Chinese Named Entity Recognition Method in History and Culture Field Based on BERT

Shuang Liu¹ · Hui Yang² · Jiayi Li¹ · Simon Kolmanič³

Received: 3 August 2020 / Accepted: 30 August 2021
© The Author(s) 2021

Abstract

With rapid development of the Internet, people have undergone tremendous changes in the way they obtain information. In recent years, knowledge graph is becoming a popular tool for the public to acquire knowledge. For knowledge graph of Chinese history and culture, most researchers adopted traditional named entity recognition methods to extract entity information from unstructured historical text data. However, the traditional named entity recognition method has certain defects, and it is easy to ignore the association between entities. To extract entities from a large amount of historical and cultural information more accurately and efficiently, this paper proposes one named entity recognition model combining Bidirectional Encoder Representations from Transformers and Bidirectional Long Short-Term Memory-Conditional Random Field (BERT-BiLSTM-CRF). First, a BERT pre-trained language model is used to encode a single character to obtain a vector representation corresponding to each character. Then one Bidirectional Long Short-Term Memory (BiLSTM) layer is applied to semantically encode the input text. Finally, the label with the highest probability is output through the Conditional Random Field (CRF) layer to obtain each character's category. This model uses the Bidirectional Encoder Representations from Transformers (BERT) pre-trained language model to replace the static word vectors trained in the traditional way. In comparison, the BERT pre-trained language model can dynamically generate semantic vectors according to the context of words, which improves the representation ability of word vectors. The experimental results prove that the model proposed in this paper has achieved excellent results in the task of named entity recognition in the field of historical culture. Compared with the existing named entity identification methods, the precision rate, recall rate, and F_1 value have been significantly improved.

Keywords History and culture · Named entity recognition · Bert pre-training model · Bidirectional long short-term memory network

1 Introduction

With the rapid development of the Internet, people's lifestyles and ways of understanding Chinese history and culture are also changing. More and more provinces in China begin to pay attention to the development of history and culture and the construction of online cultural information platform. With more and more emergence of intelligent museums and digital museum, the online information platform of network history and culture has attracted more and more attention for the public. More and more scholars began to study in this field [1]. Since China has five thousand years of history and culture, historical culture has become an indispensable part of our lives [2]. But facing massive online historical and cultural data, how to automatically extract potential knowledge from massive unstructured text data, how to extract the relevant content of historical culture and how to organize

✉ Hui Yang
yanghuimail2020@gmail.com
Shuang Liu
liushuang@dlnu.edu.cn
Jiayi Li
ygrace788@gmail.com
Simon Kolmanič
simon.kolmanic@gmail.com

¹ School of Computer Science and Engineering, Dalian Minzu University, Dalian 116600, China

² School of Hotel Management, Nanjing Institute of Tourism and Hospitality, Nanjing 211100, China

³ Faculty of Electrical Engineering and Computer Science, University of Maribor, Koroska cesta 46, 2000 Maribor, Slovenia

this useful information are key challenging problems. To extract useful information from extensive data and mine its potential internal value, natural language processing (NLP) technology is usually adopted [3]. The first step of knowledge extraction in natural language processing (NLP) is named entity recognition (NER), which is a key step in information extraction. It is also a very important step in the process of building the knowledge graph in the field of history and culture.

Named entity recognition refers to identifying specific entity information in text, such as names of people, places, organizations, etc. The recognition result has a significant influence on subsequent tasks such as information extraction, information retrieval, machine translation, automatic question answering, and knowledge graph. The identification of named entities in the field of historical culture is an important part of building a historical and cultural knowledge graph, which is also the focus of this paper.

In the research of this paper, our goal is to identify entities of several historical and cultural types, including the names of historical dynasties, historical characters, historical time, historical locations, and official titles. These entities are more important knowledge information in the discovery of historical and cultural knowledge. Among them, the task of entity recognition of historical dynasties is to find the dynasty where the characters are located in the text. The historical character name recognition task is to recognize the character names in the text. The historical event entity recognition task is to identify the time of occurrence in the text. The historical location entity recognition task is to recognize the location name in the text. The official title recognition task is to identify the official positions of the characters identified in the text.

For example:

“617年唐国公李渊于晋阳 起兵， 次年称帝建立唐朝， 定都长安。

(Which means: In 617 years, Tang Guogong Li Yuan started his army in Jinyang. The following year he was called the emperor to establish the Tang Dynasty and set Chang'an as the capital.) ” In this article, “617 years” is the historical time, “Tang Guogong” is the title of official position, “Liyuan” is the name of historical figure, “Jinyang” and “Chang'an” are the historical places and “Tang Dynasty” is the name of historical dynasty.

