applied sciences

MDPI

Article

BIBC: A Chinese Named Entity Recognition Model for Diabetes Research

Lei Yang [1], Yufan Fu [2] and Yu Dai [2,*]

[1] Computer Science and Engineering, Northeastern University, Shenyang 110169, China; yanglei@mail.neu.edu.cn
[2] Software College, Northeastern University, Shenyang 110169, China; 1971110@stu.neu.edu.cn
* Correspondence: daiy@swc.neu.edu.cn

Abstract: In the medical field, extracting medical entities from text by Named Entity Recognition (NER) has become one of the research hotspots. This thesis takes the chapter-level diabetes literature as the research object and uses a deep learning method to extract medical entities in the literature. Based on the deep and bidirectional transformer network structure, the pre-training language model BERT model can solve the problem of polysemous word representation, and supplement the features by large-scale unlabeled data, combined with BiLSTM-CRF model extracts of the long-distance features of sentences. On this basis, in view of the problem that the model cannot focus on the local information of the sentence, resulting in insufficient feature extraction, and considering the characteristics of Chinese data mainly in words, this thesis proposes a Named Entity Recognition method based on BIBC. This method combines Iterated Dilated CNN to enable the model to take into account global and local features at the same time, and uses the BERT-WWM model based on whole word masking to further extract semantic information from Chinese data. In the experiment of diabetic entity recognition in Ruijin Hospital, the accuracy rate, recall rate, and F1 score are improved to 79.58%, 80.21%, and 79.89%, which are better than the evaluation indexes of existing studies. It indicates that the method can extract the semantic information of diabetic text more accurately and obtain good entity recognition results, which can meet the requirements of practical applications.

Keywords: named entity recognition; BERT; IDCNN; diabetes dataset

Citation: Yang, L.; Fu, Y.; Dai, Y. BIBC: A Chinese Named Entity Recognition Model for Diabetes Research. *Appl. Sci.* **2021**, *11*, 9653. https://doi.org/10.3390/app11209653

Academic Editor: Agnese Magnani

Received: 8 September 2021
Accepted: 10 October 2021
Published: 16 October 2021

1. Introduction

Named Entity Recognition (NER), as a prerequisite for information extraction, is a study aimed at annotating characters in text sequences and finding out entity types. NER as a foundational task was first proposed by Grishman et al. at the sixth in a series of Message Understanding Conferences (MUC-6) [1] and then expanded in MUC-7 [2] with four types of entities: time, date, percentage, and value. In the extraction process, researchers found that extracting these valuable entities can be of great use, and NER is the basic work for tasks such as question and answer systems and building knowledge graphs.

In the medical NER task, the early approaches used are rule-based and lexicon-based approaches. That is, scholars in the related fields manually construct some rule templates to recognize entities in the text by pattern matching or string matching. In the medical field, the entities involved are mainly entity types such as drug names, disease names, drug doses, symptom names, etc. Since there are many fixed proper names in this field, researchers initially used a dictionary and rule-based approach in the medical field. As many of the rules used need to be developed by experienced experts, these efforts are quite costly and each domain has different rules that are difficult to transfer to other domains. Because of the capability to automatically learn text features of the deep learning models, deep learning based NER methods have become a hot research topic [3].

As a classical sequence labeling task, NER is a very fundamental technology in many high level NLP applications [4]. The deep learning based NER models mostly consist of

three parts. The first one is the embedding stage. In this stage, we map words into distributed representations and a series of Pre-trained Language Models (PLMs) are proposed to learn contextual word representations. GPT [5] selects Transformer [6] to extract one-way text information and achieves great performance. In addition, BERT [7] has also achieved good results under many NLP tasks which uses CBOW to train a bidirectional language model and MLM to stochastically mask entities in text input. As Chinese has its unique characteristics that differ in various fields such as syntactic relations, semantic relations, and so on [8], scholars begin to build models that fit Chinese lingual characteristics [9]. Even though Li et al. [10] try to prove that tokenization in Chinese has relatively little effect, Sun et al. [11] propose ERNIE with three masking strategies to better capture multi-granularity semantics. In 2019, Cui et al. [12] propose BERT with whole word masking which mask all characters included in one Chinese word to adapt to the natural characteristics of Chinese.

The second part is the context encoder module to extract sequence features and then capture the contextual dependencies of the input sequence. In order to make the future state also able to predict the current output, bidirectional RNNs such as Bi-LSTM [13] and Bi-GRU are proposed. According to the research of Liu et al. [14], LSTM has great potential for NER for clinical texts because it does not require hand-crafted features. However, the training speed of RNNs is limited by its temporal characteristics, and the advantages of convolutional kernel weight sharing used in CNN models can reduce the computational complexity and multiple convolutional kernels that can be computed in parallel are revalued by scholars [15,16].

The third and the last part, namely the inference module, takes representations from the second part and generate the optimal label to finish the process of NER. The Softmax function is a generalization of the binary classification function Sigmoid for multi-classification tasks, aiming at presenting the results of multi-classification in the form of probabilities. As a linear classifier, it is a popular component of many models that treat the process of sequence labeling as a set of independent classification tasks. This leads to neglecting dependencies in nearby labels. Thus, the Conditional Random Fields (CRF) is used to learn the cross-label dependencies, which has been proved to be effective and become a competitive option in the NER task [17].

However, the semantic information in the text still needs to be acquired via a more efficient approach to improve NER results. To address this problem, this paper proposes BIBC, an NER model that utilizes the latest breakthroughs in the NER field on the diabetes dataset, in order to improve the labeling performance on this task.

Our contributions can be summarized as follows:

- We make improvements based on the BiLSTM-CRF model and achieve better performance on the NER task.
- We analyze the features of the related models and Ruijin Dataset, and design a new data preprocessing method to address the long text input problem.
- We design the BIBC model to better capture both local and global sequence features to further improve the model effectiveness, and experimental results verify the performance on the Ruijin Dataset.

