End-to-End Chinese Entity Recognition Based on BERT-BiLSTM-ATT-CRF

Abstract: Traditional named entity recognition methods need professional domain knowledge and a large amount of human participation to extract features, as well as the Chinese named entity recognition method based on a neural network model, which brings the problem that vector representation is too singular in the process of character vector representation. To solve the above problem, we propose a Chinese named entity recognition method based on the BERT-BiLSTM-ATT-CRF model. Firstly, we use the bidirectional encoder representations from transformers (BERT) pre-training language model to obtain the semantic vector of the word according to the context information of the word; Secondly, the word vectors trained by BERT are input into the bidirectional long-term and short-term memory network embedded with attention mechanism (BiLSTM-ATT) to capture the most important semantic information in the sentence; Finally, the conditional random field (CRF) is used to learn the dependence between adjacent tags to obtain the global optimal sentence level tag sequence. The experimental results show that the proposed model achieves state-of-the-art performance on both Microsoft Research Asia (MSRA) corpus and people's daily corpus, with F1 values of 94.77% and 95.97% respectively.

Keywords: named entity recognition (NER); feature extraction; BERT model; BiLSTM; attention mechanism; CRF

LI Daiyi[1], TU Yaofeng[2], ZHOU Xiangsheng[2], ZHANG Yangming[2], MA Zongmin[1]

(1. Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China;
2. ZTE Corporation, Shenzhen 518057, China)

DOI: 10.12142/ZTECOM.2022S1005

http://kns.cnki.net/kcms/detail/34.1294.tn.20220119.1630.002.html, published online January 20, 2022

Manuscript received: 2021-01-13

Citation (IEEE Format): D. Y. Li, Y. F. Tu, X. S. Zhou, et al., "End-to-end chinese entity recognition based on BERT-BiLSTM-ATT-CRF," *ZTE Communications*, vol. 20, no. S1, pp. 27 – 35, Jan. 2022. doi: 10.12142/ZTECOM.2022S1005.

1 Introduction

Named entity recognition (NER) is one of the key technologies in natural language text data processing. Its main function is to identify specific types of entities from unstructured text data, such as person names, place names, organization names and domain-specific words. At present, NER is widely used in information extraction, knowledge graph construction, machine translation and intelligent question answering. The best performance of traditional methods is based on statistical models, such as hidden Markov models (HMM), support vector machines (SVM) and conditional random fields (CRF). However, these methods need professional domain knowledge and a large number of human participations to extract features, which increases the difficulty of named entity recognition in a specific domain. In recent years, the state-of-the-art English NER models are mainly constructed by combining deep learning and CRF. For example, the method of combining long short-term memory (LSTM) with CRF performs well in NER tasks[1 – 5].

Compared with English NER, the difficulties of Chinese NER mainly include the following aspects: 1) Chinese words have stronger polysemy, and the same words may have different meanings in different contexts; 2) English text contains space, initial letter upper and other identifiers to determine the entity boundary, while Chinese text does not have similar entity boundary identifiers, which increases the difficulty of entity boundary identification; 3) Chinese NER tasks usually need to be combined with Chinese word segmentation and shallow parsing, and the accuracy of these methods directly affects the effectiveness and stability of the entity recognition model. In view of the above problems, many researchers have applied the deep learning method to the research of Chinese NER, because the feature extraction of text data through deep learning not only avoids the tedious manual feature extraction, but also increases the generalization ability of the model. HAMMERTON et al.[6]

This work was supported by ZTE Industry-University-Institute Cooperation Funds under Grant No. HC-CN-20190910009, and in part by the National Natural Science Foundation of China under Grant No. 61772269.

constructed the basic framework of LSTM-CRF entity recognition model. On this basis, CHIU et al.[7] added a convolutional neural network (CNN) data preprocessing layer to the front end of the LSTM model, and obtained the F1 value of 88.83% on comll-2003 corpus; LI et al.[8] constructed the CNN-bidirectional long-term and short-term memory network (BiLSTM)-CRF named entity recognition model, and achieved significant results on the Biocreative II GM and JNLPBA2004 corpora; LUO et al.[9] embedded the attention mechanism (ATT) on the basis of the BiLSTM-CRF model, and obtained the F1 value of 91.14% on the Biocreative IV corpus; WU et al.[10] jointly trained the word segmentation and the CNN-BiLSTM-CRF model to enhance the recognition ability of the model for entity boundary, thereby improving the performance of the model's entity recognition; QIN et al.[11] constructed a CNN-BiLSTM-CRF named entity recognition model combined with feature templates, and used artificial feature templates to extract local features of text, which achieved good results on large-scale network security data; ZHANG et al.[12] proposed an LSTM model based on a lattice structure, which makes full use of word and word sequence information to improve the performance of the entity recognition model; WANG et al.[13] used segmental neural network structure to extract text features and obtained the F1 value of 92.05% on the Microsoft Research Asia (MSRA) corpus; LIU et al.[14] embedded the attention mechanism on the basis of the dense connection (DC)-BiLSTM-CRF model, and obtained the F1 value of 92.05% on the MSRA corpus; LIU et al.[15] constructed a word-character (WC)-BiLSTM-CRF model, which added word information to the beginning or end of the whole character to enhance semantic information, and obtained the F1 value of 93.74% on the MSRA corpus.

