Y. Wu, J. Huang, C. Xu, H. Zheng, L Zhang, and J. Wan, "Research on named entity recognition of electronic medical records based on roberta and radical-level feature," Wireless Communications and Mobile Computing, 2021, 2021: 1-10.

[33] M. Yin, C. Mou, K. Xiong, J. Ren, "Chinese clinical named entity recognition with radical-level feature and self-attention mechanism," Journal of biomedical informatics, 2019, 98: 103289.

[34] J. Kong, L. Zhang, M. Jiang, and T. Liu, "Incorporating multi-level CNN and attention mechanism for Chinese clinical named entity recognition, " Journal of Biomedical Informatics, 2021, 116: 103737.

[35] Q. Wang, and E. Haihong, "Bi-directional Joint Embedding of Encyclopedic Knowledge and Original Text for Chinese Medical Named Entity Recognition," 2021 2nd International Conference on Electronics, Communications and Information Technology (CECIT). IEEE, 2021: 304-309.

[36] J. Li, R. Liu, C. Chen, S. Zhou, X. Shang, and Y. Wang, "An RG-FLAT-CRF Model for Named Entity Recognition of Chinese Electronic Clinical Records, "Electronics, 2022, 11 (8): 1282.