2. Related Work

In 2015, Huang et al. [18] proposed the classical BiLSTM-CRF sequence annotation model, which uses BiLSTM to extract the sentence distant contextual information, combined with CRF to consider tag dependencies, and obtained the best results at that time. However, since Chinese characters do not have clear separators between them and face difficulties in word separation, the performance of using the classical model directly on Chinese datasets will be deviated. Thus, many scholars in China have optimized and adjusted the model for the characteristics of Chinese datasets. Zhang et al. [19] designed a Lattice LSTM network structure that combines information about the characters themselves and the words they belong to to mitigate the effects of incorrect word separation. To make full use of the

information in the text, Qiang et al. [20] also proposed to add lexical features to enhance the performance of the neural network model.

Other scholars proposed to use convolutional neural networks [21] for the task of Chinese entity recognition, using CNNs to extract local features of sequences to solve the impact of semantic deficiencies caused by unseparated Chinese data. Chiu et al. [13] proposed to use a combination of BiLSTM and CNN models, using CNNs to learn fine-grained features in characters and BiLSTMs to complete the task of sequence annotation. A combined model is proposed by combining the advantages of the models. Ding et al. [22] used graph neural networks to model the entity recognition task and used external lexicons for feature supplementation during the training process to learn the features inside automatically using the features of the model. Chiu [13] et al. argue that a dilated convolution network can improve the speed in training and prediction by overlaying CNNs, which can expand the perceptual field of the model. In addition, Strubell et al. [23] propose an Iterated Dilated Convolutional Neural Network (IDCNN), which makes remarkable improvement in speed and computationally efficiency with a SOTA-level accuracy in an NER task. However, these studies do not consider in an integrated way that fusing models together can capture more complete global and local features.

In recent years, Chinese entity recognition models have been gradually applied to the biomedical field. Entities in the biomedical domain are more specialized than those in other domains for Chinese recognition tasks, and there are a large number of entities with a mixture of numbers, Chinese, English, and symbols, e.g., medication dose entity: '500 mg/d'; the entity of the test indicator: '>12.4 mmol/L', etc. The existence of these entities makes the identification of Chinese named entities in the medical field extremely difficult.

To address these difficulties, a number of medical NER models have been investigated. In 2019, Zhang et al. [24] pre-trained BERT and used the embedding as the input feature of BiLSTM-CRF to solve the problem of medical NER for breast cancer. Li et al. [25] combined the attention mechanism and BiLSTM for entity extraction from Chinese electronic medical records. The method captured more contextual information about the entities through attention and further improved the model recognition performance using features such as medical lexicon and lexical properties. Ji et al. [26] added an entity auto-correction algorithm to the attention-Bi-LSTM-CRF model to correct the recognition of entities using historical entity information. In 2020, Li et al. [27] pre-trained the BERT model using untagged clinical text based on the clinical NER datasets of CCKS-2017 and CCKS-2018, and further improved the model performance using the introduction of dictionary features and root features of Chinese characters. In 2021, Li et al. [28] used CRF to specify identification rules in the identification of named entities in Chinese electronic medical records, solving the problem of ambiguous classification boundaries. Zhou et al. [29] proposed a label recalibration strategy and a knowledge distillation technique in the face of insufficient training sets. The label recalibration strategy improves the recall of the weakly labeled dataset in an iterative manner without introducing noise, obtaining two high-quality datasets. In addition, the knowledge distillation technique compresses the recognition models trained from the two datasets into one recognition model, which eventually achieved good results on the CDR and NCBI disease corpus. While the above studies have produced good results, most of them are based on processed medical corpus. In fact, few processed and open-sourced datasets are available due to privacy issues and specialization. In addition, a summary of the models in the related work is shown in Table 1.

Table 1. A summary of NER models.

Study	Model	Task
Huang et al. [18]	BiLSTM + CRF	NER
Zhang et al. [19]	Lattice-LSTM+CRF	Chinese NER
Qiang et al. [20]	6-Tag-Boundary-Pos	Chinese NER
Chiu et al. [13]	BiLSTM+CNN	Chinese NER
Ding et al. [22]	graph neural networks	Chinese NER
Strubell et al. [23]	IDCNNs	Sequence labeling&NER
Zhang et al. [24]	BERT+BiLSTM-CRF	Medical NER for breast cancer
Li et al. [25]	BiLSTM+Attention	Entity extraction of Chinese EMR
Ji et al. [26]	Attention-BiLSTM-CRF	Chinese EMR NER
Li et al. [27]	BERT+BiLSTM+CRF	Clinical NER
Li et al. [28]	CRF	Chinese EMR NER
Zhou et al. [29]	BioBERT-CRF	Biomedical NER

3. Method

3.1. Data Preprocessing

The dataset used in this paper is the diabetes-related literature provided by Ruijin Hospital, which includes textbooks, research papers, and clinical guidelines related to diabetes. This dataset is a chapter-level text, in which there are 15 medical entities, including examination methods, etiology, clinical manifestations, medication methods, and sites.

In NLP tasks, the sentences input to the model need to be of appropriate length. Too short sentence length will lose the feature information of the text, while too large sentence length is not conducive to the model training process. The diabetes-related medical literature used in this paper is chapter-level Chinese text, and an article with thousands or even tens of thousands of words cannot be directly input to the model for training, so it is necessary to divide the long diabetes text and transform it into data suitable for model input.

Because of formatting conversion of the original text, there will be line breaks or periods between entities in the article, which leads to a lack of good sentence boundaries. Thus, directly using punctuation to divide the sentences may have the situation of dividing one entity in two sentences, resulting in incomplete entities. To address the problem, in this paper, we propose to use a sliding window size equal to the step size to slice the sentence, and extend the window left and right by a certain length of characters. Thus, the length of the sentence can be controlled and all have certain contextual information, avoiding the case of entity segmentation. The specific division method is shown in Figure 1, in which the size of the sliding window (*window_size*) is 5, the number of characters extended to the left and right of the sentence (*pad_size*) is 3, and the final sentence length obtained is $window\_size + 2 * pad\_size = 11$.

