Research Article

Chinese Medical Entity Recognition Model Based on Character and Word Vector Fusion

Qinghui Zhang ^{1,2} **Lei Hou** ¹ **Pengtao Lv** ¹ **Mengya Zhang** ¹ and **Hongwei Yang**¹

¹Key Laboratory of Grain Information Processing and Control, Henan University of Technology, Ministry of Education, Zhengzhou 450001, China

²Henan Key Laboratory of Grain Photoelectric Detection and Control, Henan University of Technology, Zhengzhou 450001, China

Correspondence should be addressed to Pengtao Lv; pengtaolv@163.com

Received 21 August 2021; Accepted 9 November 2021; Published 24 November 2021

Academic Editor: Jianping Gou

Copyright © 2021 Qinghui Zhang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The medical information carried in electronic medical records has high clinical research value, and medical named entity recognition is the key to extracting valuable information from large-scale medical texts. At present, most of the studies on Chinese medical named entity recognition are based on character vector model or word vector model. Owing to the complexity and specificity of Chinese text, the existing methods may fail to achieve good performance. In this study, we propose a Chinese medical named entity recognition method that fuses character and word vectors. The method expresses Chinese texts as character vectors and word vectors separately and fuses them in the model for features. The proposed model can effectively avoid the problems of missing character vector information and inaccurate word vector partitioning. On the CCKS 2019 dataset for the named entity recognition task of Chinese electronic medical records, the proposed model achieves good performance and can effectively improve the accuracy of Chinese medical named entity recognition compared with other baseline models.

1. Introduction

In recent years, with the significant improvement of informatization, the development of medical and health knowledge Q&A platforms [1], and the widespread application of medical information systems [2], a large number of pieces of biomedical information in the medical field are stored in electronic text in an unstructured form and cannot be fully utilized. For these massive medical information data, it is important to identify and integrate the knowledge entities in them efficiently to build medical knowledge graphs, provide accurate medical knowledge Q&A, and perform medical knowledge reasoning [3].

Named entity recognition (NER) [4] is a vital part of natural language processing (NLP) [5]. Its purpose is to recognize various named entities, for example, names, places, and organizations, from raw text. Extracted entities can pave the way for other NLP tasks, such as relationship extraction [6] and knowledge graph construction [7].

Recently, with the rise of deep learning technology, deep neural networks have been utilized to achieve medical NER and have attracted much research attention. Using named entity recognition technology to recognize entities in the medical text is a fundamental step in transforming medical text into structured data. It also lays the foundation for mining and exploiting the rich knowledge contained in medical texts [8].

At present, Chinese medical entity recognition [9] still faces significant challenges for the following main reasons. First, considering the specificity of medical texts, there is no uniform set of nomenclature for clinical medical language [10]. Different medical personnel record and describe conditions in different ways, the medical field contains more new and uncommon words, and some drugs and symptoms have long and rare names. Second, a public Chinese medical text dataset is lacking due to the difficulty of obtaining and labeling medical texts [11]. Finally, the Chinese language is complex [12]: on the one hand, the Chinese language does

not have spaces as separators as in the English language; on the other hand, the Chinese language structure is complex, and there are numerous nested and omitted statements.

For Chinese medical entity recognition, most of the existing studies are based on character vector-based models [13] or word vector-based models [14]. In previous NER researches, the BiLSTM-CRF [15], which is the abbreviation of Bidirectional Long Short-Term Memory (LSTM) [16] joining with a conditional random field (CRF) layer [17], shows advanced performance and has become a prevalent architecture for various NER tasks. This architecture outperforms traditional methods in that it eliminates the inefficient and complex method of manually designing feature templates and utilizes a recurrent neural network (RNN) [18] to automatically capture text features. However, the entities in Chinese medical texts are relatively long and complex. The character vector-based Chinese medical entity recognition models use characters as the granularity of input, but characters do not contain as much semantic information as words. The word vector-based method segments the Chinese medical text first, but the existing word segmentation tools cannot achieve complete accuracy, which causes error transmission and thus affects the effectiveness of the model. Therefore, the performance of a single character vector or word vector-based model will be poor. This paper proposes a method based on character and word vector fusion for Chinese medical entity recognition. Our contributions are summarized as follows:

- (1) We propose a Chinese named entity recognition method based on character and word fusion that fuses Chinese character vectors and word vectors in the model, and we apply it to Chinese medical entity recognition, achieving good performance.
- (2) We improve the ELMO model in the input model of the character vector and the word vector, respectively, and get the corresponding input vector, respectively.
- (3) We propose the vector fusion method for the word vectors obtained at the input after the calculation of the BiLSTM layer. The fused feature vectors can better represent the text information.

The remainder of this paper is organized as follows. Section 2 discusses related works; Section 3 presents our framework in detail; Section 4 describes our experiments, results, and discussion; Section 5 concludes this paper.

2. Related Work

The research on Chinese medical entity recognition began later than its English counterpart. With the influence of foreign medical entity recognition evaluation conferences, related Chinese organizations have also organized conference evaluations for Chinese medical entity recognition. The more influential one is the China Conference on Knowledge Graph and Semantic Computing (CCKS) [19]. The CCKS evaluation tasks have subtasks related to Chinese medical entity recognition, and the China Health Information

Processing Conference (CHIP) also has evaluation tasks related to medical entity identification. These conferences have greatly promoted the research process of Chinese medical entity recognition [20]. The existing research approaches for Chinese medical entity recognition can be classified into three main methods, namely, rule and dictionary-based methods [21], machine learning-based methods [22], and deep learning-based methods [23].

