ABSTRACT
In order to solve the problem that traditional word vectors are difficult to express the context semantics and the feature extraction of traditional model is single, a multi-feature fusion model named BERT-BiLSTM-IDCNN-Attention-CRF for Named Entity Recognition is proposed, which uses BERT to model the context semantic relationship of word vectors and fuse the context features and local features extracted by BiLSTM and IDCNN respectively. The proposed model is tested on Chinese Electronic Medical Record (EMR) dataset issued by China Conference on Knowledge Graph and Semantic Computing 2020 (CCKS2020). Compared with the baseline models such as BiLSTM-CRF, the experiment on CCKS2020 data shows that BERT-BiLSTM-IDCNN-Attention-CRF achieves 1.27% improvement in F1. The experimental results show that the proposed model can better identify the medical entities in EMR.

CCS CONCEPTS
• Computing methodologies → Artificial intelligence; Natural language processing; Information extraction.

KEYWORDS
Named Entity Recognition, multi feature fusion, BERT, BiLSTM, IDCNN, CCKS2020

1 INTRODUCTION
Named Entity Recognition (NER) is a basic task of Natural Language Processing (NLP). The purpose is to identify and classify entities with specific meanings such as names of person (PER), location (LOC) and organization (ORG) in the corpus. Using NER can quickly obtain useful entity information from unstructured text data, and convert these unstructured data into structured data, which can efficiently analyze the data. Therefore, NER is also a basic task in knowledge graph. The electronic medical record (EMR) is generated during the clinical treatment of patients. It records the disease changes and treatment process of patients in an electronic way, and it is the basic material for clinical scientific diagnosis and treatment [1]. EMR contains a wealth of medical entities. Through using NER, various medical entities can be mined, it can be used to build a medical knowledge graph and improve the availability, intelligibility and visibility of data [2].

2 RESEARCH PROBLEM
Research on NER has gradually evolved from the early rule-based and dictionary-based methods to machine learning and deep learning. Early rule [3] or dictionary-based [4] methods require the construction of a large number of rule sets or dictionaries, and using pattern matching methods to extract entities from the corpus. With the development of machine learning technology, machine learning algorithm based on statistics is widely used in the task of NER. In this method, text features are manually selected and input into the machine learning algorithm to classify the labels corresponding to each character in the sequence.

In the last few years, deep learning methods which based on the neural network have become the mainstream method of NER task because of their powerful feature extraction capabilities. Yang [5] et al. trained an entity recognition model based on BiLSTM-CRF, and effectively extracted medical entities from admission records and discharge summary. Chiu [6] et al. used the hybrid structure of BiLSTM and CNN to obtain word-level and character-level features, which further improves the recognition performance of the model. Strubell [7] et al. applied IDCNN to NER, which greatly reduced the training time of the model. In recent years, the attention mechanism has been widely used in the field of NLP. YIN [8] et al. used CNN to extract the feature information between Chinese characters, and used the self-attention mechanism to capture the dependency characteristics between characters to identify relevant entities in the medical electronic medical record.
The above traditional methods failed to make full use of the advantages of different granular features in entity recognition, and the learning ability of the model is limited by the traditional static word vector. To solve the above problems, we use fine-tuned BERT to extract the dynamic word vectors and concatenate the part of speech and other features together as the output of the embedding layer; In the feature extraction layer, we use BiLSTM and IDCNN to extract context-dependent features and local features; Finally, the global optimal tag sequence is obtained through the CRF after the dynamic fusion of the two extracted features. The model combines two kinds of different granular features, which effectively improves the accuracy of model recognition.

3 THE FRAMEWORK OF NAMED ENTITY RECOGNITION

In this section we describe the model architecture of NER, which named BERT-BiLSTM-IDCNN-Attention-CRF. The model structure is shown in Figure 1. From the bottom up, the model includes 4 layers as following.

Embedding layer: Obtain dynamic word vectors containing rich language knowledge, and concatenate additional 5 feature embeddings.

Feature extraction layer: Use BiLSTM and IDCNN to extract context-dependent features and character local features.

Feature fusion layer: The two types of extracted features are fused using a dynamic fusion mechanism.

Feature decoding layer: Use CRF to decode the fused features to obtain the global optimal tag sequence.