The purpose of this paper is to construct a text data set in the historical and cultural field and then use these labeled data to effectively identify a large amount of unlabeled historical and cultural text data, and pave the way for the future construction of knowledge graph in the historical and cultural field. The main contribution of our work can be summarized as follows: (1) The data set of named entity

recognition in the field of ancient history and culture is constructed by acquiring the open historical and cultural text data on the Internet. (2) The named entity recognition model composed of BERT pre-trained language model, bidirectional long-term short-term memory (BiLSTM) and conditional random field (CRF) is applied to the field of ancient Chinese history and culture for the first time. (3) Constructed a knowledge graph of ancient Chinese history and culture, which mainly includes information such as names of historical dynasties, names of historical figures, historical times, historical locations, and official titles.

The rest of the paper is organized as follows. Section 2 introduces the related work in named entity recognition. Section 3 describes the method proposed in this article in detail. Section 4 introduces the data set, environment and parameter settings of the experiment in the paper. Section 5 lists the experimental results with analysis and comparison. Section 6 summarizes the paper.

2 Related Work

Many researchers have done a related study on named entity recognition (NER) from methods rule-based, based on statistical machine learning to methods based on deep learning.

The rule-based approach was first proposed in the named entity recognition task. To get named entities when recognizing the required text, this method is customized by experts in related fields combining with their own expertise. The basic idea is to match the text to be identified with a pre-defined rule template to get named entities, which is created through the expert's domain knowledge and domain-related dictionaries.

The method based on statistical machine learning builds a model by fusing language models and statistical machine learning algorithms [2], Such as Conditional Random Field (CRF) [4], etc. In traditional machine learning, CRF is regarded as the mainstream model for named entity recognition. The advantage is that CRF can use internal and contextual feature information in the process of labeling a location [5]. In the above-mentioned method based on statistical machine learning, a large amount of manual annotation training data is also required for feature extraction. Moreover, if the amount of manually labeled data is small during model training, the effect of the trained model may not be ideal [6].

In recent years, a large number of deep learning methods have been applied to named entity recognition. Named entity

recognition based on neural network is usually regarded as a task of sequence annotation, which has achieved excellent results. For example, Hammerton proposed to use unidirectional long short-term memory (LSTM) to solve the problem of named entity recognition [7]. Collobert et al. proposed for the first time to combine Convolutional Neural Networks (CNN) and CRF to conduct experiments on named entity recognition datasets in the general domain, and achieved good results [8]. In this method, each word has a fixed-size window, but it fails to consider the effective information between long-distance words. Santos et al. proposed to use the character CNN to expand the CNN-CRF model [9]. Huang et al. [10] designed a spelling feature BiLSTM-CRF model manually and achieved good results on the CONLL2003 dataset, with F_1 values reaching 88.83%. Chiu and Nichols [11] proposed a bidirectional LSTM-CNNs model, which can automatically detect the word and character-level features, with F_1 values reaching 91.62% on the CoNLL2003 dataset. Lample et al. [12] proposed a BiLSTM-CRF structure to obtain effective information from character-based embedding. Ma et al. [13] proposed a model based on LSTM-CNNs-CRF to deal with sequence labeling problems. By combining LSTM, CNN and CRF models, an end-to-end model without a large number of specific task knowledge, feature engineering and preprocessing corpus was established. Strubel et al. [14] proposed the Iterated Dilated Convolution Neural Networks (IDCNN) for named entity recognition, which improved the traditional CNN network, introduced regularization, and solved the overfitting problem caused by the increase of the number of layers of the traditional CNN network. Zhang et al. [15] proposed a new type of Lattice LSTM, which integrated potential word information into the traditional LSTM-CRF based on the word model. So the potential word information is obtained through an external dictionary. Experiments on multiple public data sets have achieved good results.

In the field of history and culture, there are also some research related to named entity recognition. For example, Liu et al. [16] used a method based on language model and conditional random field to study the recognition of named entities in the study of literature and history. The data used in its experiment comes from the book "Local Chronicles". "Local Chronicles" is a huge and most important document collection in Chinese history, which contains information about local government officials [16]. The experimental results showed that the information extracted from the text was basically consistent with the content in the Harvard Chinese Biography Database (CBDB).

Sie et al. [17] studied the development of a text retrieval and mining system for Taiwanese historical people and introduced the development of the Taiwan Historical People Database (TBDB). They described the characteristics of characters in TBDB, highlighted the current preliminary

results of TBDB database, and proposed a named entity recognition method in biography.