Rule- and dictionary-based methods were applied in early applications of medical entity recognition [24]. The rule-based method mainly relies on the manual formulation of heuristic rules. However, due to the complexity and diversity of entities in medical texts, it is often impossible to enumerate all the rules, which ultimately leads to poor recognition results [25]. The dictionary-based method mainly uses existing medical dictionaries for recognition, and it matches all the entities in the medical text according to the existing dictionaries [26]. However, there are constantly new words appearing in medical texts, and the dictionary cannot contain all medical entities, which causes the failure to recognize named entities outside the dictionary, again leading to poor recognition results [27].

Machine learning-based named entity recognition is ultimately a classification method [28], but there are two ways to perform it. One is to identify all the named entity boundaries in the medical text first and then classify these entities [29]. The second way to carry out machine learning-based named entity recognition is to use the sequence-labeling method, which aims to find a suitable label for a specific sentence sequence. Conditional random field (CRF) is the most commonly used method, which solves the problem of labeling biases by using a global normalization method [30]. Lee and Lu [31] developed a medical named entity recognition model based on CRF and rule-based text attention rules, and the model has achieved good performance in recognizing medical named entities in the admission records of stroke patients. Lei et al. [4] used a 400-point admission record and a 400-point discharge summary from the Beijing Union Medical Hospital to conduct medical entity recognition research, in which several machine learning methods, that is, support vector machines [32], maximum entropy [33], and conditional random fields, were used to conduct the named entity recognition research on medical text. The recognition result of the discharge summary had an F-value of 90.01%. Machine learning methods exhibit a good performance improvement over the rule and dictionary-based methods, but due to the characteristics of language, the migration of models between different domains can lead to significant performance problems. So, machine learning-based methods still need a large amount of labeled data to train models, which not only requires a significant amount of human involvement but also has much room for improvement in recognition accuracy.

Whether rule- and dictionary-based or machine learning-based methods, they require not only a great deal of manual labor and time but also a great deal of expertise in the medical field. In recent years, deep neural networks have been applied to natural language processing and have achieved quite good results in named entity recognition

tasks [34]. Deep learning methods automatically extract representative features through neural networks, and various neural network architectures have achieved advanced performance in the field of medical entity recognition [35]. Yang et al. [36] combined the BiLSTM network and CRF for medical entity identification in admission records and discharge summaries. Xue et al. [37] introduced the language pretraining model BERT [38] released by Google into joint learning, which greatly improved the feature representation of the shared parameter layer. Yin et al. [39] used a convolutional neural network to extract the information of Chinese character headers and fused the attention mechanism to capture the dependency relationship between characters and finally achieved excellent recognition results.

In summary, medical entity recognition methods based on deep learning are gradually being increasingly used in practical medical entity recognition tasks due to their excellent performance. However, in this paper, we propose a medical entity recognition method that fuses character and word vectors using a deep learning approach. The method considers information of both Chinese characters and word dimensions and enables the model to better represent the information of Chinese text, thus improving the performance of Chinese medical entity recognition models.

3. Model Structure

The structure of the Chinese medical entity recognition model based on character and word vector fusion proposed in this paper is shown in Figure 1. The input layer consists of two main parts: one is the Chinese character vector as the input layer into the model, and the other is the Chinese word vector as the input layer into the model. The model structure connected after the input layer is the BiLSTM layer [40, 41], where the Chinese character and word vectors from the input layer are fed into the BiLSTM layer together, and the output vectors of this layer based on character and word vectors are spliced and operated in the next step. Finally, they are sent to the CRF layer [42] for final entity recognition. The components of each part are described separately as follows.

3.1. Input Layer. The input layer of the Chinese medical entity recognition model based on character and word fusion proposed in this paper consists of two main parts. One is to convert the input text into character vectors; the other is to convert the input text into word vectors. In order to improve the performance of medical entity recognition, we use the pretraining model ELMO [43] to generate the required character and word vectors discussed in this paper.

For the input of the character vector branch, we use a corpus of existing character-level Chinese medical text datasets to generate character vectors corresponding to each character by the ELMO [44] pretraining model. When a sequence $c_1, c_2, \dots, c_n$ containing n characters is input, the character vector corresponding to each character in the sequence $e_{c_1}, e_{c_2}, \dots, e_{c_n}$ is sent as input to BiLSTM for subsequent calculation.

For the input of the word vector branch, we first use the word segmentation tool jieba for word segmentation [45], and then, we use a corpus of existing word-level Chinese medical text datasets to generate word vectors corresponding to each word by the ELMO pretraining model [46]. In order to facilitate subsequent calculations and fusion of the model, the dimensionality of the character vector generated by this method is the same as that of the word vector.

3.2. BiLSTM Layer. The basic idea of BiLSTM is to obtain contextual information from the input sequence through two LSTM networks. BiLSTM applies a forward LSTM network and a reverse LSTM network to each training, and the two LSTM networks are connected to the same output layer [47]. The output of the forward LSTM at moment t is $\vec{h}_t = \text{LSTM}(x_t, \vec{h}_{t-1})$ and that of the reverse LSTM at moment t is $\overleftarrow{h}_t = \text{LSTM}(x_t, \overleftarrow{h}_{t+1})$. Finally, the two hidden layer variables $\vec{h}_t$ and $\overleftarrow{h}_t$ are directly connected as the output of BiLSTM at moment t , given by $h_t = [\vec{h}_t; \overleftarrow{h}_t]$. The BiLSTM network structure is shown in Figure 2.