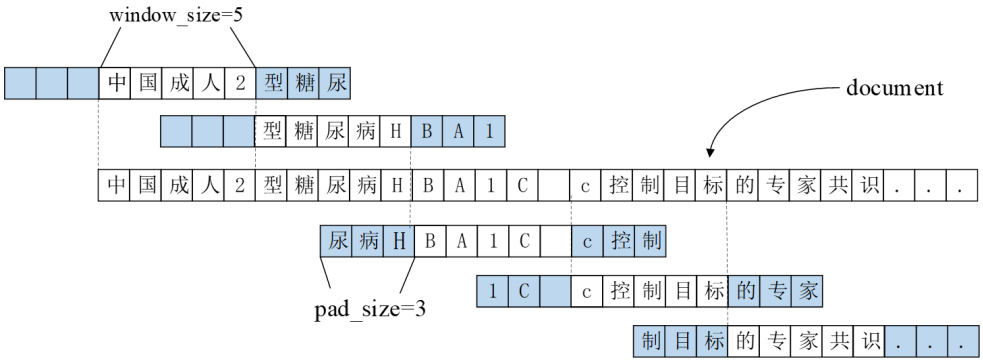

Figure 1. Schematic diagram of a sliding window.

The sentences obtained from the above division are labeled in BIO mode, and the characters in the sequence are labeled with "B-X", "I-X", or "O". The label "B-X" means the

corresponding character is at the beginning of the entity, "I-X" means the corresponding character is in the middle of the entity, and "O" means the character does not belong to any entity.

3.2. Model

The overall architecture of the proposed model named BIBC for NER is shown in Figure 2. As is shown by the figure, the whole model structure consists of four parts, namely BERT, IDCNN, bidirectional LSTM, and CRF. A modified BERT-WWM model is used to train on a massive unlabeled corpus for feature supplementation, which greatly enhances the semantic representation between words or phrases through deep and continuous learning. In addition, the IDCNN layer is added to make the model take into account both local and global features by IDCNN and bidirectional LSTM, thus making the whole model achieve better results on diabetes medical literature. The specific implementation process of the BIBC-based named entity recognition model consists of the following parts:

1. Firstly, the sentences are input to the BERT-WWM pre-trained language model based on whole-word masking for pre-training to obtain word vectors that better match the Chinese expressions.
2. The obtained vectors are then sent into the IDCNN layer to obtain the local information of the sentences.
3. The vector sequences obtained from IDCNN layer become the input to BiLSTM to further encode the global sequence features of the sentences.
4. Finally, the entities are labeled using the label transfer function of CRF and the feature information extracted by each network to obtain a globally optimal label sequence.

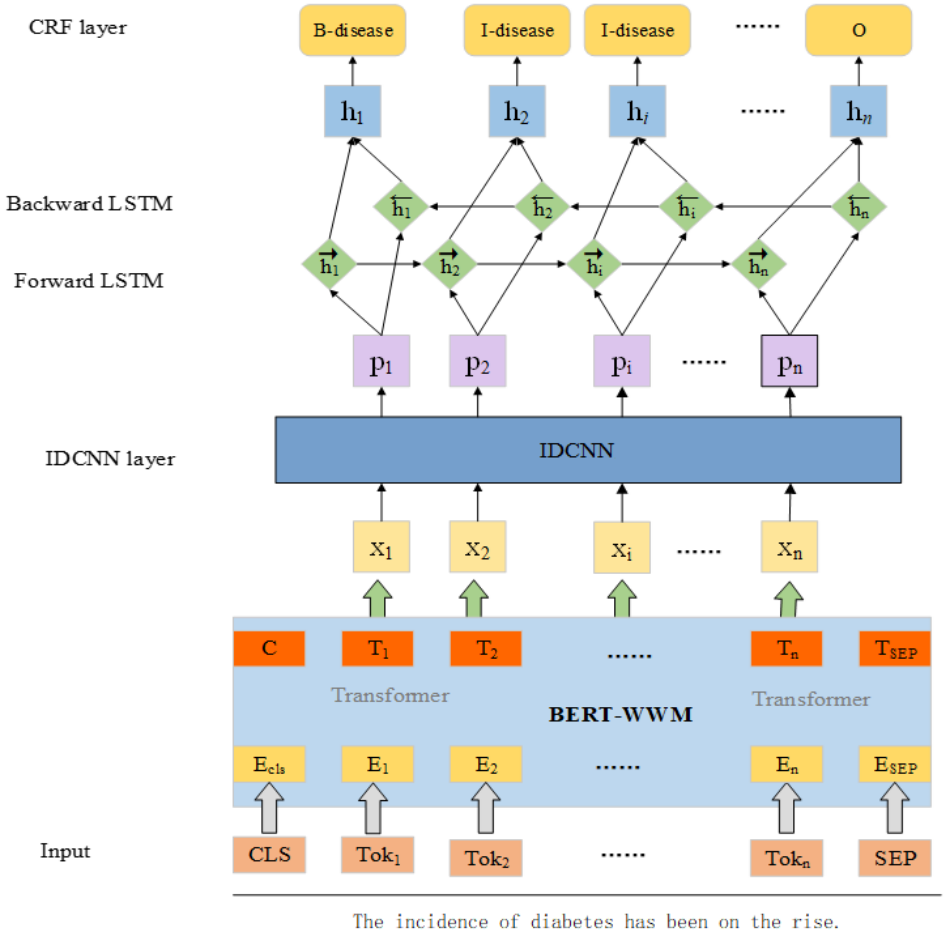

Figure 2. The model structure overview of BIBC.

3.2.1. Pre-Trained Language Model BERT

The BERT model uses a deep bi-directional Transformer network structure. For each word in the sentence, it can well calculate the interrelationship and importance with other words, and then obtain a new feature of a word. This feature vector contains more relationships and weights, so it can represent the polysemy of words.

BERT precedes each input sentence with a [CLS] marker symbol, which is used to represent the features of the whole sequence of the input, and we use this representation in the downstream target task as well. Between two sentences is the [SEP] symbol, which is used to split the two sentences. BERT is able to obtain some temporal features through the position vector, so that features are learned for different word contexts, and finally the multisense words are represented by different vectors.