However, there are differences between Chinese characters and words. The above methods focus on the feature extraction of characters and words, but ignore the polysemy problem in Chinese. In order to solve this problem, DEVLIN et al.[16] constructed encoder representations from transformers (BERT) pre-training language model to obtain the semantic vector of words, which enhanced the generalization ability of the word vector model, enriched the syntax and grammatical information in the sentence, and effectively solved the problem of polysemy representation of a word. For example, in the sentence "道可道，非常道 (The Dao/way that can be told is not the usual Dao/way)," the two "Dao" characters have different meanings, but in Word2vec[17] and Glove[18], the vector representations of the two "Dao" characters are the same, which is inconsistent with the objective facts. The BERT model can obtain the semantic vector of words according to context information, represent the polysemy of words, and enhance the semantic representation of sentences. In order to automatically extract the depth features of Chinese text and solve the problem of characterizing the polysemy representation, this paper constructs a Chinese NER model based on the BERT-BiLSTM-ATT-CRF network structure. The model uses the BERT model to train the word vector based on the context information of a word, and then inputs the trained word vector sequence into the BiLSTM-ATT model for further training, to capture the most important semantic information in the sentence, and finally the entity recognition result is marked by the CRF layer. The experimental results show that the proposed model achieves state-of-the-art performance on both the MSRA corpus and people's daily corpus, with the F1 values of 94.77% and 95.97% respectively.

The innovations of this paper are mainly as follows:

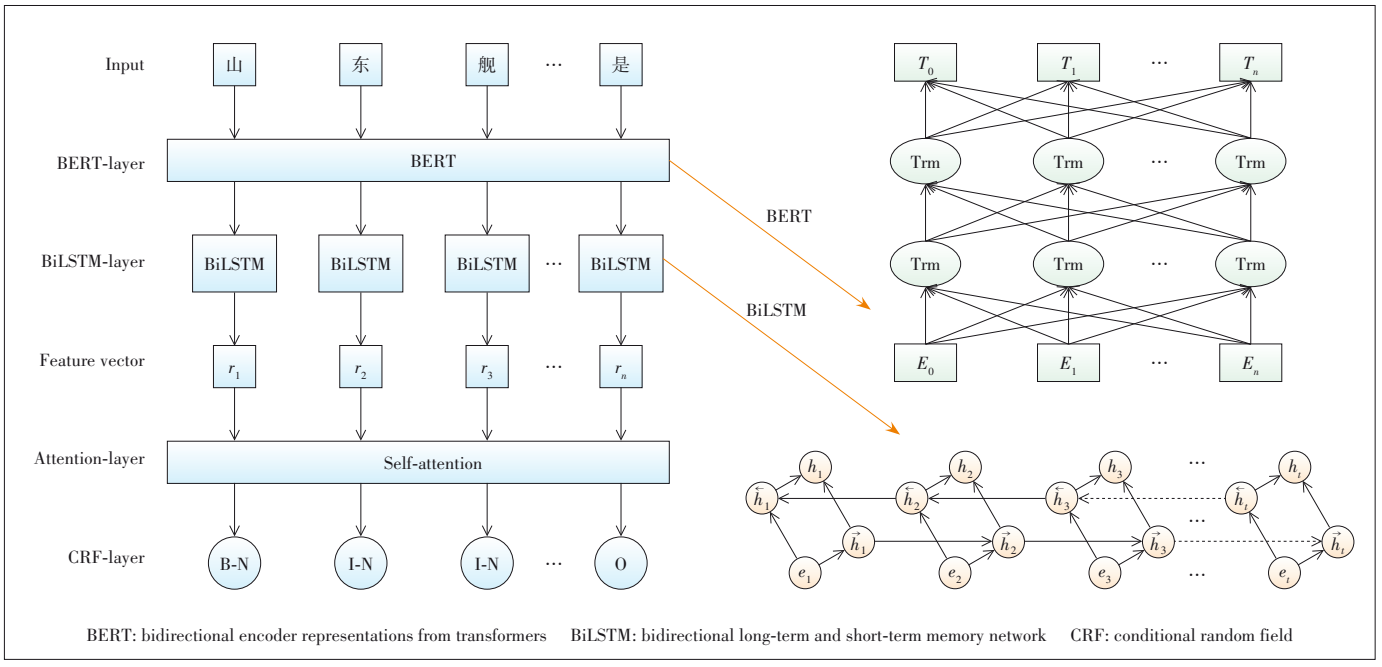
1) This paper applies the BERT pre-training language model to Chinese NER, which can obtain the semantic information of Chinese words in different contexts according to the context information of words, which effectively solves the problem of oversimplification of word vector representation. The BERT model is obtained by replacing the feature extractor in the ELMO model with a transformer. The experimental results show that the performance of the entity recognition model is improved effectively by introducing the BERT model.

2) The attention mechanism is embedded into the BiLSTM model and construct a BiLSTM-ATT module, which can selectively give different weights to different words in the text, and then the context-based semantic association information is used to effectively make up for the lack of deep neural network in obtaining local features, so as to highlight the importance of specific words to the whole text.

3) In the BERT-BiLSTM-ATT-CRF model proposed in this paper, the BERT model is only used to obtain the vector representation of the words in the text. The parameters of the model remain unchanged in the whole training process. The word vectors trained by the BERT are classified and recognized through the BiLSTM-ATT-CRF model, which can maintain the polysemy of words and reduce the training practice parameters.

2 Proposed BERT-BiLSTM-ATT-CRF Model

In recent years, converting traditional named entity recognition problems into sequence labeling tasks is the basic idea of the deep learning model for Chinese NER. The overall structure of the proposed BERT-BiLSTM-ATT-CRF model is shown in Fig. 1. The whole model is divided into three layers: the BERT layer, BiLSTM-ATT layer and CRF layer. Firstly, the annotated corpus is represented by the word vector based on context information through the BERT layer, and then the word vector is input into the BiLSTM-ATT layer for further training to obtain the important semantic features in the sentence. Finally, the output result of the BiLSTM-ATT layer is decoded by CRF to obtain the tag sequence of the optimal sentence level, and then extracting and classifying each entity in the sequence is conducted and classified to complete the task of Chinese entity recognition.