In this paper, we use two independent BiLSTM structures to compute and process the character vector-based input and word vector-based input, respectively, and after the BiLSTM computation, we obtain two sets of feature vectors, that is, character-based feature vectors and word-based feature vectors, which are sent to the next connection layer for vector stitching and fusion.

3.3. Fusion Layer. The main purpose of the fusion layer is to connect the features extracted from the character-level BiLSTM and word-level BiLSTM and perform feature fusion, so that the final CRF layer can use both Chinese character vector information and word vector information to improve the recognition accuracy when performing the final Chinese medical entity recognition [48, 49].

The model in this paper obtains two different feature vectors after the embedding of the input layer and the feature extraction and operation in the BiLSTM layer. The two feature vectors are recorded as character-based feature vectors h_{c_i} and word-based feature vectors h_{w_i} . In order to explore the best way to fuse character and word vectors, this paper compares the following three fusion methods and finally determines the most effective one for subsequent experiments:

- (a) Direct addition of character and word vectors. Since the input layer of this model controls the generation of input character and word vectors with the same dimension, the dimension of the obtained feature vectors after the operation of the BiLSTM layer is also the same, and the new feature vectors of the fused character and word vectors can be obtained for subsequent calculations by adding the character and word feature vectors directly. The specific calculation process is shown in (1) as follows [50]:

$$h_t = h_{c_i} + h_{w_i}. \quad (1)$$

FIGURE 1: Overall structure of the proposed model.

FIGURE 2: BiLSTM model structure.

In (1), h_{ci} represents the output feature vector of the character vector at position i in a sentence through the BiLSTM layer, and h_{wi} denotes the word vector of this word in this sentence passing through the output feature vector of the BiLSTM layer.

- (b) Direct splicing of character and word vectors. The feature vectors h_{ci} and h_{wi} obtained from the BiLSTM layer of this model are directly spliced to obtain a combined feature vector with twice the dimension of

the original single feature vector. The specific calculation process is shown in (2) as follows:

$$h_i = \text{con}(h_{ci}, h_{wi}). \quad (2)$$

In (2), h_{ci} represents the output feature vector of the character vector at position i in a sentence through the BiLSTM layer, h_{wi} denotes the word vector of this word in this sentence passing through the output

feature vector of the BiLSTM layer, and con represents the splicing operation of the vector.

- (c) Use of the feed-forward neural network to weight the character and word vectors. The feature vectors h_{ci} and h_{wi} obtained from the BiLSTM layer of this model are first directly added together, and then, the results obtained are applied to weighting operations using feed-forward neural networks. The specific operation process is shown in (3) as follows [51]:

$$h_i = w \cdot (h_{ci} + h_{wi}) + b. \quad (3)$$

In (3), h_{ci} represents the output feature vector of the character vector at position i in a sentence through the BiLSTM layer, h_{wi} denotes the word vector of this word in this sentence passing through the output feature vector of the BiLSTM layer, and w and b stand for the parameters that must be learned in the neural network.

After the subsequent experimental results are discussed, the experiments are presented in Section 4.4, and the best results can be obtained by directly splicing the feature vectors, so all the subsequent experiments in this paper are conducted in this way to ensure the best results.

Next, we use an example to describe the specific process of character and word vector fusion. For example, with the same input sentence “左下肺感染” in the branch based on the character vector, after the ELMO pretraining model, we obtain five character vectors expressed as $e_{c1}, e_{c2}, e_{c3}, e_{c4}$, and e_{c5} . These five character vectors are input to the BiLSTM layer for subsequent training to obtain five feature vectors $h_{c1}, h_{c2}, h_{c3}, h_{c4}$, and h_{c5} . In the branch based on the word vector, the input text is segmented first, the results of word segmentation are “左下肺” and “感染,” and the two word vectors are expressed as e_{w1} and e_{w2} , respectively. After the BiLSTM layer, we also obtain the two corresponding feature vectors h_{w1} and h_{w2} .

Next, the two sets of feature vectors obtained are spliced to obtain the final feature vector. After the experiments, we use direct splicing to splice the feature vectors obtained based on character and word vectors. The calculation process is shown in (4) as follows:

$$\begin{aligned} h_1 &= \text{con}(h_{c1}, h_{w1}), \\ h_2 &= \text{con}(h_{c2}, h_{w1}), \\ h_3 &= \text{con}(h_{c3}, h_{w1}), \\ h_4 &= \text{con}(h_{c4}, h_{w2}), \\ h_5 &= \text{con}(h_{c5}, h_{w2}). \end{aligned} \quad (4)$$

In (4), h_{ci} represents the output feature vector of the character vector at position i in a sentence through the BiLSTM layer, h_{wi} denotes the word vector of this word in this sentence passing through the output feature vector of the BiLSTM layer, and con represents the splicing operation of the vector.

After calculation, we obtain the corresponding five final feature vectors h_1, h_2, h_3, h_4 , and h_5 , which are fused with the Chinese character and word information, and the

fused feature vectors are sent to the CRF layer for processing to obtain the final medical entity recognition results.

3.4. CRF Layer. LSTM only considers the long-term dependency information of the sentence but not the dependency relationship between tags [52]. For example, in medical entity labeling, B-disease cannot appear after I-disease. CRF can ensure the validity of tags by learning the adjacent relationship between tags. The CRF layer introduces the transition matrix W as a parameter of the CRF layer, and maximum likelihood estimation is used to train the CRF layer. Thus, for the sentence X , the probability of the model labeling sequence $Y = (y_1, y_2, \dots, y_n)$ is

$$\text{score}(X, Y) = \sum_{i=1}^n P_{i, y_i} + \sum_{i=0}^n W_{y_i, y_{i+1}}, \quad (5)$$

$$P(Y|X) = \frac{\exp(\text{score}(x, y))}{\sum_{y'} \exp(\text{score}(x, y'))}. \quad (6)$$

In (5) and (6), $P_{i,j}$ represents the probability value of the i -th character classified to the j -th label, and $W_{i,j}$ expresses the state transition score of the i -th leading to the j -th label. X stands for the input sentence $(x_1, x_2, \dots, x_n)$, Y stands for the sentence tag sequence $(y_1, y_2, \dots, y_n)$, and $\exp$ stands for the exponential function of the natural constant e .