One of the tasks of BERT pre-training is Masked language model, which masks a word in a sequence and then predicts it according to the context. The smallest token in Chinese data are a word, and a word consists of many Chinese characters, which contains a lot of information, so, for Chinese data, it is often necessary to mask a word and then predict it. In this part, we follow Cui et al. [12] using BERT-WWM, a derivative model of BERT. The model makes improvements to alleviate the shortcomings of BERT in pre-training by masking only part of the tokens, and uses a full word masking method in Chinese text.

3.2.2. IDCNN Layer

Dilated Convolutional Neural Networks (DCNN) [30] were initially applied in the field of image segmentation, and it mainly changes the use of convolutional kernels of CNNs. The DCNN expands the perceptual field of view by adding a width to the convolutional kernel, ignoring only the information on the convolutional kernel, and no longer performs convolutional operations on the continuous region of the input matrix like the ordinary CNN's convolutional kernel.

As is shown in Figure 3a, it is a 3×3 1-dilated convolution, which convolves on a contiguous region of the input data like the general CNN convolution kernel; Figure 3b corresponds to a 2-dilated convolution, where the size of the convolution kernel remains unchanged, but the width is increased, so the perceptual field of view is widened as well. The 4-dilated convolution shown in Figure 3c increases the perceptual field of view while keeping the convolution kernel unchanged. In the three figures, only the data in the nine red dot regions are focused on, and the weight of the data in the middle part is set to 0.

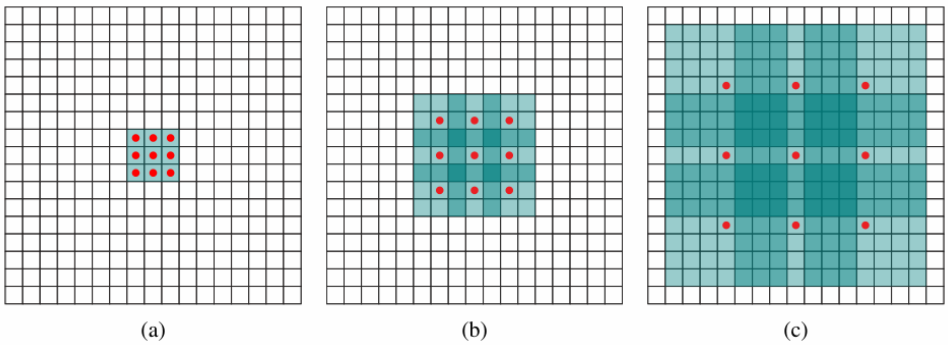

Figure 3. Schematic (**a**) is a 3×3 1-dilated convolution (**b**) is a 3×3 2-dilated convolution (**c**) is a 3×3 4-dilated convolution diagram of DCNN.

From our analysis, the convolutional operation of traditional CNN, when stride is set to 1, the three layers of 3×3 convolutional kernel can act in a range of $(kernel - 1) \times layer + 1 = 7$, and the perceptible area and the number of layers are linearly related. However, in DCNN, the convolutional kernel with width has an exponential relationship between the perceived area and the number of layers, and the perceived field of view expands rapidly. In this way, more information can be obtained with as few layers

as possible, solving the problem that ordinary CNNs can only obtain a small amount of information.

In the sequence labeling task, any word in the sentence may have an impact on the current word. CNN is to cover more information in the sequence by increasing the number of layers of convolutional layers, making the network structure more and more complex. Dropout, for example, is also used to prevent overfitting from bringing more hyperparameters, which makes the whole model training even more difficult. DCNN makes the perceptual field of view larger without increasing parameters, and solves the problem of too many parameters without information loss. Generally stacking Dilated CNNs can obtain more information, but this method leads to overfitting, so we choose IDCNN (Iterated Dilated CNN) as an alternative. IDCNN applies the same DCNN structure several times to increase the generalization ability of the model by reusing the same parameters in a cyclic manner. IDCNN structure is composed of four identical Dilated CNN blocks stitched together, where each Dilated CNN block has three convolutional layers inside with dilation widths of 1, 1, and 2, respectively.

In the NER model proposed in this paper, sentences are first pre-trained by BERT-WWM to obtain word vectors, and then input to Iterated Dilated CNN. IDCNN extracts local feature information of sentences by convolutional operations, and the entire perceptual field of view grows exponentially due to the increased width of the convolutional kernel, which reduces a large number of features without losing parameters and increases the training speed of the model. The feature vector with local information obtained by IDCNN is used as the input to BiLSTM to further extract the features of the whole sequence. In the internal IDCNN structure, the word vectors obtained from BERT training are firstly input into the first layer of DCNN, and the results of the first layer processing are output to the second layer of DCNN and other DCNN layers, respectively, and finally the output feature vectors are spliced.

3.2.3. BiLSTM Layer

LSTM can only obtain the forward information of the sentence, but most of the time the backward dependency of the sentence also carries a lot of information, so using BiLSTM can obtain both the forward and backward information of the sentence, and each layer of it has the same structure as LSTM, only the order of calculation is different. Thus, each word can contain the complete contextual information, and the network structure of BiLSTM is shown in Figure 4.

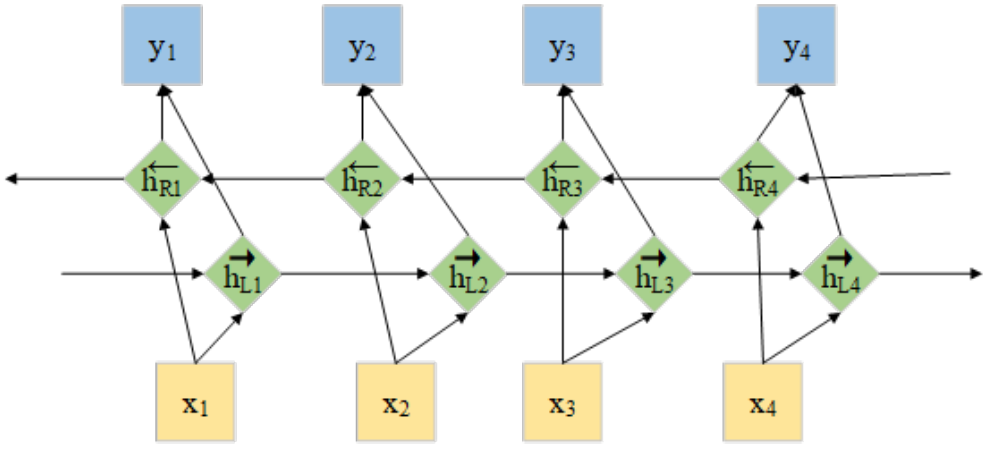

Figure 4. Neural network structure diagram of BiLSTM.