Zhang et al. [1] proposed a semi-supervised cultural relic entity recognition model Semi-supervised model for Cultural Relics' Named Entity Recognition (SCRNER), which is composed of BiLSTM and CRF. The experimental data set is a collection of historical cultural relics data from the National Museum of China Online (<http://www.chnmuseum.cn/>) with manually annotations after preprocessing. In addition, Zhang et al. [1] also used the Embeddings from Language Models (ELMo) to represent the dynamically acquired words as the input of the model, which solved the problems of fuzzy boundary of Chinese objects and Chinese characteristics in the field of cultural relics. In their experiments, the F_1 value of the self-built dataset is 86.1%.

In the field of named entity recognition of other languages, Ajees et al. [19] proposed one improved BiLSTM-CRF model, which was successfully applied to the recognition of named entities in Indian achieving good results. Gorla et al. [20] used dictionary features, context, word level and corpus features to identify people's names, place names and organizational names of Southeast Asian languages such as Telugu.

In other fields, Yue et al. [18] proposed a Chinese named entity recognition method based on BERT in the field of police situation, which solved the problem that the key entity information in the police text is difficult to recognize. Seti et al. [21] proposed the use of character-level graph convolutional networks (GCN) to identify entity information in sports texts in the field of sports culture. The author conducted experiments on four datasets, and the results showed that the proposed method was effective and significantly increased the F_1 value. Dai et al. [22] compared several NER neural models, including BiLSTM, with two pre-training language models, word2vec and BERT, based on the data set offered by CCKS 2019 and the Chinese electronic medical records provided by the Second Affiliated Hospital of Soochow University. The experimental results show that the methods used in this paper are effective.

3 Models

One NER data set containing information about ancient Chinese history and culture is constructed in this paper, and BERT is applied to data sets to extract historical and cultural entities. Then performance of different models (IDCNN-CRF, BiLSTM-CNN-CRF and BiLSTM-CRF layer) on entity recognition in historical and cultural data sets are studied. Based on this, the fine-tuned BERT model with BiLSTM-CRF layer is selected as the final model for entity recognition. To find the most effective evaluation method of

entity recognition results, different evaluation indexes are put forward too.

3.1 Model Architecture

The BERT-BiLSTM-CRF model is mainly composed of three parts, namely the BERT pre-training language model, BiLSTM layer and CRF layer. The model structure is shown in Fig. 1. BERT pre-training language model is first used to encode a single character to obtain a vector representation corresponding to each character. Then, one BiLSTM layer is designed to semantically encode the input text. Finally, the label is output with the highest probability through the CRF layer, thereby obtaining each character Category. Details of each module, including BERT, LSTM, and CRF layers, are described below.

3.1.1 Bert

The most important part of the BERT model is the bidirectional Transformer coding structure. Transformer is an improvement to the coding-masking, and it chooses to replace the Recurrent Neural Network (RNN) with a form based on the attention mechanism. The Transformer coding unit is shown in Fig. 2.

In the attention mechanism, each word corresponds to three different vectors, namely Query vector (Q), Key

Fig. 1 BERT-BiLSTM-CRF model structure. “

“曹、丕、死、于、洛、阳

(means:Cao Pi died in Luoyang)” are the input character sequence of the model and “Cao Pi” is a person’s name and “Luoyang” is a place name. $p_1, p_2, \dots, p_6$ are the output of the BiLSTM hidden layer

Fig. 2 Transformer coding unit. explanation: the drawing is drawn according to reference [26]

vector (K), and Value vector (V). These three vectors are obtained by multiplying the embedding vector by three different weight matrices w_q, w_k, w_v . Then each word is scored by multiplying the Query vector and Key vector. Attention value is to use *softmax* to smooth the score item just obtained and then multiply the result with the Value vector.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \quad (1)$$

In addition, Transformer coding unit adds residual network and layer normalization to improve the degradation problem.

$$\text{LN}(x_i) = \alpha * \frac{x_i - \mu_L}{\sqrt{\sigma_L^2 + \epsilon}} + \beta, \quad (2)$$

$$\text{FNN} = \max(0, xW_1 + b_1)W_2 + b_2, \quad (3)$$

where α and β are the parameters to be learned, μ and σ are the mean and variance of the input.

The input of BERT model is mainly composed of three parts: Token embeddings, Segment embeddings, and Position embeddings. Among them, Token embeddings are the most important information about words in the model. Segment embeddings, or segment vectors, are used to distinguish two sentences. In addition, there is a [CLS] character at the beginning of a sentence and a [SEP] character at the end of a sentence. Position embeddings are used to remember the location information of words. The final input is the sum of three vectors. In the BERT-BiLSTM-CRF model, the input characters are first semantically represented by the BERT module. After the vector

representation of each word in the sentence is obtained, the sequence of the word vectors output by the BERT layer is input into the second BiLSTM module for semantics encoding.