▲**Figure 1. Overall architecture of BERT-BiLSTM-ATT-CRF model**

Algorithm 1 is the algorithm flow of the BERT-BiLSTM-ATT-CRF model.

Algorithm 1. The algorithm flow of BERT-BiLSTM-ATT-CRF model

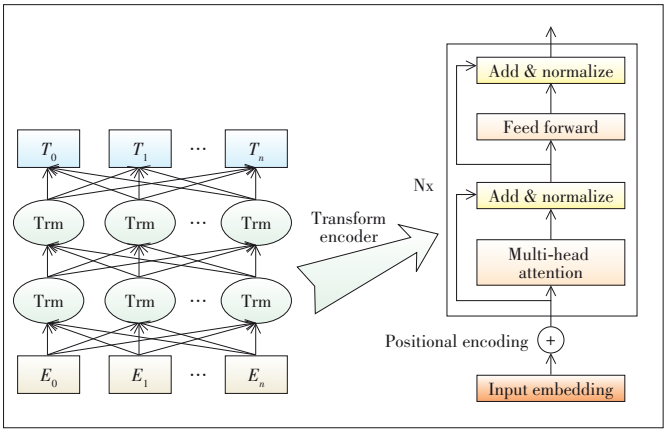
Input: A sentence sequence S, a radical information matrix $\boldsymbol{A}$.
Output: The entity list Y.
1: Preprocessing the dataset. The output embedding of each word in the sequence consists of three parts: token embedding (E^t), segment embedding (E^s) and position embedding (E^p);
2: The generated sequence vector $\boldsymbol{X} = E^t \oplus E^s \oplus E^p$ is input into bidirectional transformer encoder for feature extraction, and the sequence vector with rich semantic preferential energy is obtained;
3: The word vectors generated by BERT training are input into the BiLSTM-ATT module to obtain the sequence depth features. $\boldsymbol{H} = BiLSTM(\boldsymbol{X})$, $\boldsymbol{H}' = Attention(\boldsymbol{H})$;
4: The probability and loss score of tag sequence y were calculated by CRF model;
5: **if** not converge **then**
Repeat lines 2 – 4;
6: **end if**
7: **return** the label sequence Y by using the Viterbi algorithm.

2.1 BERT Module

In the field of natural language processing (NLP), word embedding is used to map a word into a low dimensional space, which can effectively solve the problem of text feature sparseness, so that similar words in the semantic space have a closer distance. Traditional word vector generation methods, such as one hot, word2vec and Elmo[19], and other pre-trained language models are mostly independent of the context information of words, so it is difficult to accurately represent the polysemy of words. However, the BERT model proposed by JACOB et al. can be used to represent words according to their context information in an unsupervised way, which can effectively solve the problem of polysemy representation.

The structure of the BERT model is shown in Fig. 2. The multi-layer bidirectional transformer[20] is used as the encoder in BERT model, and each unit is composed of feed-forward neural network (Feed Forward) and multi-head attention mechanism, so that the representation of each word can integrate the information of its left and right sides.

As shown in Fig. 2, the key part of the BERT model is the self-attention mechanism module in the transformer encoder. The function of the attention mechanism is to calculate every word in an input sentence to obtain the degree of correlation



▲**Figure 2. Structure of bidirectional encoder representations from transformers (BERT) model**

between words in the sentence and then adjust the weight coefficient matrix to obtain the representation of words. The calculation is as follows:

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{1}$$

where $(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})$ is the input word vector matrix, d_k is the dimension of the input vector, and QK^T is the relationship between input word vectors.

The calculation of the transformer adopts the multi-head attention mechanism[20] to project though multiple linear transformation pairs for enhancing the model ability of focusing on different positions. The calculation is shown in Eqs. (2) and (3):

$$\text{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Concat}(\text{head}_1,\text{head}_2,...,\text{head}_k)W^0, \tag{2}$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V), \tag{3}$$

where $\boldsymbol{W}^0$ is the additional weight matrix of the model to obtain different spatial position information. At the same time, in order to deal with the degradation problem in deep learning, the residual network and normalization layer are added into the transformer coding unit. The calculation is as follows:

$$LN(x_i) = \alpha \times \frac{x_i - u_i}{\sqrt{\sigma_L^2 + \varepsilon}} + \beta, \tag{4}$$

$$FFN(Z) = \max(0, ZW_1 + b_1)w_2 + b_2, \tag{5}$$

where α and β are learning parameters, and μ and σ are the mean and variance of the input layer. The representation of fully connected feedforward network (FFN) is shown in Eq. (5), where the output of the multi-head attention mechanism is denoted as Z and $\boldsymbol{b}$ is the bias vector.

2.2 BiLSTM-ATT Module

LSTM[21-22] is a variant of recurrent neural network (RNN). It can effectively solve the gradient explosion or gradient disappearance during RNN training. Since the LSTM model cannot process context information at the same time, GRAVES et al.[23] proposed the BiLSTM model, whose basic idea is to obtain the context information of input sequence through two hidden layers of LSTM. The specific operation is to connect the output vectors of the two hidden layers of LSTM to generate the context vector. The structure of LSTM unit is shown in Fig. 3, which consists of input gate, forgetting gate and output gate.