The input sentence sequence is sent into the forward LSTM to obtain a feature vector with sentence forward dependencies $\vec{h_l} = \{h_{L1}, h_{L2}, ..., h_{Ln}\}$. The feature vector with sentence backward dependencies obtained from the reverse LSTM is denoted as

$\vec{h_r} = \{h_{R1}, h_{R2}, ..., h_{Rn}\}$. The acquired bi-directional feature vectors are stitched together to obtain the new vectors:

$$Y = \{[h_{L1}, h_{R1}], [h_{L2}, h_{R2}], ..., [h_{Ln}, h_{Rn}]\} = \{y_1, y_2, ..., y_n\} \quad (1)$$

Vector Y in Equation (1) carries sentence-level contextual information and then it is input to CRF to obtain the sequence of sentence labels.

3.2.4. CRF Layer

We add the CRF model to solve the dependency problem between labels by adding many mandatory constraints between the output sequence labels. For instance, $I - Drug$ can only be the middle part of the drug entity instead of the beginning of the entity, and the label after it cannot be $I - Reason$, O is used to indicate that the character does not belong to any entity, etc. For input sentence $X = \{x_1, x_2, ..., x_n\}$, the probability matrix is denoted by P, and the fractional matrix is the output of BiLSTM. P is a matrix of $n \times k$, where k denotes the total number of the labels and $P_{i,j}$ denotes the probability that the i-th word in the sentence is labeled as the j-th tag. For output sequence $y = \{y_1, y_2, ..., y_n\}$, the score can be calculated from Equation (2):

$$s(X, y) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i} \quad (2)$$

where A denotes the label transfer matrix and $A_{y_i, y_{i+1}}$ denotes probability of label i transfer to label $i + 1$. A is a 1×2 matrix with start flag y_0 and end flag y_{n+1} added to the label set. The probability of y is calculated by softmax for all possible tags in the sentence, as shown in Equation (3):

$$p(y|X) = \frac{e^{s(X,y)}}{\sum_{\tilde{y} \in Y_X} e^{s(X,\tilde{y})}} \quad (3)$$

In the training process, maximizing the log probability of correctly labeled sequences is used as the optimization objective which is calculated in Equation (4):

$$\log(p(y|X)) = s(X|y) - log(\sum_{\tilde{y} \in Y_X} e^{s(X,\tilde{y})}) \quad (4)$$

where Y_X denotes all possible label sequences for a given sequence X (including labeled sequences that do not conform to the BIO format). The final output sequence obtained uses the Vibit algorithm to calculate the label sequence with the largest score following Equation (5):

$$y^{\star} = \arg\max_{\tilde{y} \in Y_X} s(X, \tilde{y}) \quad (5)$$

Finally, a dynamic optimization algorithm is used in the prediction phase to solve for the optimal sequence of labels $y^{\star}$.

In our proposed NER model, the input sentences are first pre-trained with BERT-WWM to generate a corresponding pre-trained word embedding from the sentences. Afterwards, this word embedding will be sent to the Iterated Dilated CNN to extract multiple local feature information through inflated convolution to obtain the local feature vector of the sentences. To further extract the global features of the whole sentence, BiLSTM is used to obtain the global features of the sentence from the sentence local feature vector combined with the context, and then obtain the classification probability of each word. Finally, it is plugged into the CRF layer containing the sequence transfer probabilities, and constraints are added to the final prediction labels to ensure the correctness of the final annotation results. The input sentence is processed as described in Figure 5 in the proposed BIBC model.

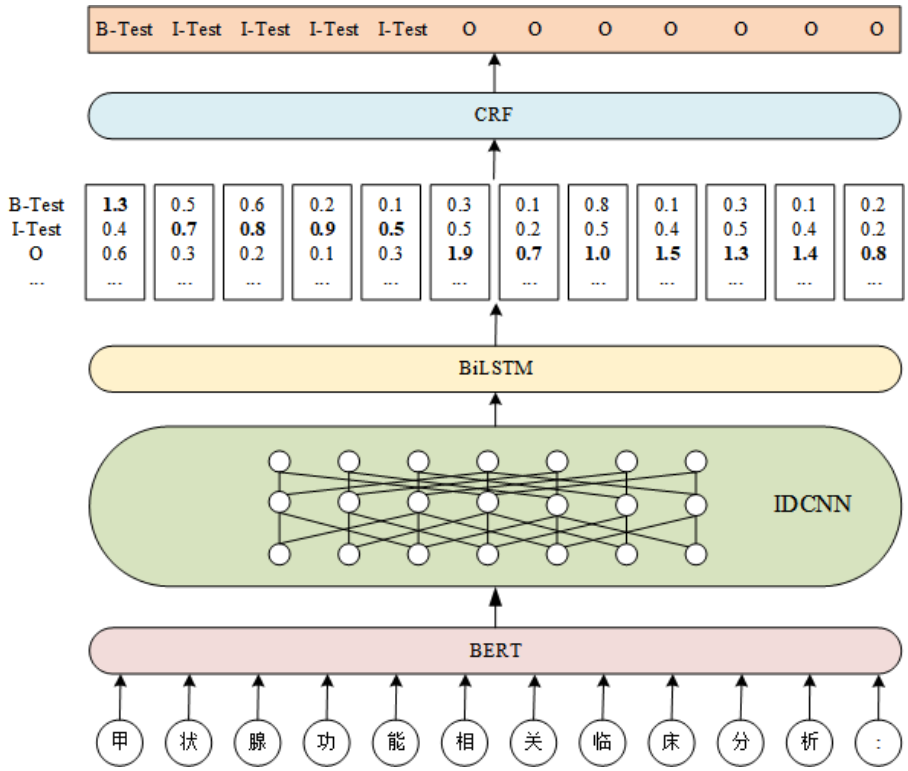

Figure 5. Input example of the BIBC model.