3.1.2 BiLSTM Layer

LSTM is an improved version of RNN designed to solve the problem of vanishing gradients. Basically, an LSTM unit consists of three multiplication gates. These gates control the proportion of information forgotten and passed to the next time step. These gate control units are Input Gate, Forget Gate, and Output Gate. Figure 3 shows the basic structure of the LSTM unit.

The input of the forget gate is the hidden layer state $h_{(t-1)}$ at the previous moment, the input word x_t at the current moment, and the output value is f_t . The calculation formula is as follows:

$$f_t = \sigma(W_f h_{t-1} + U_f x_t + b_f). \quad (4)$$

The input gate input is the hidden layer state h_{t-1} at the previous moment, the input word x_t at the current moment, and the output value is i_t and $\tilde{C}_t$, the calculation formula as follows:

$$i_t = \sigma(W_i h_{t-1} + U_i x_t + b_i) \quad (5)$$

$$\tilde{C}_t = \tanh(W_c h_{t-1} + U_c x_t + b_c). \quad (6)$$

The state update of the memory unit consists of two parts. The first part is the product of C_{t-1} and the output of the forget gate f_t , and the second part is the i_t and $\tilde{C}_t$ of the input gate. The product of, the calculation formula is as follows:

$$C_t = C_{t-1} \odot f_t + i_t \odot \tilde{C}_t. \quad (7)$$

The input of the output gate is the hidden layer state h_{t-1} at the previous moment and the input word x_t at the current moment, the cell state at the current moment C_t , and the output value is o_t and the hidden layer state h_t , the calculation formula is as follows:

$$o_t = \sigma(W_o h_{t-1} + U_o x_t + b_o) \quad (8)$$

$$h_t = o_t \odot \tanh(C_t). \quad (9)$$

Finally, the hidden state sequence $[h_1, h_2, \dots, h_n]$ with the same sentence length will be obtained.

Among them, i_t is the input gate, f_t is the forget gate, $\tilde{C}_t$ is the new memory unit, c_t is the final memory unit, o_t is the output gate, and h_t is the hidden Layer, σ represents the *sigmoid* activation function, $\tanh$ represents the hyperbolic tangent activation function, $\odot$ represents the dot product of the corresponding elements, W_i, W_f, W_c, W_o represent the weight matrix of the hidden layer, b_i, b_f, b_c, b_o represents the deviation vector.

In the named entity recognition model based on the BERT-BiLSTM-CRF model in Fig. 1, the vector output through the BERT layer is input to the BiLSTM network for encoding, the forward LSTM network learns future features, and the reverse LSTM network learns historical features. At the moment t , the hidden layer state sequence in the forward and backward directions will be obtained.

The forward hidden layer state sequence is:

$$\vec{h}_t = [\vec{h}_1, \vec{h}_2, \vec{h}_3, \dots, \vec{h}_n]. \quad (10)$$

The state sequence of the backward hidden layer is:

$$\overleftarrow{h}_t = [\overleftarrow{h}_1, \overleftarrow{h}_2, \overleftarrow{h}_3, \dots, \overleftarrow{h}_n]. \quad (11)$$

Then, the vector output of the forward LSTM and the backward LSTM are combined to get the output result of the BiLSTM. details as follows:

$$[h_t = [\vec{h}_t : \overleftarrow{h}_t]. \quad (12)$$

In this paper, the entity recognition task is transformed into a sequence labeling task in the historical and cultural field. The data used in the experiment has many long sentences. The characters before and after the long sentence can form a concentrated semantic feature. Each entity mentioned in the text sequence can depend on the long-distance information text [1]. The bidirectional long short-term memory network can learn the output weight of the previous moment and the input of each sequence at the current moment. In addition, the forward network and the backward network in BiLSTM can simultaneously capture the forward and backward information of the sentence sequence to obtain context

Fig. 3 LSTM cell structure. explanation: the drawing is drawn according to reference [29]

information. Therefore, this method is used to capture all information during the modeling of long sentence sequences [23].

In short, LSTM is chosen to obtain long-distance dependence between entities in the historical text.

3.1.3 CRF Layer

After obtaining the hidden layer vector output by BiLSTM, the result of BiLSTM needs to be input into the CRF layer, and the label with the highest probability is output through the CRF layer to obtain the category of each character.