3.1 Embedding Layer

3.1.1 BERT Word Vector Embedding. The model cannot directly process characters. Before feeding data into the model, the original text data needs to be transformed into vector representation. The existing solution is to use word embedding which can map the word to a continuously-valued low dimensional space. Related research [9] has proved that as a basic procedure of NLP, word embedding can encode the internal semantic and syntactic information of words, and it significantly improves the performance in NER tasks. However, the traditional word embedding methods such as Word2Vec [10], GloVe [11] have the fatal disadvantage: the embedding of the same word in different sentences are identical. BERT [12] can solve these problems well. Like ELMo [13] and GPT [14], BERT is designed as a pre-training language model which can significantly improve the performance of many NLP tasks and reduce the large number of labeled data needed for training stage. BERT also includes two stages of training. First, through the pre-training process, BERT can not only learn rich prior semantic knowledge from a large number of unlabeled data, but also model contextual semantic knowledge for each word in the sequence through multi-layer transformer, the core idea of transformer is to use attention mechanism which can calculate the relationship between each word in a sentence and all the words in the sentence, so that the embeddings for same word in different semantic sentences are different. Second, BERT can be fine-tuned using labeled data from the downstream tasks, in order to modify the network parameters obtained in the pre-training stage, so that the model can be integrated with domain knowledge and adapt the requirement of downstream tasks.

In this paper, we will use BERT to train Chinese character embeddings. The publicly available RoBERTa [15] pre-trained model on general domain is served as our based model. Then we choose to fine-tune its parameters to obtain the fine-tuned RoBERTa-FT model and frozen its parameters. BERT is only used as a feature generator of word vectors to convert the input text sequence into word vector sequence [16], and on this basis, the word embedding which generated by the lookup table in the downstream network structure is used as a supplement to the non-training part of the BERT generated word vector.

3.1.2 Feature Embedding. In the text of EMR, named entities such as "解剖部位" (body), "疾病与诊断" (disease) and "药物" (drug) are mostly noun, while "影像检查" (check) and "手术" (operation) are usually preceded by the verb ‘行’ to indicate the occurrence of this activity. Therefore, we use the fastHan (https://github.com/fastnlp/fastHan/) to extract the part of speech (POS) features and word boundary features of words. Because traditional radicals are more explanatory than simplified radicals in the font and structure, and small in quantity. We build the traditional radicals mapping table to obtain the radical of each character.

In order to improve the recognition accuracy of specific entities, we have made a drug dictionary. First, we download the drug name dictionary from the Sougou library (https://pinyin.sogou.com/dict cate/), then manually remove non-drug name and add all drug entities which in the training set. Finally, we use Bi-directional Maximum Match (BMM) algorithm, match the entities that appear in the dictionary from the EMR text and tagging them. The annotation examples of feature embedding are shown in Table 1

![Figure 1: Structure of Medical Named Entity Recognition Model](image-url)
3.2 Feature Extraction Layer

3.2.1 BiLSTM. Recurrent Neural Network (RNN) can memorize the information at the previous time and participate in the output calculation of the current time, which is suitable for processing sequence data such as text. However, RNN is prone to the problem of gradient disappearance and explosion when calculating the gradient in the backpropagation process, and produce long-distance dependency problems and lose “Memory” ability. Long Short-Term Memory (LSTM) is an improvement of RNN. As shown in Figure 2, LSTM [17] adds an extra memory cell ‘c’ to the hidden layer to store long-term state information, and introduces Input gate (I_t), Forget gate (F_t) and Output gate (O_t). It uses gating mechanism to selectively let information pass through, effectively alleviating the problem of gradient disappearance or explosion in RNN.

We choose the BiLSTM which is composed of forward and backward LSTM to model the context information of the input sequence, so as to get a more global representation of the sequence.

\[ H_{f,t} = h_{f,t} @ h_{b,t} \]  (2)

Where \( h_{f,t} \) is the output feature of the forward LSTM, \( h_{b,t} \) is the output feature of the backward LSTM, and \( H_{f,t} \) is the result of the concatenate in the last dimension of the two features.

3.2.2 Local Feature Extraction. 1. IDCNN-CNN uses the method of local connection to convolute the feature map to extract local features. For sequence, the relationship between adjacent characters is relatively close, while the relationship between distant characters is relatively weak. CNN can perceive the local information of sequence, and then synthesize the local information at a higher level through stacking layers, so as to obtain the global information. Therefore, CNN can obtain the global features of sequence while taking into account the local features. In this paper, we using Iterated Dilated Convolutions [18] (IDCNN) to extract local features, to improve the performance of medical entity recognition. IDCNN expands the receptive field of convolution kernel by adding an expansion width on the convolution kernel, so that each convolution output contains a large range of information with fewer parameters, and the receptive field increases exponentially with the increase of the number of layers. The effect of covering all input sequences can be achieved by stacking several convolution layers.