4. Experiment and Results

4.1. Dataset. The dataset used for Chinese medical entity recognition in this study is from the subtask of CCKS 2019 evaluation task 1 “Medical entity recognition and attribute extraction for Chinese electronic medical records,” and the EMRs were manually annotated by a professional medical team.

The manual annotation of the medical entity recognition dataset is mainly carried out by the professional medical team. After the medical team members have obtained a unified standard, they annotate in detail with the predefined entity categories, the start index, and the end index of each entity, respectively. The sequence numbers of the characters at the start index and end index need to be counted by the annotator. Each annotation text contains two parts: raw text and annotation information. The annotation information consists of several triples, which are formed of entity start index, entity end index, and entity category. Through entity start and end indices, we can extract the entity from annotated texts. Of course, you can also use some annotation tools, which can directly generate character serial numbers, and it will save a part of manpower. The common annotation tools are YEDDA [53] annotation tools and text annotation tools of the Jingdong Zhongzhi Wise open annotation platform. YEDDA annotation tool is a graphical annotation tool based on Python, does not rely on other libraries outside of Python, and supports two efficient annotation methods: shortcut keys and command-line annotation, fast and

efficient. The text annotation tool of Jingdong Zhongzhi Wise open annotation platform can be used directly online without downloading, and the operation method and interface are designed to be very friendly.

The dataset used in this study is divided into six pre-defined medical entity categories, which are disease and diagnosis, imaging examination, laboratory test, operation, drug, and anatomy. The annotated dataset was divided into a training set and a test set. The training set contained 1000 EMRs, and the test set contained 379 EMRs. The statistics for the six types of entities are presented in Table 1.

In this study, we choose the BIO (begin, inside, and output) annotation system to annotate the training and test data, where the specific formats are B-X, I-X, and O. B represents the character at the start position of the medical entity, I represents the character of the remaining part of the medical entity, and O represents the character of the nonmedical entity. X represents the medical entity categories, that is, disease and diagnosis, imaging examination, laboratory test, operation, drug, and anatomy, which are denoted as DISE, IMAG, LABO, OPER, DRUG, and ANAT, respectively. Thus, these six types of medical entities have a total of 13 different labels in this task, and various types of medical entity labeling symbols and examples are presented in Table 2.

Although the CCKS 2019 dataset used in this paper has been manually annotated by a professional medical team, problems inevitably exist, such as inconsistent entity category labeling and incorrect entity start index and end index labeling, and these errors cause error transfer, which affects the effect of named entity recognition. In this paper, we use manual error correction in the preprocessing stage of the data for the mislabeled entities. The entities found to be incorrectly annotated are manually corrected, thereby improving the effectiveness of entity recognition.

4.2. Evaluation Criteria. We use precision (P), recall (R), and F1 value (F1) to evaluate the performance of the model proposed in this paper. The calculations are as follows:

$$P = \frac{TP}{TP + FP}, \quad (7)$$

$$R = \frac{TP}{TP + FN}, \quad (8)$$

$$F1 = \frac{2 \times P \times R}{P + R}. \quad (9)$$

In equations (7)–(9), TP indicates the number of named entities correctly recognized (correctly identified), FP indicates the number of named entities incorrectly recognized (incorrectly identified), and FN indicates the number of named entities not recognized (incorrectly identified).

4.3. Parameter Setting. The main focus of this study is the impact of the method based on character and word vector fusion on the performance of Chinese medical entity recognition models. During the experiments, it was found that

changes in certain parameters also had a significant effect on the experimental model results. Therefore, the experiments in this paper first explored the effects of some parameters of the model on the experimental results, and then, we explored the research topic of this paper by comparing the experimental results of different models. In this section, we empirically use direct splicing of character and word vectors for all parameter tuning experiments.

The number of epochs in a deep neural network requires manual adjustment of the parameters, and different tasks often have different settings for the number of epochs. Among them, too few epochs can cause the model to fail to converge to local minima and appear to be underfitting. On the contrary, if the number of epochs is too large, it will increase the training time of the model and also cause the phenomenon of model overfitting; thus, the generalization ability of the model will be lost. In order to explore the influence of the number of epochs on model performance, the graph of the evaluation index of the proposed model with the number of iterations is shown in Figure 3. As can be seen from the figure, the model in this paper undergoes fast convergence before seven iterations, after which the performance of the model improves slowly with the increase in the number of iterations. After approximately 15 iterations, the model performance changes in a small range with the number of iterations. So, the final number of iterations selected in this paper is 25.

The learning rate is also an important parameter that affects the performance of a deep learning model. A large learning rate leads to a nonconverging model, while a small learning rate leads to a particularly slow convergence or failure to learn. In order to investigate the effect of learning rate on model performance, the learning rates of 0.0005, 0.0008, 0.001, 0.003, 0.005, and 0.008 were selected empirically for the experiments. The experimental results are presented in Table 3, from which it can be observed that the model performance is the best when the learning rate is equal to 0.005. So, the learning rate selected in this paper is 0.005.