For the specified sequence $X = (x_1, x_2, x_3, \dots, x_n)$, the corresponding label is $y = (y_1, y_2, y_3, \dots, y_n)$, and the score is defined as:

$$S(X, y) = \sum_{i=0}^n A_{y_i, y_{i+1}} + \sum_{i=1}^n P_{i, y_i}. \quad (13)$$

Among them, A is the transfer score matrix, and A_{ij} represents the score transferred from label i to label j .

The maximum probability of the sequence label y can be calculated using the *softmax* function, namely:

$$P(y|X) = \frac{e^{S(X, y)}}{\sum_{\tilde{y} \in Y_X} e^{S(X, \tilde{y})}}, \quad (14)$$

$$\log(P(y^*|X)) = S(X, y^*) - \log\left(\sum_{\tilde{y} \in Y_X} e^{S(X, \tilde{y})}\right), \quad (15)$$

where Y_X is all possible label sequences for the input sentence X .

During the decoding process, the Viterbi algorithm [24] is used to decode, and finally the sequence with the highest predicted total score is output as the final optimal sequence:

$$y^* = \underset{\tilde{y} \in Y_X}{\operatorname{argmax}}(S, \tilde{y}). \quad (16)$$

Compared with other traditional deep learning named entity recognition methods, the introduction of the BERT pre-training language model is the main difference of this model. This model needs to be learned and researched on a large amount of corpus. It can calculate the vector representation of the word according to the context information, can characterize the ambiguity of the word, and enhance the semantic representation of the sentence.

Table 1 Data labeling specifications

Entity category	Marking specification
Name of historical Dynasty	DYN (Dynasty)
Names of historical figures	PER (Person)
Historical time	TIME (Time)
Historical location	LOC (Location)
Official title	CH (ChengHao)

Fig. 4 'BIO' tagging format

4 Experimental Settings

This section mainly introduces the experimental setup of our model training. It mainly includes data set introduction, annotation strategy, evaluation index, experiment environment and parameters setting.

4.1 Datasets and Annotation Strategies

Data sets are a key part of named entity recognition. It determines whether the model trained on the data set is suitable for practical problems [28]. However, for the current Chinese named entity recognition data set, there is no proprietary data set that belongs to the specific field of historical culture. Therefore, our goal is to create a data set suitable for the identification of named entities in the historical and cultural field. This paper will use the manual labeling method to label the data set. The specific data labeling specifications are shown in Table 1.

HistoryNER is a custom historical and cultural field named entity recognition data set, and the data source is mainly from domestic public history and cultural website information and related encyclopedia website information. For example, History Chunqiu Website (<http://www.lishi-chunqiu.com/>), Sohu History (<https://history.sohu.com/>), etc. The data set used in this paper consists of more than 200,000 words of historical and cultural texts. The entities to be identified include historical dynasty name, historical character name, historical time, historical location and official title.

The labeling of the data in this paper uses the BIO three-segment notation. For each entity of the text, the

Table 2 Data labeling specifications

Entity category	Train_Number	Test_Number
DYN	2596	673
PER	3000	924
TIME	4771	596
LOC	5867	822
CH	2166	864

first word is marked as “B- (entity name)”, and the subsequent mark is “I- (entity name)”. Irrelevant words are all marked as “O”. The marked data is divided into 11 categories, namely “B-DYN”, “I-DYN”, “B-PER”, “I-PER”, “B-TIME”, “I-TIME”, “B-LOC”, “I-LOC”, “B-CH”, “I-CH”, “O”. An example of BIO formatting can be seen in Fig. 4, where PER representing person.

After the above-mentioned labeling method was used to label the data set of more than 200,000 characters obtained in this article by manual labeling, more than 20,000 entities were labeled in total. In the whole data set labeling process, all the labels are completed by one person, and the labeling time is about 35 days. The advantage of being done independently by an individual is that it can avoid different tagging people’s understanding of different words from producing different tags. The number distribution of each entity in the training set and test set is shown in Table 2.

During the experiment, this paper divides the manually annotated data set into a training set and a test set. Among them, the ratio of training set to test set is 4:1.

4.2 Evaluation Metrics

Evaluation indicators for named entity recognition include *Precision*, *Recall*, and F_1 score (F_1). Here T_p denotes the number of true classes, F_p denotes the number of false positive classes, F_n denotes the number of false negative classes, and T_n denotes the number of true negative classes. The specific formula is as follows:

$$\text{Precision} = \frac{T_p}{T_p + F_p}, \quad (17)$$

$$\text{Recall} = \frac{T_p}{T_p + F_n}, \quad (18)$$

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (19)$$

Table 3 Experimental environment

Configuration item	Parameter
Operating system	Ubuntu 16.04
GPU	RTX2080Ti (11)
Python	3.7
Development tools	Pycharm IDE
Deep learning library	Keras, TensorFlow

4.3 Experimental Environment

To ensure the smooth progress of the entire experiment, this paper uses the following environment configuration for experiment, which is shown in Table 3.