In this paper, 4 convolution blocks of the same size are stacked together. The expansion width of each convolution block is 1, 1, 2, and the number of convolution kernel is 128. In order to control the information flow better, a gating mechanism is added to the last layer of each convolution block, so that the model can select more useful features for prediction to pass through with greater weight. We also add residual connection to solve the problem of gradient disappearance. The structure of IDCNN model is shown in Figure 3.

Convolution operation is performed on the input of the last layer of each convolution block, and the obtained feature map is activated by sigmoid function. Since the sigmoid function value has a range of [0,1], it can be regarded as a “gate” to control the flow. The input \( C_{t,3} \) make a point multiplication with the “gate” to achieve the function of controlling the flow of information. The calculation of gating is shown in Eq. 3), where \( C_{t,3} \) is the third layer convolution output of the \( t \)-th convolution block.

\[ Z_t = C_{t,3} \times Conv(C_{t,3}) \]  (3)
2. RefineNet: The outputs of IDCNN are four feature maps of the same size, which represent the features from the low level to the high level. Direct concatenate two features will lead to the dimension of the final feature map too large, only taking the last layer will lose the low-level features; these two methods can’t make full use of the information of each layer. In this paper, we select RefineNet [19] network, input the output features of IDCNN into the RefineNet network, and fuse high-level semantic features with fine-grained low-level features.

\[ Z = \text{RefineNet}([Z_1, Z_2, Z_3, Z_4]) \]  

3. Multi-Head Self-Attention: The core goal of attention mechanism in NLP is to select the information that is more critical to the current task from a large number of text information, and assign more weight to these information, so as to quickly extract the important features of sparse data. Self-Attention [20] mechanism is an improvement of attention mechanism. It uses the same source query \( Q \) and key value \( K \) to calculate the attention within the sequence, and calculates the association between each word and all other words. Multi-Head Self-Attention can divide the model into multiple heads to form multiple subspaces, so that the model can capture the internal correlation of the sequence form multiple angles, so as to capture longer distance dependence.

In this paper, we use the improved Multi-Head Self-Attention mechanism in TENER [21] to calculate the attention of the output features of RefineNet, which can effectively capture the relative position and direction information, and strengthen the features that play an important role in the recognition of the named entities.

\[ R_{t-j} = [\ldots \sin \left( \frac{t-j}{10000^{\frac{2k}{d_k}}} \right) \cos \left( \frac{t-j}{10000^{\frac{2k}{d_k}}} \right) \ldots]^T \]  

\[ A^{rel}_{t,j} = Q_t K_j^T + Q_t R_{t-j} + u K_j^T + v R_{t-j} \]  

\[ \text{Attention}(Q, K, V) = \text{Softmax}(A^{rel}) \]  

Where Eq. 5 represents the query vector \( Q \), key vector \( K \) and value vector \( V \) generated by input vector \( Z \). Eq. 6) \( R_{t-j} \) is the relative distance between \( t-th \) and \( j-th \) characters calculated by sine and cosine functions. Eq. 7) \( A^{rel}_{t,j} \) is the similarity between \( t-th \) and \( j-th \) characters, \( u \) and \( v \) are the bias terms respectively. Eq. 8) shows that the similarity matrix \( A^{rel} \) between the calculated characters is normalized to the probability distribution by using Softmax function, and multiplied by the corresponding value vector \( V \) to get the single head attention score. Multi-Head Self-Attention is to concatenate the attention generated by multiple subspaces.

From the perspective of feature dimensions and distribution, the extracted local feature is input to a layer of private BiLSTM, which will give the output feature with the same dimension as the extracted global features.

3.3 Feature Fusion Layer

In order to make full use of the extracted two types of features, this paper adopts the method of dynamic feature fusion. The model calculates the corresponding weight for each type of feature, and adaptively fuses the two types of features. First, map each type of feature into 1 dimension, and then concatenate the two types of features in the last dimension and input the Softmax layer to get the weight corresponding to each type of feature. The last two types of feature vectors are multiplied by the weight and then added together. The fusion result is obtained, as shown in Eq. 9, Eq. 10, and Eq. 11. Then input the fused feature into the shared BiLSTM, and further extract the fused semantic features.

\[ a_i = \text{Dense}_{\text{unit}=1}(\text{output}_i) \]  

\[ a_i = \frac{e^{a_i}}{\sum_n e^{a_n}} \]  

\[ \text{Output} = \sum_i a_i \cdot \text{output}_i \]  

3.4 Feature Decoding Layer

In named entity recognition, there is a constraint relationship between tags and tags. For example, after ‘I-drug’, it can’t be followed by ‘B-drug’, but can only be followed by ‘I-drug’ or ‘E-drug’. Conditional Random Field (CRF) [22] can make full use of the context information of entity tags, and train a transition probability matrix to learn the transition relationship between tags, so as to solve the tag bias problem. Therefore, this paper uses CRF to model the dependency between sequence tags and obtain the optimal tag sequence.