After continuous experiments and a comprehensive evaluation of the results, the best parameter values for the experiments presented in this paper were finally determined. The parameters of the character vector-based recognition model and word vector-based recognition model are set to the same values in the experiment, and the specific parameter settings are presented in Table 4.

4.4. Fusion Strategy Test. In order to explore the effect of fusion strategies on the experimental results, this section describes experiments on the three character and word fusion strategies mentioned above separately, after which the best one for subsequent experiments was chosen. The three fusion strategies that must be experimented on are direct addition of character and word vectors (denoted as model-add), direct splicing of character and word vectors (denoted as model-con), and the use of the feed-forward neural network to weight the character and word vectors (denoted as model-wb). The experimental results are presented in

TABLE 1: Entity category of dataset.

	Text	Disease and diagnosis	Imaging examination	Laboratory test	Operation	Drug	Anatomy	Overall
Training data	1000	2116	222	318	765	456	1486	5363
Test data	379	682	91	193	140	263	447	1816

TABLE 2: Various types of medical entity labeling symbols and examples.

Medical entity	Labeling symbol	Labeling example
Disease and diagnosis	B-DISE/I-DISE	肺 B-DISE 癌 I-DISE
Imaging examination	B-IMAG/I-IMAG	M B-IMAG I I-IMAG
Laboratory test	B-LABO/I-LABO	尿 B-LABO 培 I-LABO 养 I-LABO
Operation	B-OPER/I-OPER	胃 B-OPER 癌 I-OPER 根 I-OPER 治 I-OPER 术 I-OPER
Drug	B-DRUG/I-DRUG	替 B-DRUG 加 I-DRUG 氟 I-DRUG
Anatomy	B-ANAT/I-ANAT	肝 B-ANAT 脏 I-ANAT
Nonmedical entities	O	人 O

FIGURE 3: Model performance with the number of epochs.

Table 5. According to the experimental results, the direct splicing of character and word vectors is the most effective among the three fusion strategies. So, this fusion strategy is used in the subsequent experiments.

4.5. Experimental Results. In order to enable the model to achieve better performance, the experimental selection of parameters was carried out with the optimal combination of parameters obtained from the abovementioned experiments, and direct splicing of character and word vector was selected for the experimental fusion strategy. Table 6 presents the precision (P), recall (R), and F1 values (F1) of the six entities after the experiment on the character and word vector fusion-based BiLSTM-CRF model (denoted CW-BiLSTM-CRF) proposed in this paper. From the table, we can observe that the recognition results for different types of entities are different, which is normal, because the number and length of each type of entity are different.

The F1 values on “Imaging examination” and “Drug” are smaller than those on other types of entities. The F1 values on “Imaging examination” are only 79.65%, which may be caused by the small number of “Imaging examination” entities. The poor performance of the “Drug” entity is caused by the small number of entities, and it may be caused by the fact that the name of the entity often contains unusual and proprietary words. “Disease and diagnosis” has the highest number of entity types among the six entities, but its precision, recall, and F1 value are relatively low, probably due to the fact that entities in the “Disease and diagnosis” category are generally long, for example, “慢性肾脏病5期” and “两肺感染性病变.” So, the problem of incorrect boundary prediction arises when predicting these entities, which leads to incorrect entity prediction.

4.6. Comparison with Other Models. In order to evaluate the performance of the model proposed in this article, we conducted further experiments to compare our proposed

TABLE 3: Model performance with different learning rates.

Learning rate	Precision (P)	Recall (R)	F1 value (F1)
0.0005	84.56	84.13	84.34
0.0008	84.49	86.38	84.91
0.001	83.85	82.92	83.38
0.003	84.23	83.58	83.90
0.005	84.36	85.16	84.76
0.008	84.21	85.75	84.46

TABLE 4: Experimental parameter settings.

Parameters	Value
Character embedding	200
Word embedding	200
LSTM dim	128
Dropout rate	0.5
Learning rate	0.005
Batch size	50
Epoch	25
Optimizer	Adam

TABLE 5: Results of fusion strategy test.

Model	Precision (P)	Recall (R)	F1 value (F1)
Model-add	82.78	85.86	83.32
Model-con	84.36	85.16	84.76
Model-wb	83.69	84.92	84.30

TABLE 6: Different types of named entity recognition results.

Model	Types	Precision (P)	Recall (R)	F1 value (F1)
CW-BiLSTM-CRF	DISE	80.16	81.62	80.88
	IMAG	79.31	80.00	79.65
	LABO	83.71	83.92	83.81
	OPER	87.95	89.88	88.90
	DRUG	81.07	79.81	80.46
	DISE	92.29	93.11	92.85
	Overall	84.36	85.16	84.76

BiLSTM-CRF model based on character and word fusion (CW-BiLSTM-CRF) with the following model. The models selected in this paper include the LSTM model based on word vector (C-LSTM), the LSTM-CRF model based on character vector (C-LSTM-CRF), the BiLSTM-CRF model based on word vector (W-BiLSTM-CRF), and the BiLSTM-CRF model based on character vector (C-BiLSTM-CRF).

4.6.1. Comparison of Convergence Rate. We monitored the change of F1 values for the five models with the number of epochs. The results are shown in Figure 4, from which it can be seen that the model proposed in this paper obtains higher F1 values at the beginning of training, its convergence speed is faster than the other models, and its F1 values are generally better than the other models. The number of epochs of each model converges quickly in the interval of 0–7 and with the increase in the number of

epochs. After approximately 15 epochs, the performance of the model fluctuates in a small range with the number of epochs, and the basic area is stable.