The vector representation of the output of the hidden layer of LSTM model is defined as follows:

$$f_t = \sigma(\boldsymbol{W}_{fx}\boldsymbol{x}_t + \boldsymbol{W}_{fh}h_{t-1} + \boldsymbol{b}_f), \tag{6}$$

$$i_t = \sigma(\boldsymbol{W}_{ix}x_t + \boldsymbol{W}_{ih}h_{t-1} + \boldsymbol{b}_i), \tag{7}$$

$$\hat{c}_t = \tanh(w_{cx}\boldsymbol{x}_t + w_{ch}h_{t-1} + \boldsymbol{b}_c), \tag{8}$$

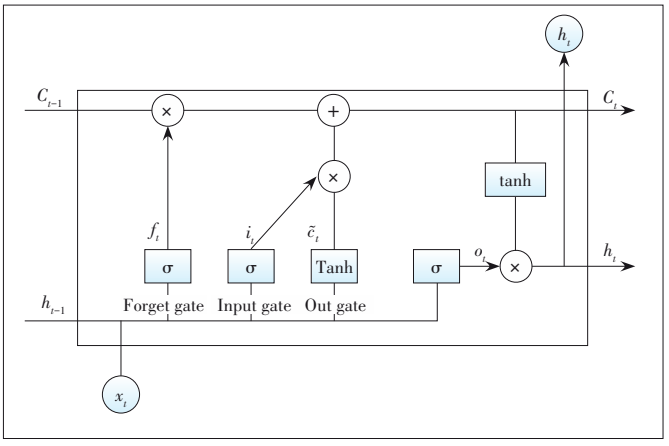
$$c_t = f_t * c_{t-1} + i_t * \hat{c}_t, \tag{9}$$

$$o_t = \sigma(\boldsymbol{W}_{ox}x_t + \boldsymbol{W}_{oh}h_{t-1} + \boldsymbol{b}_o), \tag{10}$$

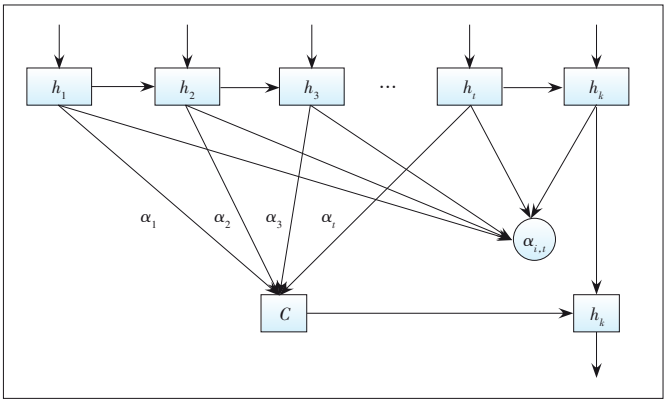
$$h_t = o_t * \tanh(c_t), \tag{11}$$

where $\boldsymbol{W}$ and $\boldsymbol{b}$ respectively represent the weight matrix and bias auto vector connecting the two hidden layers; σ is sigmoid activation function; $\boldsymbol{x}_t$ represents the input vector at the time t; f_t, i_t and o_t represent the input gate, forgetting gate and output gate at the time t respectively; * represents the point multiplication operation, and h_t represents the output of LSTM unit at the time t.

The core idea of attention mechanism is to focus on important information at a specific time, while ignoring other non-important information[24-25]. The integration of attention mechanism and BiLSTM model (Fig. 4) can effectively highlight the



▲ **Figure 3. Long short-term memory (LSTM) unit structure**



▲ **Figure 4. Embedding attention mechanism into bidirectional long-term and short-term memory network (BiLSTM) model**

role of keywords. The purpose of embedding attention mechanism in the BiLSTM neural network is to selectively give different weights to different words in the text, and then using context based semantic association information can effectively make up for the deficiency of deep neural network in obtaining local features.

In this paper, the calculation of the attention mechanism constructed can be summarized as follows:

1) Suppose $\boldsymbol{h}_i$ represents the feature vector output from the hidden layer of the BiLSTM model containing the context information of word w_i; then h_i is transformed into u_i through the full connection layer, where u_i is defined as follows:

$$u_i = \tanh(\boldsymbol{W}h_i + \boldsymbol{b}), \tag{12}$$

where $\boldsymbol{W}$ and $\boldsymbol{b}$ represent the weight matrix and bias auto vector of attention mechanism respectively.

2) The similarity between u_i and the context vector $\boldsymbol{u}_t$ is calculated, and the normalized weight $\alpha_{i,t}$ is obtained by the Softmax function, where $\alpha_{i,t}$ is defined as follows:

$$\alpha_{i,t} = \frac{\exp(u_i^T \boldsymbol{u}_t)}{\sum_i \exp(u_i^T \boldsymbol{u}_t)}, \tag{13}$$

where $\alpha_{i,t}$ represents the importance of the corresponding word in the whole sentence; u_t represents the contribution of the corresponding word to the sentence, which is mainly obtained through random initialization and training.

3) The h_i obtained by each word is multiplied by the corresponding attention weight $\alpha_{i,t}$ to obtain the global vector $\boldsymbol{C}$ of the sentence, where $\boldsymbol{C}$ is defined as follows:

$$\boldsymbol{C} = \sum_{j=1}^{T} a_{i,t}\boldsymbol{h}_j. \tag{14}$$

The sentence-level global vector $\boldsymbol{C}$ and the BiLSTM layer output h_t of the target word are combined into a vector $[\boldsymbol{C}; h_t]$, which is fed to a tanh function as the output of attention layer. The output z_t of the attention layer is defined as follows:

$$z_t = \tanh(\boldsymbol{W}[\boldsymbol{C}; \boldsymbol{h}_t]). \tag{15}$$

2.3 CRF Module

In the task of entity recognition, the BiLSTM model only obtains the word vector containing context information, but cannot deal with the interdependence between adjacent tags. Therefore, we use the CRF model[26-27] to obtain an optimal prediction sequence through the relationship of adjacent tags, which is used to make up for the shortcomings of the BiLSTM model. The main operations of the CRF layer are as follows:

1) The parameter of the CRF layer is a $(k+2)\times(k+2)$ matrix $\boldsymbol{A}$. The A_{ij} represents the transfer score from the i-th tag to the j-th tag, and then the previously labeled tags can be used when labeling a position. The reason for adding 2 is to add a start state for the beginning of a sentence and a termination state for the end of the sentence. If you remember a tag sequence $y = (y_1, y_2, ..., y_n)$ whose length is equal to the length of the sentence, the model scores the tag of sentence x, which is equal to y. The specific calculation is shown in Eq. (16).

$$\text{Score}(x,y) = \sum_{i=1}^{n} \boldsymbol{P}_{i,y_i} + \sum_{i=1}^{n+1} A_{y_{i-1},y_i}. \tag{16}$$

Among them, $\boldsymbol{P}$ is the score matrix output by the BiLSTM-ATT module, and the size of P is $n \times k$, where n is the number of words and k is the number of tags; P_{ij} is the score of the i-th tag corresponding to the j-th word; $\boldsymbol{A}_{ij}$ is the transfer score matrix, A is the score of tag i transferred to tag j, and the size of A is $k+2$.

2) Obviously, the score of the whole sequence is equal to the sum of the scores of each position. The score of each position is divided into two parts: one part is determined by the score matrix $\boldsymbol{P}$ output by the LSTM model and the other is determined by the transfer matrix $\boldsymbol{A}$ of the CRF model. Therefore, the normalized probability can be obtained by the Softmax function as

$$\boldsymbol{P}(y|x) = \frac{\exp(\text{Score}(x,y))}{\sum_{y' \in Y_x} \exp(\text{Score}(x,y'))}, \tag{17}$$

where x is the true label value, y' is the predicted label value, and Y_x is the set of all possible labels. During the training process, the maximum likelihood probability of the correct label sequence is as follows:

$$\log(p(y|x)) = S(x,y) - \sum_{y' \in Y_x}^{n} S(x,y'). \tag{18}$$

3) Finally, the Viterbi algorithm[28] is used to obtain the sequence with the highest total score of prediction on all sequences, which is taken as the annotation result of the final entity recognition. The sequence with the highest score is as follows:

$$Y^* = \arg\max(x,y'), (y' \in Y_x). \tag{19}$$

3 Experiments

Our experiments on different datasets show that the proposed BERT-BiLSTM-ATT-CRF entity recognition model is effective in different fields. In addition, we compare the existing NER models with state-of-the-art performance, and further verify that our entity recognition model is effective and stable.

3.1 Datasets

In this paper, we mainly use the Chinese annotated the People's Daily corpus[29] and MSRA corpus[30] as the experimental data sets. These two data sets are Chinese evaluation data sets

in the domestic public news field. They mainly include three types of entities: person names, place names and organizations. In order to ensure the fairness of the comparison, we use the same data segmentation method as CHEN et al. used[31], and divide the data into three parts: the training set, verification set and test set. The specific scales of the corpora are shown in Table 1.

3.2 Data Annotation and Evaluation Metrics

The commonly used labeling modes of NER include BIO (B-begin, I-inside, O-outside), BIOE (B-begin, I-inside, O-outside, E-end), BIOES (B-begin, I-inside, O-outside, E-end, S-single), etc. In this experiment, we choose to use the BIO labeling mode, and there are seven prediction labels, which are "O", "B-PER", "I-PER", "B-ORG", "I-ORG", "B-LOC" and "I-LOC". In order to evaluate the performance of the proposed model, the precision (P), recall (R) and F-measure (F) are used as the evaluation criteria. The definitions of P, R and F are shown as follows:

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN},$$

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \quad (\beta^2 \in [0, +\infty]). \tag{20}$$

Among them, TP is the number of positive samples correctly predicted; TN is the number of negative samples predicted correctly; FP is the number of negative samples predicted incorrectly; FN is the number of positive samples predicted incorrectly. P is the precision rate and R is the recall rate.

3.3 Hyper-Parameter Settings

We select the optimal hyper-parameter values of the model through model training and consideration of previous work in the literature. There are two kinds of pre-training language models: BERT-Base and BERT-Large. Some parameters of these two models are different. In this experiment, we choose to use the pre-training language model of BERT-Base. The model has a total of 12 layers and 768 dimensions of the hidden layer; it adopts a 12-head mode, including 110 million parameters. During the training, the maximum sequence length is set to 128, the size of batch size is 64, the number of hidden layers of BiLSTM is 200, and the Adam optimizer[32] is used to select the appropriate learning rate to 0.001 5. In order to prevent over-fitting of the model, the dropout technology[33] is introduced into the model and its value is set to 0.5. The specific parameter settings are shown in Table 2.

3.4 Experimental Results and Analysis

3.4.1 Compared with Traditional Neural Network

In order to make a more objective evaluation on the performance of the proposed BERT-BiLSTM-ATT-CRF model, we use the People's Daily corpus and MSRA corpus to evaluate the performance of different models, and use the values of P, R and $F1$ to evaluate the performance of model entity recognition. The specific experimental results are shown in Tables 3 and 4.