Specifically, for the given observation sequence \( X = x_1, x_2, \ldots, x_n \), a prediction label sequence \( Y = y_1, y_2, \ldots, y_n \) can be obtained by CRF. The total score of a sentence is calculated as the following Eq. 12).

\[ s(X, y) = \sum_{i=0}^{n} A_{y_i y_{i+1}} + \sum_{i=1}^{n} P_{y_i} \]  

Where \( P \) is the emission score, which is the output matrix of the feature extraction layer, \( P_{y_j} \) is the score generated when \( x_i \) in the sequence is classified to the \( j-th \) tag; \( A \) is the transition matrix, \( A_{y_j y_{j+1}} \) is the transfer score from the \( i-th \) tag to the \( j-th \) tag.

Then, the normalized probability is obtained by Softmax, the calculation is shown in Eq. 13).

\[ p(y|X) = \frac{e^{s(X,y)}}{\sum_{y \in Y_X} e^{s(X,y)}} \]  

The CRF parameters are estimated by using the maximum conditional likelihood estimation in the training process, and the calculation is shown in Eq. 14).

\[ \log (p(y|X)) = s(X,y) - \log \left( \sum_{\hat{y} \in Y_X} e^{s(X,\hat{y})} \right) \]  

Where \( Y_X \) represents the set of all possible tag sequences for the input sequence \( X \).

\[ y^* = \arg\max_{\hat{y} \in Y_X} s(X,\hat{y}) \]  

4 DATE SOURCE AND PREPROCESSING

4.1 Date Source

The dataset is CCKS2020 Chinese EMR, which is marked with six types, including "疾病和诊断" (disease), "解剖部位" (body), "实验室检验" (test), "影像检查" (check), "手术" (operation), and "药物" (drug), including 1 050 labeled text.
Table 2: The Type and Number of Entities in Training and Test Set

<table>
<thead>
<tr>
<th>Entity Type</th>
<th>Entity Definition</th>
<th>Training Set</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease</td>
<td>疾病和诊断</td>
<td>3 158</td>
<td>880</td>
</tr>
<tr>
<td>Body</td>
<td>解剖部位</td>
<td>3 985</td>
<td>939</td>
</tr>
<tr>
<td>Check</td>
<td>影像检查</td>
<td>803</td>
<td>216</td>
</tr>
<tr>
<td>Test</td>
<td>实验室检查</td>
<td>1 022</td>
<td>266</td>
</tr>
<tr>
<td>Operation</td>
<td>手术</td>
<td>748</td>
<td>172</td>
</tr>
<tr>
<td>Drug</td>
<td>药物</td>
<td>1 573</td>
<td>410</td>
</tr>
</tbody>
</table>

Table 3: Data Annotation Example

<table>
<thead>
<tr>
<th>复查</th>
<th>C</th>
<th>T</th>
<th>提示</th>
<th>肝</th>
<th>脏</th>
<th>较大</th>
<th>增大</th>
</tr>
</thead>
<tbody>
<tr>
<td>O</td>
<td>O</td>
<td>B-check</td>
<td>E-check</td>
<td>O</td>
<td>O</td>
<td>B-body</td>
<td>E-body</td>
</tr>
</tbody>
</table>

4.2 Date Preprocessing

The CCKS2020 is manually annotated by professionals. By checking the annotation of the dataset, we found that there are a lot of problems such as inconsistent labels, missing labels, and wrong labels. Therefore, for all datasets, we performed the following processing. First, we modified the annotation manually to get a relatively clean dataset. In addition, we unify the case of letters and punctuation in Chinese and English in the dataset. Finally, under the premise of ensuring relatively complete semantics, we segment the sentences, and set the longest to 202 for each sentence and the shortest to 20. After data preprocessing, the type and number of entities in the training set and test set are shown in Table 2.