4.6.2. Overall Comparison. Table 7 presents the experimental results of the model proposed in this paper with other baseline models. We can observe that the CW-BiLSTM-CRF model proposed in this paper achieves the best performance among all models. The C-LSTM-CRF model performs better than the C-LSTM model because the addition of conditional random fields provides better sequence features to the model, which improves the performance of the model. The C-BiLSTM-CRF model performs better than the LSTM-CRF model based on a character vector model because the BiLSTM can capture contextual information from the input sequences, thus improving the performance of the model.

FIGURE 4: F1 values of different models varying with the number of epochs.

TABLE 7: Named entity recognition results of different models.

Models	Precision (P)	Recall (R)	F1 value (F1)
C-LSTM	77.49	76.28	76.88
C-LSTM-CRF	79.63	80.46	80.04
W-BiLSTM-CRF	82.75	81.29	82.01
C-BiLSTM-CRF	83.17	85.32	84.23
CW-BiLSTM-CRF	84.36	85.16	84.76

TABLE 8: Result comparison with other models on six types of entities.

	Disease and diagnosis			Imaging examination		
	P	R	F1	P	R	F1
C-LSTM	74.36	73.26	73.81	73.69	78.56	76.05
C-LSTM-CRF	76.82	77.23	77.02	75.36	73.52	74.43
W-BiLSTM-CRF	80.58	79.37	79.97	73.23	77.23	75.18
C-BiLSTM-CRF	81.28	76.36	78.74	78.39	81.53	79.93
CW-BiLSTM-CRF	80.16	81.62	80.88	79.31	80.00	79.65
	Laboratory test			Operation		
	P	R	F1	P	R	F1
C-LSTM	78.32	76.89	77.60	80.25	83.62	81.90
C-LSTM-CRF	76.35	79.36	77.83	84.23	81.25	82.71
W-BiLSTM-CRF	79.68	81.57	80.61	85.36	84.89	85.12
C-BiLSTM-CRF	82.63	82.69	82.66	88.32	86.58	87.44
CW-BiLSTM-CRF	83.71	83.92	83.81	87.95	89.88	88.90
	Drug			Anatomy		
	P	R	F1	P	R	F1
C-LSTM	71.26	76.35	73.72	85.32	88.60	86.93
C-LSTM-CRF	75.38	74.19	74.78	86.37	87.46	86.91
W-BiLSTM-CRF	79.26	76.52	77.87	89.23	91.25	90.23
C-BiLSTM-CRF	80.46	78.29	79.36	91.82	92.56	92.19
CW-BiLSTM-CRF	81.07	79.81	80.46	92.29	93.11	92.85

The model proposed in this paper improves the accuracy by 1.19% and the F1 value by 0.53% compared to the C-BiLSTM-CRF model. The performance of this model is also improved compared to that of the W-BiLSTM-CRF model. The medical entity recognition model proposed in this paper that fuses both character and word information can better represent the information, and the performance is improved to a certain extent compared to that of the previous models that use a single character vector or a word vector.

4.6.3. Comparison of Each Category. For further comparison, we experimented with six categories of entities separately. The final results of the experiments are presented in Table 8. From the results in the table, it can be noted that the accuracy of the model proposed in this paper is the highest, except for “Disease and diagnosis” and “Operation.” While the F1 value that is not the highest is only “Imaging examination,” its overall result is slightly lower than that of the C-BiLSTM-CRF model. The model proposed in this paper combines character and word vectors, and it needs more data to better extract the information in medical texts. Owing to the small number of “Imaging examination” entities, the model proposed in this paper cannot perform well on them. However, it performs very well on the rest of the entities and is able to recognize medical entities more accurately, which leads to the conclusion that our model still has the best overall recognition effect compared to the rest of the models.

5. Conclusions

In this study, we propose a Chinese medical entity recognition model based on character and word vector fusion that has a significant impact on improving the accuracy of Chinese medical entity recognition. The key to this model is to express the Chinese text as character and word vectors separately in the input layer, then fuse the two feature vectors obtained after the BiLSTM layer, and use the fused feature vectors to obtain the final recognition results. This approach avoids the problems of missing information of character vectors and inaccurate word separation of word vectors. On the CCKS 2019 dataset named entity recognition task for Chinese EMRs, the model proposed in this study achieves better performance than other baseline models.

In the future, we will explore the effectiveness of the character and word vector fusion strategy in other advanced models, so as to further improve the performance of the Chinese medical entity recognition model based on character and word vector fusion. Alternatively, by incorporating information from other dimensions, the model can realize better recognition performance in the face of longer entities; the fused character and word vectors model can also be applied to datasets from other domains to verify its generality.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62073123), the Open Project of the Key Laboratory of Food Information Processing and Control of Ministry of Education (KFJJ-2020-109), Key Scientific Research Projects of Higher Education Institutions in Henan Province (21A520008), and Henan Province Science and Technology Research Project (212102210058).