As shown in Tables 3 and 4, we compare the proposed BERT-BiLSTM-ATT-CRF model with the traditional classical neural network model. Firstly, the experimental results of LSTM-CRF and BiLSTM-CRF show that the $F1$ value of the latter is higher than that of the former on the People's Daily corpora and MSRA corpora. The main reason is that LSTM only considers the above information, while BiLSTM can obtain context sequence information by using bidirectional structure and extract more effective features. Secondly, the experimental results of BiLSTM and BiLSTM-CRF show that the $F1$ value of BiLSTM-CRF model increases by 5.04% and 7.99% respectively on the two corpora after adding the CRF module. The main reason is that the CRF module can make full use of the interdependence between adjacent tags while considering

▼Table 1. Statistics of datasets

Dataset	Type	Train	Dev	Test
People's Daily	Sentence	17.6k	0.9k	1.7k
MSRA	Sentence	46.4k	Null	4.4k

MSRA: Microsoft Research Asia corpus

▼ Table 2. Optimal hyper-parameter values of BERT-BiLSTM-ATT-CRF model

Layer	Parameter	Value
BERT	Transformer layer number	12
	Hidden layer dimension	768
	Head number	12
BiLSTM	Optimizer	Adam
	Batch size	32
	Dropout rate	0.5
	Learning rate	0.001 5
	Hidden layer number	200

ATT: attention mechanism
BERT: bidirectional encoder representations from transformers
BiLSTM: bidirectional long-term and short-term memory network
CRF: conditional random field

▼Table 3. Test results on People's Daily corpus

Model	P/%	R/%	$F1$/%
LSTM-CRF	84.20	80.20	82.00
BiLSTM	81.08	79.21	80.05
BiLSTM-CRF	87.21	83.21	85.09
BERT-BiLSTM-CRF	96.04	95.30	95.67
BERT-BiLSTM-ATT-CRF	96.28	95.67	95.97

ATT: attention mechanism
BERT: bidirectional encoder representations from transformers
BiLSTM: bidirectional long-term and short-term memory network
CRF: conditional random field
LSTM: long short-term memory

▼Table 4. Test results on MSRA corpus

Model	*P*/%	*R*/%	*F*1/%
LSTM-CRF	83.45	80.20	82.00
BiLSTM	78.72	79.21	80.05
BiLSTM-CRF	86.79	83.21	85.09
BERT-BiLSTM-CRF	94.38	94.92	94.65
BERT-BiLSTM-ATT-CRF	94.52	95.02	94.77

ATT: attention mechanism
BERT: bidirectional encoder representations from transformers
BiLSTM: bidirectional long-term and short-term memory network
CRF: conditional random field
LSTM: long short-term memory

the context information, so as to obtain the global optimal tag sequence.

At the same time, the results in Tables 3 and 4 show that the performance of the entity recognition model is improved when the attention mechanism is added to the BERT-BiLSTM-CRF model. The main reason is that the attention mechanism is embedded in the BiLSTM neural network, so that the model can selectively give different weights to different words in the text, and then use the context based semantic association information to effectively make up for the lack of deep neural network in obtaining local features.

In order to improve the performance of the Chinese entity recognition model, some researchers have introduced the BERT model to preprocess the word vector on the basis of the BiLSTM-CRF model. The experimental results show that the *F*1 values of the BERT-BiLSTM-CRF model on the two corpora are 94.74% and 94.21% respectively, which is much higher than that of the BiLSTM-CRF model on the same corpus. The main reason is the addition of the BERT model, which can obtain the semantic vector of the word according to the context information of the word to represent the polysemy of the word, so that the generated word vectors can better represent the semantic information in different contexts, thus enhancing the generalization ability of the model and improving the performance of the model entity recognition. In this paper, the attention mechanism is introduced on the basis of the BERT-BiLSTM-CRF model, which effectively highlights the role of keywords in sentences, thereby improving the entity recognition ability of the model. The comparison of the experimental results of BERT-BiLSTM-CRF and BERT-BiLSTM-ATT-CRF shows that after adding attention, the *F*1 value of the model obtained by BERT-BiLSTM-ATT-CRF is higher than that of the former in both corpora, which proves the effectiveness of the model proposed in this paper.

3.4.2 Comparison with Previous Works

In order to further verify the effectiveness and stability of the proposed BERT-BiLSTM-ATT-CRF model, we compare it with the existing advanced models. The results are shown in Table 5.

As shown in Table 5, the MSRA corpus is used as the data

▼Table 5 Different models compared on MSRA corpus

Model	*P*/%	*R*/%	*F*1/%
CHEN et al. (2006)[31]	91.22	81.71	86.20
ZHANG et al. (2006)[32]	92.20	90.18	91.18
ZHOU et al. (2013)[33]	91.86	88.75	90.28
LU et al. (2016)[34]	NULL	NULL	87.94
Radical-BiLSTM-CRF (2016)[35]	91.28	90.62	90.95
IDCNN-CRF (2017)[36]	89.39	84.64	86.95
Lattice-LSTM-CRF (2018)[12]	93.57	92.79	93.18
CNN-BiLSTM-CRF(2019)[10]	91.63	90.56	91.09
WC-LSTM-pertain (2019)[15]	Null	Null	93.74
BERT-IDCNN-CRF (2020)[36]	94.86	93.97	94.41
BERT-BiLSTM-CRF (2020)[37]	94.38	94.92	94.65
HanLP (BERT)[38]	94.79	95.65	95.22
BERT-BiLSTM-ATT-CRF	94.52	95.02	94.77

BERT: bidirectional encoder representations from transformers
BiLSTM: bidirectional long-term and short-term memory network
CNN: convolutional neural network
CRF: conditional random field
HanLP: Han Language Processing
IDCNN: Iterated Dilated Convolutional Neural Network
LSTM: long short-term memory
WC: word-character

set to evaluate the performance of the entity recognition model. CHEN et al.[31–34] constructed a statistical model using manual features and character embedding features. The Radical-BiLSTM-CRF[35] model uses bidirectional LSTM to extract the feature vector of the root sequence and then joins it with the character vector to form the model input, which improves the performance of model entity recognition. The Lattice-LSTM-CRF model[12] improves the traditional LSTM unit to grid LSTM, and then makes full use of the information between words and the word order, effectively avoiding the error of word segmentation and obtaining better results of entity recognition. The CNN-BiLSTM-CRF model[10] extracts glyph embedding with morphological features from each Chinese character by CNN, and connects it with the word embedding of semantic feature information to form the input of the model, which obtains good results. The WC-LSTM-CRF[15] model uses word information to strengthen semantic information and reduce the influence of word segmentation errors. The *F*1 value reaches 93.74%.