4.4 Parameter Setting

The models used in this paper are all built using TensorFlow. The BERT pre-training language model set by default uses a 12-head attention mechanism Transformer with a hidden layer of 768 dimensions and a total of 110 M parameters. The maximum sequence length is 100, batch size is 32, LSTM hidden unit is 150. Adam is used as an optimizer, dropout is set to 0.5, and learning rate is set to $1e - 5$.

5 Experimental Results and Analysis

To better evaluate the performance of our method in the named entity recognition task in the field of ancient Chinese history and culture, some comparison experiments are done by comparing the performance of the proposed model with the baseline model. At the same time, our model is compared with the model in the previously published papers on historical and cultural named entity recognition.

5.1 Comparison with Baseline Models

The baseline model used in this paper is the BiLSTM-CRF model, which is the most widely used in named entity recognition applications. Based on the baseline model, a CNN layer is added to extract character-level features from historical and cultural text data. The character-level features extracted through the CNN layer are stitched with the trained word vector. The spliced vector is input into BiLSTM, and the output result is transferred to CRF after training is completed, and the optimal sequence label is selected in the CRF layer. Therefore, the BiLSTM-CNN-CRF model is taken as the baseline model. BERT-BiLSTM-CRF model finally used in this paper is compared with two major baseline models. As can be seen from the histogram in Fig. 5, the precision, recall rate and F_1 value of BiLSTM-CRF model using

Fig. 5 Performance comparison of our method with the baseline model

Table 4 Performance comparison with different LSTM units dimensions

Dimensions	Precision	Recall	F1
50	0.8760	0.8774	0.8764
100	0.8700	0.8840	0.8765
150	0.8933	0.9210	0.9067
200	0.8736	0.8696	0.8712

CNN to extract character level features are 4.06%, 5.8% and 4.97%, which is obviously higher than that of BiLSTM-CRF model.

However, the performance of the model in this paper is 4.6%, 15.64% and 12.17% higher than that of the BiLSTM-CRF using CNN to extract character features, and 8.66%, 9.84% and 7.2% higher than that of the BiLSTM-CRF. In summary, our model has made a significant improvement in performance.

5.2 Comparison of LSTM Units Dimensions

The dimensions of hidden units in LSTM can change the training parameters of the model, thereby affecting the overall performance and computational complexity. To get the best hyper-parameters, hidden units of different dimensions are set up, including 50, 100, 150 and 200. It can be seen from Table 4 that when the dimension of the hidden unit is 150, the *F1* score of the model is the highest at 90.67%. When the hidden unit dimension is less than 150, the performance of the model as a whole is reduced by 3% points in *F1* value performance compared to the model with a hidden unit dimension of 150. When the dimension of the hidden unit is higher than 150, the overall performance of the model is worse than that of the hidden unit with a dimension of

Table 5 Performance comparison for whether to use Dropout

Dropout	Precision	Recall	<i>F1</i>
No using	0.8729	0.8802	0.8755
Using	0.8933	0.9210	0.9067

Table 6 Performance comparison between different epochs

Epoch	Precision	Recall	<i>F1</i>
50	0.8819	0.8953	0.8878
80	0.8893	0.8949	0.8919
100	0.8855	0.9157	0.9001
150	0.8934	0.9210	0.9067
200	0.8852	0.9222	0.9030
300	0.8809	0.8957	0.8880

150, and the *F1* score is 3.5% points lower than that of the hidden unit with a dimension of 150. Based on analysis, it is learned that too few hidden units will result in insufficient feature capturing ability, so that the model performs poorly in overall performance. On the contrary, as the number of dimensions increases, the increase in training parameters leads to an increase in computational complexity with poor performance. Therefore, the hidden unit dimension of LSTM is set as 150.

5.3 Comparison of Using Dropout

To verify the effectiveness of using Dropout in the experiment, a set of comparative experiments is added here. The experimental results are shown in Table 5. It shows the performance comparison of whether to use Dropout.

The final experimental results show that the overall performance of the experiment using Dropout has been improved, compared with 3.12% without Dropout.