4.3 Entity Annotation

Named entity recognition can be regarded as a sequence tagging problem, which needs to process the original annotation corpus into sequence tagging form. In this paper, we use the BIOES annotation scheme to map the label given by the dataset to each character and mark it at the character level [23]. Where B, I and E represent the beginning, middle and end of entity respectively, O represents non-entity and S represents single character entity. An example of data annotation format is shown in Table 3.

5 EXPERIMENTAL RESULTS AND ANALYSIS

5.1 Experimental Parameters

We leverage the RoBERTa\textsubscript{base}−wwm pre-trained model as the basic model in the experiment. The dimension of output vector is 728 and 128 after dimension reduction. Epoch set to 55, batch_size set to 10, dropout set to 0.5, clip set to 5, LSTM units set to 128. Learning rate using warmup mechanism, among them, the learning_rate set to 0.003, and decay_steps set to 3000, decay_rate set to 0.90, etc. For evaluation, we use entity-level Precision\(P\), Recall\(R\) and F1 score as the metric.

5.2 Experimental Parameters

This paper compares the results of CRF++, BiLSTM-CRF, BiLSTM-CRF+feature embedding, BiLSTM-CRF+ feature embedding (drug_dict), BiLSTM-IDCNN-Attention-CRF, BERT-BiLSTM-IDCNN-Attention-CRF model, and experimental comparison results are shown in Table 4.

It can be seen from Table 4 that compared with the machine learning algorithm CRF++, the recognition effect of the deep learning algorithm BiLSTM-CRF has been greatly improved. It can be seen from method 3 that when the features of POS, word boundaries, radicals are added to the embedding layer, the recognition effect is improved; Method 4 is added to the drug dictionary feature, which assists the identification of drug categories, so the F1 value is improved to some extent, which proves the effectiveness of feature embedding. Method 5 integrates the local features extracted by IDCNN on the basis of BiLSTM extraction features and uses attention to strengthen the weight of important characters, which greatly improves the model recognition effect and proves the effectiveness of fusing local features. Method 6 uses BERT in the embedding layer to generate dynamic word vectors on the basis of method 5. Benefiting from the rich semantic information contained in the BERT word vectors, the model recognition effect is improved to a certain extent. However, considering the size of the dataset and the quality of labeling, the method of finally using BERT to generate word vectors is still relatively limited in the recognition of medical electronic medical records.

6 SUMMARY

In this paper, we use BERT as the embedding layer to generate dynamic word vectors with rich semantic information. In view of the lack of local feature extraction ability of single BiLSTM, we use IDCNN to extract local features of the text, and integrate the extracted multi-layer features through RefineNet to make full use of the extracted information of each layer. Then the Self-Attention mechanism is used to enhance the features which play an important role in entity recognition and improve the performance of model recognition. Finally, the extracted two kinds of features are sent to the CRF decoding layer by using the dynamic fusion method to obtain the optimal tag sequence. By testing the CCKS2020 medical
Table 4: Comparison of Experimental Results

<table>
<thead>
<tr>
<th>Id</th>
<th>Model</th>
<th>P(%)</th>
<th>R(%)</th>
<th>F1(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CRF++</td>
<td>87.80</td>
<td>83.23</td>
<td>85.45</td>
</tr>
<tr>
<td>2</td>
<td>BiLSTM-CRF</td>
<td>88.60</td>
<td>88.21</td>
<td>88.41</td>
</tr>
<tr>
<td>3</td>
<td>BiLSTM-CRF+Feature embedding</td>
<td>89.18</td>
<td>88.43</td>
<td>88.80</td>
</tr>
<tr>
<td>4</td>
<td>BiLSTM-CRF+ Feature embedding(drug_dict)</td>
<td>89.13</td>
<td>88.86</td>
<td>88.99</td>
</tr>
<tr>
<td>5</td>
<td>BiLSTM-IDCNN-Attention-CRF</td>
<td>89.52</td>
<td>89.40</td>
<td>89.46</td>
</tr>
<tr>
<td>6</td>
<td>BERT- BiLSTM-IDCNN-Attention-CRF</td>
<td>89.36</td>
<td><strong>90.00</strong></td>
<td>89.68</td>
</tr>
</tbody>
</table>

electronic medical record dataset, the results show that the multi feature fusion model based on BERT proposed in this paper can significantly improve the medical Named entity recognition.

ACKNOWLEDGMENTS

This work was supported by National Key Research & Development Program of China(No.2016YFB1001103), Hongyan YUN is the corresponding author of this paper.

REFERENCES