The above-mentioned entity recognition models greatly improve the value of *F*1, but the improved models always focus on the extraction of character and word features and ignore the problem of polysemy in Chinese. LI et al.[36] and XIE et al.[37] used the BERT pre-training language model to represent the vector, which enhanced the generalization ability of the word vector model and enriched the syntactic and grammatical information in the sentence. This model effectively solved the representation problem of polysemy of a word. In order to further improve the performance of entity recognition model, we construct the BERT-BiLSTM-ATT-CRF model based on the

research in Ref. [33]. The model can effectively capture the most important semantic information in the sentence while ensuring the polysemy representation of a word. Although the model proposed in this paper is not different from the model of BERT-BiLSTM-CRF and BERT-IDCNN-CRF, the $F1$ value of the model on the MSRA corpus reaches 94.77%. The experimental results show that the proposed model achieves state-of-the-art performance on both the MSRA corpus and People's Daily corpus.

4 Conclusions

Traditional named entity recognition methods require professional domain knowledge and a large amount of human participation to extract features. Meanwhile, there are some problems in Chinese entity recognition tasks, such as polysemy and Chinese sentences without entity boundary identifiers. Firstly, we use the BERT pre-training language model to obtain the semantic features containing the contextual information of the word, which effectively solves the problem of polysemy representation of a word; Secondly, the classic neural network model BiLSTM is embedded with the attention mechanism, which can extract the most important semantics in the sentence features; Finally, we use the CRF model to obtain an optimal prediction sequence through the relationship of adjacent tags, which is used to make up for the shortcomings of the BiLSTM model. In order to verify the effectiveness of the proposed BERT-BiLSTM-Att-CRF model, the People's Daily corpus and MSRA corpus are used as the data sets for model performance evaluation. Compared with other models, the BERT-BiLSTM-ATT-CRF model shows the best results on both the corpora.

The biggest advantage of the BERT-BiLSTM-ATT-CRF model is that it can conduct pre-training according to the semantic information of the word context and obtain the word level features, syntactic structure features and semantic information features of context, which makes the model have better performance than the other models. At the same time, the attention mechanism is embedded into the BiLSTM model to enhance the extraction of key information features in sentences. Combined with CRF, it can take advantage of the interdependence between adjacent tags to further improve the ability of Chinese entity recognition. Our next work plan is to study the construction method of domain specific NER, and test the performance and generalization ability of the proposed model in multi-domain NER tasks.

References

[1] GRIDACH M. Character-level neural network for biomedical named entity recognition [J]. Journal of biomedical informatics, 2017, 70: 85 – 91. DOI: 10.1016/j.jbi.2017.05.002
[2] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition [C]//Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2016. DOI: 10.18653/v1/n16-1030
[3] MA X Z, HOVY E. End-to-end sequence labeling via Bi-directional LSTM-CNNs-CRF [C]//54th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2016: 1064 – 1074
[4] SHIN Y, LEE S G. Learning context using segment-level LSTM for neural sequence labeling [J]. IEEE/ACM transactions on audio, speech, and language processing, 2020, 28: 105 – 115. DOI: 10.1109/TASLP.2019.2948773
[5] DONG D Z, OUYANG S. Optimization Techniques of Network Communication in Distributed Deep Learning Systems [J]. ZTE technology journal, 2020, 26(5): 2-8. DOI: 10.12142/ZTETJ.202005002
[6] HAMMERTON J. Named entity recognition with long short-term memory [C]// Proceedings of the seventh conference on natural language learning at HLT-NAACL. Association for Computational Linguistics, 2003: 172 – 175. DOI: 10.3115/1119176.1119202
[7] CHIU J P C, NICHOLS E. Named entity recognition with bidirectional LSTM-CNNs [J]. Transactions of the association for computational linguistics, 2016, 4: 357 – 370. DOI: 10.1162/tacl_a_00104
[8] LI L S, GUO Y K. Biomedical named entity recognition based on CNN-BLSTM-CRF model [J]. Journal of Chinese information processing, 2018, 32(1): 116 – 122. DOI: 10.3969/j.issn.1003-0077.2018.01.015
[9] LUO L, YANG Z H, YANG P, et al. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition [J]. Bioinformatics, 2017, 34(8): 1381 – 1388. DOI: 10.1093/bioinformatics/btx761
[10] WU F Z, LIU J X, WU C H, et al. Neural Chinese named entity recognition via CNN-LSTM-CRF and joint training with word segmentation [C]//The World Wide Web Conference. ACM, 2019: 3342 – 3348. DOI: 10.1145/3308558.3313743
[11] QIN Y, SHEN G W, ZHAO W B, et al. Network security entity recognition method based on deep neural network [J]. Journal of Nanjing university (natural science), 2019, 55 (1): 29 – 40
[12] ZHANG Y, YANG J. Chinese NER using lattice LSTM [EB/OL]. (2018-07-05) [2020-05-01]. https://arxiv.org/abs/1805.02023
[13] WANG L, XIE Y, ZHOU J S, et al. Fragment level Chinese named entity recognition based on neural network [J]. Journal of Chinese information processing, 2018, 32 (3): 84 – 90, 100. DOI: 10. 3969/j.issn.1003-0077.2018.03.012
[14] LIU X J, GU L C, SHI X Z. Named entity recognition based on BiLSTM and attention mechanism [J]. Journal of luoyang institute of technology, 2019, 29 (1): 65 – 70
[15] LIU W, XU T G, XU Q H, et al. An encoding strategy based word-character LSTM for Chinese NER [C]//Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2019: 2379 – 2389
[16] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [EB/OL]. (2018-10-11)[2020-05-01]. https://arxiv.org/abs/1810.04805
[17] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [EB/OL]. (2013-09-07)[2021-05-01]. https://arxiv.org/abs/1301.3781
[18] PENNINGTON J, SOCHER R, MANNING C. Glove: global vectors for word representation [C]//Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2014. DOI: 10.3115/v1/d14-1162
[19] PETERS M E, NEUMANN M, IYYER M, et al. Deep Contextualized Word Representations [C]//Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2018: 2227 – 2237. DOI: 10.18653/v1/N18-1202
[20] VASWANI A, SHAZEER N, PARMAR N. et al. Attention is all you need [C]// Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems. NIPS, 2017: 5998 – 6008
[21] JOZEFOWICZ R, ZAREMBA W, SUTSKEVER I. An empirical exploration of recurrent network architectures [C]//32nd International Conference on Ma-