5.4 Performance Comparison Between Different Epochs

To obtain better parameters of the model, different epochs are set, which are respectively 50, 80, 100, 150, 200 and 300. Table 6 gives a comparison of model performance under different epochs. It can be found from the table that when training 150 epochs, the model has obtained the optimal values in the precision rate and *F1* value evaluation indicators, which are 89.33% and 90.67% respectively, and the recall rate reaches the most highest when the epoch is equal to 200. Therefore, epoch is equal to 150 is chosen as the final training round of the model.

Fig. 6 Performance comparison of our method with the baseline model

5.5 Comparison Between Different Methods

At the same time, IDCNN-CRF and BiLSTM-Attention-CRF are used to test the performance of the self-built historical entity domain named entity recognition dataset. IDCNN in IDCNN-CRF adds dilation width on the basis of CNN. When it acts on the input matrix, it skips the input data in the middle of all dilation widths. The size of the filter matrix itself remains unchanged, so that the filter obtains the data on the broader input matrix, so as to obtain a better sequence labeling effect. BiLSTM-Attention-CRF is based on BiLSTM-CRF to increase the attention mechanism. The attention mechanism is adopted to assign different weight coefficients to the vectors of different features in the text to better extract features, thereby improving the name entity recognition performance.

The experimental results are shown in Fig. 6. Through experimental comparison, it is found that the performance of the model used in this paper is improved by 5.63% compared with BiLSTM-Attention-CRF in *F1* value performance. It is 1.48% higher than IDCNN-CRF in *F1* value performance.

From the histogram in Fig. 6, it is found that the difference between IDCNN-CRF in the precision rate and recall rate of the model used in this paper is not too large, about 1% point. Compared with BiLSTM-CRF, the precision and recall rate differ by about 5% points. In general, the model used in this paper has achieved the best results in performance evaluation index. This also means that the model used in this paper is effective in the task of named entity recognition in the field of ancient Chinese history and culture.

6 Conclusions and Future Work

Aiming at the problem of entity recognition in the field of historical culture, a named entity recognition method based on BERT-BiLSTM-CRF network is proposed. First, the model uses the BERT pre-trained language model to encode a single character to obtain a vector representation corresponding to each character. Second, it uses the BiLSTM layer to semantically encode the input text, and finally uses the CRF layer to output the label with the highest probability to obtain each character category. Experimental results show that the named entity recognition framework proposed in this paper can effectively mine the contextual semantic information of historical and cultural texts. Compared with traditional named entity recognition algorithms, the precision rate, recall rate and *F1* value have been greatly improved. It also has a high recognition rate for entities in historical text in practical applications. However, the method proposed in this paper is based on small-scale data sets for testing. Therefore, the model has certain limitations. In the future work plan, the data scale will be expanded. In addition, an attention mechanism will be also introduced to improve the performance of the model in the future.

Acknowledgements Thank all reviewers and editors for their helpful suggestions for the improvement of this paper. Thanks to Economic and social development research project of Liaoning province in 2021 (No. 2021lslybkt-022) for supporting this article. Thanks to Jiangsu Provincial Social Science Applied Research Excellence Project (No. 21SYB-138); Research Innovation Team Grant Project (No. 2021KYTD04) for funding this article.

Declarations

Conflict of interest The authors declared that they have no conflicts of interest to this work.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Zhang, M., Geng, G., Chen, J.: Semi-Supervised Bidirectional Long Short-Term Memory and Conditional Random Fields