4. Experiment

4.1. Evaluation Metrics

In the task of named entity recognition, precision (P), recall (R), and F1-value ($F1$) are usually used to evaluate the performance of the model.

Precision is a relatively intuitive evaluation that indicates the percentage of correct results among all the results obtained and is calculated as shown in Equation (6):

$$P = \frac{TP}{TP + FP} \tag{6}$$

Recall represents the proportion of correct results identified among all correct results, and the coverage of the model identification results is measured and calculated as shown in Equation (7):

$$R = \frac{TP}{TP + FN} \tag{7}$$

In order to evaluate the performance of the model, the $F1$ value is combined with the precision and recall to make a comprehensive evaluation of the results, which is calculated as shown in Equation (8):

$$F1 = \frac{2 \times P \times R}{P + R} \tag{8}$$

where TP denotes the number of correct predicted results, FP denotes the number of incorrect predicted results, and FN denotes the number of results that were not predicted among all correct results in the data set.

4.2. Experiment on the Data Preprocessing Method

In this paper, we use the sliding window segmentation method to process the Ruijin dataset in the NER task, and experimentally verify that this method can actually improve the entity recognition results compared to the segmentation method using punctuation.

In particular, different *window_sizes* are applied for the text segmentation for comparison between different settings.

As shown in Figure 6, the validation on the BERT-BiLSTM-CRF model shows that a *window_size* value around 180 is relatively good in F1 score. If the sliding window value is too small, the sentence contains too little contextual information, which will lead to the degradation of the effect; and, if the value is large, the sentence length is too large, which is not conducive to the training of the model, so, in this paper, we select a *window_size* of 180. Our analysis of the data characteristics shows that the maximum length of the entity in the dataset is 14. In order to avoid the entity being divided and retain some contextual information, the *pad_size* is set to 30, so the size of the divided sentences is $window\_size + 2 * pad\_size = 240$.

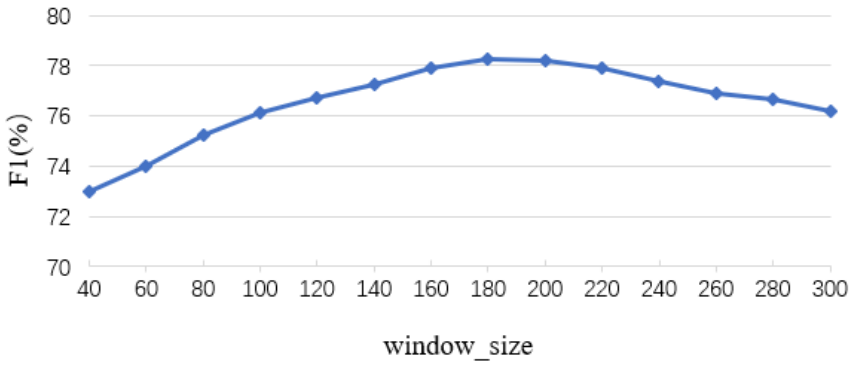

Figure 6. The experimental results of different *window_size*.

In order to demonstrate the effect of different division methods on the results, comparison experiments were conducted on the models BiLSTM_CRF and BERT_BiLSTM_CRF, respectively, and the specific experimental results are shown in Figure 7.

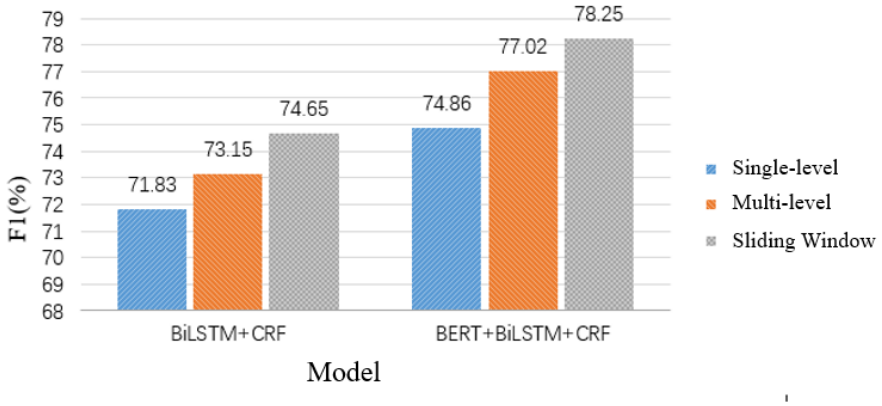

Figure 7. NER results based on different partition methods.

As the experimental comparison using different data processing methods on different models can be seen from Figure 7, the method using sliding windows is better than the one using punctuation symbols to divide the data, and the worst results are one level of punctuation symbols to divide the data. It is because the part of sentences obtained by using a single level of punctuation symbols division is too long, which affects the training performance of the model. In addition, the direct interception of sentences with too large length will lead to a decrease in information capture. The data set is not very standardized because of the format conversion, and using multi-level punctuation may divide two entities in two sentences, resulting in the entity segmentation not being recognized correctly.

All the data used in the following models are experimentally validated using the processing that works best.

4.3. Experiment on BIBC

4.3.1. Experiment Settings

Our experiments use data obtained by the sliding window partitioning method to complete the proposed NER task. We use Bayesian optimization in scikit-optimize and select the best combination of parameter values by nested k-fold cross-validation. The final determined parameter settings are shown in Table 2.

Table 2. Model parameter configuration.

Parm Type	Parm Value
Max seq length	256
Batch size	64
Learning rate	5×10^{-6}
Epoch	20
Dropout	0.5
Clip	5
filter	100
Bi-LSTM Hidden Layer Size	128

In order to verify the effectiveness of the proposed method in this paper for NER on the Ruijin diabetes dataset, this section conducts comparative experiments using the partial model and the overall model respectively to verify the effect of each part of the model on the results.