References

- [1] A. Alasmari and L. Zhou, “How multimorbid health information consumers interact in an online community Q&A platform,” *International Journal of Medical Informatics*, vol. 131, p. 103958, 2019.
- [2] A. L. Rector, W. D. Solomon, W. A. Nowlan, T. W. Rush, P. E. Zanzetta, and W. M. Claassen, “A Terminology Server for medical language and medical information systems,” *Methods of Information in Medicine*, vol. 34, no. 1, pp. 147–157, 1995.
- [3] R. C. Wasserman, “Electronic medical records (EMRs), epidemiology, and epistemology: reflections on EMRs and future pediatric clinical research,” *Academic pediatrics*, vol. 11, no. 4, pp. 280–287, 2011.
- [4] J. Lei, B. Tang, X. Lu, K. Gao, M. Jiang, and H. Xu, “A comprehensive study of named entity recognition in Chinese clinical text,” *Journal of the American Medical Informatics Association*, vol. 21, no. 5, pp. 808–814, 2014.
- [5] R. K. Taira and S. G. Soderland, “A statistical natural language processor for medical reports,” *Proceedings/AMIA. Annual Symposium. AMIA Symposium*, vol. 1, pp. 970–974, 1999.
- [6] X. Yu, “Relationship extraction,” *Encyclopedia of Social Network Analysis & Mining*, vol. 1, pp. 1562–1572, 2011.
- [7] L. Li, P. Wang, J. Yan et al., “Real-world data medical knowledge graph: construction and applications,” *Artificial Intelligence in Medicine*, vol. 103, no. 19, pp. 101817–101827, 2020.
- [8] C. Wang, X. Ma, J. Chen, and J. Chen, “Information extraction and knowledge graph construction from geoscience literature,” *Computers & Geosciences*, vol. 112pp. 112–120, Mar, 2018.
- [9] J. Qiu, Y. Zhou, Q. Wang, T. Ruan, and J. Gao, “Chinese clinical named entity recognition using residual dilated convolutional neural network with conditional random field,” *IEEE Transactions on NanoBioscience*, vol. 18, no. 3, pp. 306–315, 2019.
- [10] Q. Wei, T. Chen, R. Xu, Y. He, and L. Gui, “Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks,” *Database: The Journal of Biological Databases and Curation*, vol. 2016, 2016.
- [11] M. Seok, H.-J. Song, C.-Y. Park, J.-D. Kim, and Y.-s. Kim, “Named entity recognition using word embedding as a feature,” *International Journal of Software Engineering and its Applications*, vol. 10, no. 2, pp. 93–104, 2016.
- [12] F. Liao, L. Ma, J. Pei, and L. Tan, “Combined self-attention mechanism for Chinese named entity recognition in military,” *Future Internet*, vol. 11, no. 8, pp. 180–191, 2019.
- [13] J. Yin, S. Luo, Z. Wu et al., “Chinese named entity recognition with character-level BLSTM and soft attention model,”