- Model for named-entity recognition using embeddings from language models representations. *Entropy* **22**, 252 (2020)
2. Wang, Z.N., Jiang, M., Gao, J.L., CHEN, Y.X.: Chinese named entity recognition method based on BERT[J]. *Comput. Sci.* **46**(11A), 138–142 (2019)
 3. Li, L.F., Yang, J.Q., Li, B.S., Du, Y.X., Hu, W.J.: Named entity recognition of Chinese EMR Based on Bert [J]. *J. Inner Mongolia Univ. Sci. Technol.* **39**(01), 71–77 (2020)
 4. Marcińczuk, M.: Automatic construction of complex features in conditional random fields for named entities recognition. In: *Proceedings of the International Conference Recent Advances in Natural Language Processing*, Hissar, Bulgaria, 7–9 (2015)
 5. Chen, S.D., Ouyang, X.Y.: Overview of named entity recognition technology [J/OL]. *Radio Commun. Technol.*, 1–11 [2020-05-10] [2020-05-19]. <http://kns.cnki.net/kcms/detail/13.1099.TN.20200414.1436.002.html>
 6. Saimaiti, A., Wang, L., Yibulayin, T.: Learning subword embedding to improve uyghur named-entity recognition. *Information* **10**, 139 (2019)
 7. Hammerton, J.: Named entity recognition with long short-term memory. *HLT-NAACL. Association for Computational Linguistics.* **4**, 172–175 (2003)
 8. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* **12**, 2493–2537 (2011)
 9. Santos, C.N., Guimaraes, V.: Boosting named entity recognition with neural character embeddings[J]. *arXiv preprint arXiv:1505.05008* (2015)
 10. Huang, Z.H., Xu, W., Yu, K.: Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991* (2015)
 11. Chiu, J.P.C., NICHOLS, E.: Named entity recognition with bidirectional LSTM-CNNs[J]. *Trans. Assoc. Comput. Linguist.* **4**, 357–370 (2016)
 12. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. *arXiv arXiv:1603.01360* (2016)
 13. Ma, X., HOVY, E.: End-to-end Sequence labeling via Bidirectional LSTM-CNNs-CRF [C]. In: *Proceedings of the 54th annual meeting of the association for computational linguistics (Volume 1: Long Papers)*, p. 10644074. Association for Computational Linguistics, Germany (2016)
 14. Strubell, E., Verga, P., Belanger, D., McCallum, A.: Fast and Accurate Entity Recognition with Iterated Dilated Convolutions. 2670–2680. <https://doi.org/10.18653/v1/D17-1283.2017>
 15. Zhang, Y., Yang, J.: Chinese ner using lattice lstm[J]. *arXiv preprint arXiv:1805.02023* (2018)
 16. Liu, C., Huang, C., Wang, H., Bol, P. K.: Mining local gazetteers of literary Chinese with CRF and pattern based methods for biographical information in Chinese history. In: *2015 IEEE International Conference on Big Data (Big Data)*, Santa Clara, CA, pp. 1629–1638 (2015) <https://doi.org/10.1109/BigData.2015.7363931>
 17. Sie, S., Ke, H., Chang, S.: “Development of a text retrieval and mining system for Taiwanese historical people” *Pacific Neighborhood Consortium Annual Conference and Joint Meetings (PNC)*. Tainan **2017**, 56–62 (2017). <https://doi.org/10.23919/PNC.2017.8203522>
 18. Wang, Y., Wang, M.X., Zhang, S., Du, L.: Recognition of named entity of police text based on BERT [J]. *Comput. Appl.* **40**(02), 535–540 (2020)
 19. Ajees, A.P., Mary Idicula, S.: An improved word representation for deep learning based NER in Indian languages. *Information* **10**, 186 (2019)
 20. Gorla, S., Neti, L.B.M., Malapati, A.: Enhancing the performance of Telugu named entity recognition using Gazetteer features. *Information* **11**, 82 (2020)
 21. Seti, X., Wumaier, A., Yibulayin, T., Paerhati, D., Wang, L., Saimaiti, A.: Named-entity recognition in sports field based on a character-level graph convolutional network. *Information* **11**, 30 (2020)
 22. Dai, Z., Wang, X., Ni, P., Li, Y., Li, G., Bai, X.: Named entity recognition using BERT BiLSTM CRF for chinese electronic health records. In: *12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. Suzhou, China **2019**, 1–5 (2019). <https://doi.org/10.1109/CISP-BMEI48845.2019.8965823>
 23. Mesnil, G., He, X., Deng, L., Bengio, Y.: Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. *Interspeech* **8**, 3771–3775 (2013)
 24. Viterbi, A.J., Wolf, J.K., Zehavi, E., et al.: A pragmatic approach to trellis-coded modulation[J]. *IEEE Commun. Mag.* **27**(7), 0–19 (1989)
 25. Mikolov, T., Chen, K., Corrado, G., et al.: Efficient estimation of word representations in vector space[J]. *arXiv preprint arXiv:1301.3781* (2013)
 26. Devlin, J., Chang, M.W., Lee, K. et al.: Bert: Pre-training of deep bidirectional transformers for language understanding[J]. *arXiv preprint arXiv: 1810.04805* (2018)
 27. Yang, P., Dong, W.Y.: Chinese named entity recognition method based on BERT embedding [J]. *Comput. Eng.* **46**(04), 40-45+52 (2020)
 28. Han, X., Zhang, Y., Zhang, W., Huang, T.: An attention-based model using character composition of entities in Chinese relation extraction. *Information* **11**, 79 (2020)
 29. Christopher, O.: Understanding LSTM networks[J/ OL]. colah's blog. [2015-08-27]. [2020-05-18]. <https://colah.github.io/posts/2015-08-Understanding-LSTMs>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.