4.3.2. Results and Analysis

The results of our ablation studies performed on the diabetes dataset are shown in the following Table 3.

Table 3. Comparison of entity recognition results.

Model	Precision	Recall	F1-Score
CRF	68.31	66.93	67.61
LSTM	70.82	72.32	71.56
BiLSTM	72.35	73.09	72.72
BiLSTM-CRF	73.41	75.93	74.65
IDCNN-CRF	72.95	73.47	73.21
BERT-CRF	76.53	77.85	77.18
BERT-BiLSTM-CRF	78.35	78.15	78.25
BERT(wwm)-BiLSTM-CRF	79.01	78.83	78.92
Our model	79.58	80.21	79.89

From the result, we can discover that single models are generally less effective than composite models. Among the three single model, deep learning based methods have better results than statistical methods, namely CRF. In addition, BiLSTM achieves the best performance in single models which means that bidirectional context information does improve the effectiveness. As is shown in Table 3, BiLSTM-CRF, a mainstream model, works better than single models. This is because each layer of the composite model focused on different features. For example, BiLSTM-CRF compared with Bi-LSTM has an additional CRF layer, the main purpose of which was to learn the constraints of the sentence and reduce the sequence of incorrect predictions. Meanwhile, the introduction of BERT increases the score by 2.53%. This comparison between BiLSTM-CRF and BERT-BiLSTM-CRF indicates that the word vectors obtained from BERT are relatively better than vectors obtained from Word2vec, and the problem of polysemy is alleviated. Thus, the

pre-trained model has already extracted the linguistic knowledge and encoded it into the network, which can introduce a large amount of external knowledge from unsupervised data compared to the other two methods.

The result of IDCNN-CRF surpasses that of BiLSTM-CRF, while BiLSTM is capable of extracting more distant dependency information and CNN focuses more on local information. It shows that the effect of using convolutional neural network alone is not as good as using bidirectional LSTM. The adjustment of the masking method in the modified BERT(BERT-WWM) makes the model work better than BERT-base, which means, in Chinese NLP tasks, that the whole word masking setting can effectively improve the model performance. The improvement achieved in our work could benefit the research on diabetes and other related medical fields.

The accuracy, recall, and F1 values of the proposed BIBC model are improved by 0.57%, 1.38%, and 0.97%, respectively, compared with BERT(wwm)-BiLSTM-CRF. This experimental result indicates that BERT combined with IDCNN can obtain more local information, BiLSTM can obtain long-range contextual information, and the final feature vector contains a large amount of semantic information of the text, which improves the effect of the final NER task. In summary, the BIBC-based named entity recognition model works better than other aforementioned statistical or deep learning models, and the effectiveness of the model is verified through our experiments.

In the phrase "A blood glucose level below 3.9 mmol/L (70 mg/d1) in people with diabetes is known as hypoglycaemia", the BIBC identifies "hypoglycaemia" as a Disease, but it is actually classified as a Symptom. The reason for this may be that there is some ambiguity in the entity and the presence of "is known as" makes it appear that the entity "hypoglycaemia" can also be classified as a Disease. Another reason is that the entity of "hypoglycemia" belongs to multiple categories. For example, in the sentence "Geriatric patients should have individualized glucose-lowering goals, taking into account their functional status, co-morbidities or co-morbidities, and especially the occurrence of cardiovascular disease, hypoglycemia, and microvascular complications", the "hypoglycaemia" entity is classified as a Disease. A similar situation exists for other entities in this dataset, such as "Insulin" entity may belong to the categories of Reason, Anatomy, Drug, Method, or Amount. This shows that the BIBC model is still lacking in understanding the contextual background, and in the future we can improve the model effect by introducing more feature information such as lexicality and external description of the corresponding language in the context of identifying the entity language category.

5. Conclusions and Future Work

In this paper, we use the current mainstream model BiLSTM-CRF as the base model, and improve on this model to enhance the performance of NER. To address the polysemy phenomenon, we further introduce BERT to train the word vector, which alleviates this problem and also solves the lack of medical data and the non-standardized annotation problem. The performance of NER is greatly improved by using BERT combined with BiLSTM-CRF on the diabetes dataset. We proposed the BIBC model to reduce training parameters and expand the perceptual field, which ensures the combination of both local and global features.

Due to the raw format of the dataset, it took a great amount of work in data preprocessing. The imperfections in this process remain to be added to our study in the future, such as a more precise method to divide the sentences and to consider the sentence grammar, lexicality, and external description, especially medical expert knowledge, etc.