- Journal of Beijing Institute of Technology (Social Sciences Edition)*, vol. 29, no. 103, pp. 63–74, 2020.
- [14] Z. Tang, B. Wan, and L. Yang, “Word-character graph convolution network for Chinese named entity recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1520–1532, 2020.
  - [15] G. Wu, G. Tang, Z. Wang, Z. Zhang, and Z. Wang, “An attention-based BiLSTM-CRF model for Chinese clinic named entity recognition,” *IEEE Access*, vol. 7, pp. 113942–113949, 2019.
  - [16] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
  - [17] J. Yu, J. Jiang, and R. Xia, “Entity-Sensitive attention and fusion network for entity-level multimodal sentiment classification,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 429–439, Article ID 19404836, 2020.
  - [18] C. Lyu, B. Chen, Y. Ren, and D. Ji, “Long short-term memory RNN for biomedical named entity recognition,” *BMC Bioinformatics*, vol. 18, no. 1, p. 462, 2017.
  - [19] L. Gong, Z. Zhang, and S. Chen, “Clinical named entity recognition from Chinese electronic medical records based on deep learning pretraining,” *Journal of Healthcare Engineering*, vol. 2020, no. 2, p. 8829219, 2020.
  - [20] M. Alam and S. U. Hussain, “Deep learning based roman-Urdu to Urdu transliteration,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 35, no. 04, pp. 373–382, 2021.
  - [21] Q. Wang, Y. Xia, Y. Zhou, D. Gao, Y. Xia, and P. He, “Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition,” *Journal of Biomedical Informatics*, vol. 92, 2019.
  - [22] A. Cameron, M. Stumptner, N. Nandagopal, W. Mayer, and T. Mansell, “Rule-based peer-to-peer framework for decentralised real-time service oriented architectures,” *Science of Computer Programming*, vol. 97, no. 2, pp. 202–234, 2015.
  - [23] J. Li, A. Sun, J. Han, and C. Li, “A survey on deep learning for named entity recognition,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 99, p. 1, 2020.
  - [24] R. Alfred, L. C. Leong, C. K. On, and P. Anthony, “Malay named entity recognition based on rule-based approach,” *International Journal of Machine Learning and Computing*, vol. 4, no. 3, pp. 300–306, 2014.
  - [25] J. S. Lee, E. Paintsil, V. Gopalakrishnan, and M. Ghebremichael, “A comparison of machine learning techniques for classification of HIV patients with anti-retroviral therapy-induced mitochondrial toxicity from those without mitochondrial toxicity,” *BMC Medical Research Methodology*, vol. 19, no. 1, p. 216, 2019.
  - [26] Y. Wu, M. Jiang, J. Lei, and H. Xu, “Named entity recognition in Chinese clinical text using deep neural network,” *Studies in Health Technology and Informatics*, vol. 216, pp. 624–628, 2015.
  - [27] P. Ma, B. Jiang, Z. Lu, N. Li, and Z. Jiang, “Cybersecurity named entity recognition using bidirectional long short-term memory with conditional random fields,” *Tsinghua Science and Technology*, vol. 26, no. 3, pp. 11–17, 2021.
  - [28] Y. Li, G. Du, Y. Xiang et al., “Towards Chinese clinical named entity recognition by dynamic embedding using domain-specific knowledge,” *Journal of Biomedical Informatics*, vol. 106, p. 103435, Article ID 103435, 2020.
  - [29] Y. Wu, M. Jiang, and J. Xu, “Clinical named entity recognition using deep learning models,” *AMIA Annu Symp Proc*, vol. 2017, pp. 1812–1819, 2018.
  - [30] C. Zhang, X. Xiao, and C. Wu, “Medical fraud and abuse detection system based on machine learning,” *International Journal of Environmental Research and Public Health*, vol. 17, no. 19, Article ID 3027884, 2020.
  - [31] L.-H. Lee and Y. Lu, “Multiple embeddings enhanced multi-graph neural networks for Chinese healthcare named entity recognition,” *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 7, pp. 2801–2810, 2021.
  - [32] Y. Gao, Y. Wang, P. Wang, and L. Gu, “Medical named entity extraction from Chinese resident admit notes using character and word attention-enhanced neural network,” *International Journal of Environmental Research and Public Health*, vol. 17, no. 5, Article ID 32131522, 2020.
  - [33] J. Liang, X. Xian, X. He et al., “A novel approach towards medical entity recognition in Chinese clinical text,” *Journal of Healthcare Engineering*, vol. 2017, p. 4898963, Article ID 4898963, 2017.
  - [34] Q. Wang and M. Iwaihara, “Deep Neural Architectures for Joint Named Entity Recognition and Disambiguation,” *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 2019.
  - [35] 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.
  - [36] H. M. Yang, L. Li, and R. D. Yang, “Named entity recognition based on bidirectional long short-term memory combined with case report form,” *Chinese Journal of Tissue Engineering Research*, vol. 22, pp. 3237–3242, 2018.
  - [37] K. Xue, Y. Zhou, and Z. Ma, “Fine-tuning BERT for joint entity and relation extraction in Chinese medical text,” in *Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 892–897, IEEE, San Diego, CA, USA, Nov. 2019.
  - [38] Z. Zhang, T. Zhou, Y. Zhang, and Y. Pang, “Attention-based deep residual learning network for entity relation extraction in Chinese EMRs,” *BMC Medical Informatics and Decision Making*, vol. 19, no. 55, pp. 55–177, 2019.
  - [39] M. Yin, C. Mou, and K. Xiong, “Chinese clinical named entity recognition with radical-level feature and self-attention mechanism,” *Journal of Biomedical Informatics*, vol. 98, 2019.
  - [40] F. A. Gers and E. Schmidhuber, “LSTM recurrent networks learn simple context-free and context-sensitive languages,” *IEEE Transactions on Neural Networks*, vol. 12, no. 6, pp. 1333–1340, 2001.
  - [41] E. Kiperwasser and Y. Goldberg, “Simple and accurate dependency parsing using bidirectional LSTM feature representations,” *Transactions of the Association for Computational Linguistics*, vol. 4, no. 1, pp. 313–327, 2016.
  - [42] D. C. Wintaka, M. A. Bijaksana, and I. Asror, “Named-entity recognition on Indonesian tweets using bidirectional LSTM-CRF,” *Procedia Computer Science*, vol. 157, no. 1, pp. 221–228, 2019.
  - [43] Y. Du, Q. Li, and L. Wang, “Biomedical-domain pre-trained language model for extractive summarization,” *Knowledge-Based Systems*, vol. 199, Article ID 105964, 2020.
  - [44] M. Peters, M. Neumann, and M. Iyyer, “Deep contextualized word representations,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2227–2237, Association for Computational Linguistics, New Orleans, Louisiana, June 2018.
  - [45] T. Qian, M. Zhang, and Y. Lou, “A joint model for named entity recognition with sentence-level entity type Attentions,”

- IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, no. 99, pp. 1468–1448, 2021.
- [46] B. Ji, S. Li, and J. Yu, “Research on Chinese medical named entity recognition based on collaborative cooperation of multiple neural network models,” *Journal of Biomedical Informatics*, vol. 104, Article ID 103395, 2020.
  - [47] E. Ford, J. A. Carroll, H. E. Smith, D. Scott, and J. A. Cassell, “Extracting information from the text of electronic medical records to improve case detection: a systematic review,” *Journal of the American Medical Informatics Association*, vol. 23, no. 5, pp. 1007–1015, 2016.
  - [48] F. Ye, Y. Chen, G. Zhou, and H.-M. Li, “Intelligent recognition of named entity in electronic medical records,” *Chinese Journal of Biomedical Engineering*, vol. 30, no. 2, pp. 256–262, 2011.
  - [49] Na Ye, X. Qin, and L. Dong, “Chinese named entity recognition based on character-word vector fusion,” *Wireless Communications and Mobile Computing*, vol. 2020, no. 3, Article ID 8866540, 2020.
  - [50] N. Zhang, G. Xu, Z. Zhang, and F. Li, “MIFM: multi-granularity information fusion model for Chinese named entity recognition,” *IEEE Access*, vol. 7, pp. 181648–181655, 2019.
  - [51] Y. Jize, L. Senlin, and W. Zhouting, “Chinese named entity recognition with character-level BLSTM and soft attention model,” *Journal of Beijing Institute of Technology (Social Sciences Edition)*, vol. 29, no. 103, pp. 63–74, 2020.
  - [52] J. P. C. Chiu and E. Nichols, “Named entity recognition with bidirectional LSTM-CNNs,” *Transactions of the Association for Computational Linguistics*, vol. 4, no. 1, pp. 357–370, 2016.
  - [53] J. Yang, Y. Zhang, and L. Li, “YEDDA: a lightweight collaborative text span annotation tool,” in *Proceedings of the ACL 2018, System Demonstrations*, pp. 31–36, Association for Computational Linguistics, Melbourne, Australia, July 2018.