chine Learning. JMLR, 2015: 2342 – 2350
[22] GUO D, ZHENG Q F, PENG X J, et al. Face detection detection, alignment alignment, quality assessment and attribute analysis with multi-task hybrid convolutional neural networks [J]. ZTE Communications, 2019, 17(3): 15 – 22. DOI: 10.12142/ZTECOM.201903004
[23] GRAVES A, SCHMIDHUBER J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures [J]. Neural networks, 2005, 18(5/6): 602 – 610. DOI: 10.1016/j.neunet.2005.06.042
[24] TAN Z X, WANG M X, XIE J, et al. Deep semantic role labeling with self-attention [EB/OL]. (2017-12-05)[2020-05-01]. https://arxiv.org/abs/1712.01586
[25] SHEN T, ZHOU T Y, LONG G D, et al. DiSAN: directional self-attention network for RNN/CNN-free language understanding [EB/OL]. (2017-11-20)[2020-05-01]. https://arxiv.org/abs/1709.04696
[26] LAFFERTY J, MCCALLUM A, PEREIRA F. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C]//18th International Conference on Machine Learning 2001 (ICML 2001). ACM, 2001: 282 – 289
[27] ZHU Y Y, WANG G X, KARLSSON B F. CAN-NER: Convolutional attention network for Chinese named entity recognition [EB/OL]. (2019-04-30)[2020-05-01]. https://arxiv.org/abs/1904.02141
[28] VITERBI A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm [J]. IEEE transactions on information theory, 1967, 13(2): 260 – 269. DOI: 10.1109/TIT.1967.1054010
[29] SI N W, WANG H J, LI W, et al. Chinese part of speech tagging model based on attentional long-term memory network [J]. Computer science, 2018, 45 (4): 66 – 70
[30] LEVOW G A. The third international Chinese language processing bakeoff: word segmentation and named entity recognition [C]//Fifth SIGHAN Workshop on Chinese Language Processing. Association for Computational Linguistics, 2006: 108 – 117
[31] CHEN A T, PENG F C, SHAN R, et al. Chinese named entity recognition with conditional probabilistic models [C]//Fifth SIGHAN Workshop on Chinese Language Processing. Association for Computational Linguistics, 2006: 173 – 176
[32] ZHANG S X, QIN Y, WEN J, et al. Word segmentation and named entity recognition for sighan bakeoff3 [C]//Fifth SIGHAN Workshop on Chinese Language Processing. Association for Computational Linguistics, 2013: 158 – 161
[33] ZHOU J S, QU W G, ZHANG F. Chinese named entity recognition via joint identification and categorization [J]. Chinese journal of electronics, 2013, 22 (2): 225 – 230
[34] LU Y N, ZHANG Y, JI D H. Multiprototype Chinese character embedding [C]// Tenth International Conference on Language Resources and Evaluation. Association for Computational Linguistics, 2016: 855-859
[35] DONG C H, ZHANG J J, ZONG C Q, et al. Character-based LSTM-CRF with radical-level features for Chinese named entity recognition [M]//Natural Language Understanding and Intelligent Applications. Cham, witzerland: Springer International Publishing, 2016: 239 – 250. DOI: 10.1007/978-3-319-50496-4_20
[36] LI N, GUAN H M, YANG P, et al. Chinese named entity recognition method based on BERT-IDCNN-CRF [J]. Journal of shandong university (science edition), 2020, 55 (1): 102 – 109
[37] XIE T, YANG J N, LIU H. Chinese entity recognition based on BERT-BiLSTM-CRF model [J]. Computer systems & applications, 2020(7): 48 – 55
[38] HE H. HanLP: Han language processing [EB/OL]. (2020-04-30)[2020-07-01]. https://github.com/hankcs/HanLP

Biographies

LI Daiyi (lidaiyi@nuaa.edu.cn) received his master's degree from School of computer and Communication Engineering, Zhengzhou University of Light Industry, China in 2018. He is studying for a doctor's degree in the School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, China. His main research interests are knowledge graphs and big data.

TU Yaofeng received his Ph.D. degree from Nanjing University of Aeronautics and Astronautics, China. He is a researcher at ZTE Corporation. His research interests include big data, database and machine learning.

ZHOU Xiangsheng is an expert and senior R&D manager in the AI field of ZTE Corporation. His research fields mainly include NLP, NAS, training acceleration, etc.

ZHANG Yangming is a software engineer at ZTE Corporation. His research interests mainly focus on natural language processing, knowledge engineering and acoustic signal processing.

MA Zongmin received his Ph. D. degree from the City University of Hong Kong, China and is a full professor with Nanjing University of Aeronautics and Astronautics, China. His research interests mainly include big data and knowledge engineering. He has published more than 100 papers in highly cited international journals and authored five monographs published by Springer. He is the Fellow of IFSA and Fellow of IET.