Author Contributions: Conceptualization, Y.D. and L.Y.; investigation, Y.F.; methodology, Y.D. and Y.F.; supervision, Y.D. and L.Y.;visualization, Y.F. and L.Y.; writing—original draft preparation, Y.D., Y.F. and L.Y.; writing—review and editing, Y.D., Y.F. and L.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The model is trained on the Ruijin diabetes dataset, which comes from authoritative journals in the field of diabetes in Chinese. The time span reaches seven years, covering the most extensive research content and hotspots in the field of diabetes in recent years. The annotators of the dataset all have medical backgrounds. This dataset aims to do diabetes literature mining and build a diabetes knowledge graph through diabetes-related textbooks and research papers. In the Tianchi Ruijin Hospital MMC Ai-aided Knowledge Map challenge, the organizing committee provides 362 labeled canto-level data as the training set with pre-defined categories, including disease, reason, symptom, test, test value, drug name, frequency, amount, method of administration, non-drug therapy, operation, location, severity, and duration of adverse reactions, and each canto-level data contain about 480 annotated data on average. In addition, the organizing committee provides another 117 unlabeled canto-level data as the training set for the model evaluation. In our experiments, we split the 362 labeled canto-level data into two partitions, 326 canto-level data as the training set and 36 canto-level data as the validation set, and the 117 unlabeled canto-level data as the test set. Data link: https://tianchi.aliyun.com/dataset/dataDetail?dataId=88836, accessed on 21 January 2021.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Grishman, R.; Sundheim, B. Message Understanding Conference 6: A Brief History. In Proceedings of the 16th International Conference on Computational Linguistics, Copenhagen, Denmark, 5–9 August 1996; Volume 96, pp. 466–471. [CrossRef]
2. Krupke, G.; Hausman, K. Isoquest Inc: Description of the NetOwl(TM) extractor system as used for MUC7. In Proceedings of the Seventh Message Understanding Conference (MUC-7), Fairfax, VA, USA, 29April–1 May 1998.
3. Yadav, V.; Bethard, S. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. *arXiv* **2019**, arXiv:1910.11470.
4. Wang, Y.; Sun, Y.; Ma, Z.; Gao, L.; Xu, Y.; Sun, T. Application of Pre-training Models in Named Entity Recognition. *arXiv* **2020**, arXiv:2002.08902.
5. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. In Technical Report, OpenAI. 2018. Available online: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf (accessed on 8 October 2021).
6. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
7. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
8. Liu, C.L.; Jin, G.; Liu, Q.; Chiu, W.Y.; Yu, Y.S. Some Chances and Challenges in Applying Language Technologies to Historical Studies in Chinese. *arXiv* **2012**, arXiv:1210.5898.
9. Cui, Y.; Che, W.; Liu, T.; Qin, B.; Wang, S.; Hu, G. Revisiting Pre-Trained Models for Chinese Natural Language Processing. *arXiv* **2020**, arXiv:2004.13922.
10. Li, X.; Meng, Y.; Sun, X.; Han, Q.; Li, J. Is Word Segmentation Necessary for Deep Learning of Chinese Representations? *arXiv* **2019**, arXiv:1905.05526.
11. Zhang, Z.; Han, X.; Liu, Z.; Jiang, X.; Sun, M.; Liu, Q. ERNIE: Enhanced Language Representation with Informative Entities. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019.
12. Cui, Y.; Che, W.; Liu, T.; Qin, B.; Yang, Z.; Wang, S.; Hu, G. Pre-Training with Whole Word Masking for Chinese BERT. *arXiv* **2019**, arXiv:1906.08101.
13. Chiu, J.; Nichols, E. Named Entity Recognition with Bidirectional LSTM-CNNs. *Comput. Sci.* **2015**._a_00104.
14. Liu, Z.; Yang, M.; Wang, X.; Chen, Q.; Tang, B.; Wang, Z.; Xu, H. Entity Recognition from Clinical Texts via Recurrent Neural Network. *BMC Med. Inform. Decis. Mak.* **2017**, *17*, 53–61. [CrossRef] [PubMed]
15. Hwang, K.; Sung, W. Single stream parallelization of generalized LSTM-like RNNs on a GPU. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015. [CrossRef]
16. Gao, Y.; Chen, Y.; Wang, J.; Lu, H. Reading Scene Text with Attention Convolutional Sequence Modeling. *arXiv* **2017**, arXiv:1709.04303.
17. Ma, X.; Hovy, E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. *arXiv* **2016**, arXiv:1603.01354.
18. Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv* **2015**, arXiv:1508.01991.
19. Zhang, Y.; Yang, J. Chinese NER Using Lattice LSTM. *arXiv* **2018**, arXiv:1805.02023

20. Qiang, B.H.; Huang, J.; Wang, Y.F.; Wang, S.; Wang, Y. Research on Chinese named entity recognition using combined boundary-PoS feature. In Proceedings of the 2015 International Conference on Design, Manufacturing and Mechatronics (ICDMM2015), Wuhan, China, 17–18 April 2015; pp. 839–848. [CrossRef]
21. Dong, X.; Qian, L.; Guan, Y.; Huang, L.; Yu, Q.; Yang, J. A multiclass classification method based on deep learning for named entity recognition in electronic medical records. In Proceedings of the 2016 New York Scientific Data Summit (NYSDS), New York, NY, USA, 14–17 August 2016; pp. 1–10. [CrossRef]
22. Ding, R.; Xie, P.; Zhang, X.; Lu, W.; Si, L. A Neural Multi-digraph Model for Chinese NER with Gazetteers. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019.
23. Strubell, E.; Verga, P.; Belanger, D.; Mccallum, A. Fast and Accurate Entity Recognition with Iterated Dilated Convolutions. *arXiv* **2017**, arXiv:1702.02098.
24. Zhang, X.; Zhang, Y.; Zhang, Q.; Ren, Y.; Qiu, T.; Ma, J.; Sun, Q. Extracting comprehensive clinical information for breast cancer using deep learning methods. *Int. J. Med Inform.* **2019**, *132*, 103985. [CrossRef] [PubMed]
25. Li, L.; Zhao, J.; Hou, L.; Zhai, Y.; Cui, F. An attention-based deep learning model for clinical named entity recognition of Chinese electronic medical records. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 235. [CrossRef] [PubMed]
26. Ji, B.; Liu, R.; Li, S.; Yu, J.; Wu, Q.; Tan, Y.; Wu, J. A hybrid approach for named entity recognition in Chinese electronic medical record. *Bmc Med. Inform. Decis. Mak.* **2019**, *19*, 149–158. [CrossRef] [PubMed]
27. Li, X.; Zhang, H.; Zhou, X.H. Chinese Clinical Named Entity Recognition with Variant Neural Structures Based on BERT Methods. *J. Biomed. Inform.* **2020**, *107*, 103422. [CrossRef] [PubMed]
28. Li, Y.; Ma, Q.; Wang, X. Medical Text Entity Recognition Based on CRF and Joint Entity. In Proceedings of the 2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC), Dalian, China, 14–16 April 2021; pp. 155–161. [CrossRef]
29. Zhou, H.; Liu, Z.; Lang, C.; Xu, Y.; Hou, J. Improving the recall of biomedical named entity recognition with label re-correction and knowledge distillation. *BMC Bioinform.* **2021**, *22*, 1–16. [CrossRef] [PubMed]
30. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. *arXiv* **2015**, arXiv:1511.07122.