NAMED ENTITY RECOGNITION IN THANGKA FIELD BASED ON BERT-BiLSTM-CRF-a

Xiaoran GUO[1], Sujie CHENG [2], Weilan WANG*[3]

Thangka is one of the precious intangible cultural heritages, which is closely related to Tibetan Buddhism. However, Tibetan Buddhism has a complex system, and the naming patterns of various deities are not fixed and difficult to identify from Chinese texts. In this paper, we propose a multi-neural network fusion named entity recognition model BERT-BiLSTM-CRF-a which is based on the BERT pre-training language model, Bidirectional Long-and-Short Term Memory (BiLSTM) and Conditional Random Field (CRF). Specifically, the model uses the BERT to enhance the dynamic representation ability. Then, a weighting method from attention mechanism is introduced to weight the forward and backward BiLSTM hidden layer vectors before concatenating to further improve the effective utilization of context features. Finally, CRF model is used to output the global optimal annotation results. Experimental results on the test sets show that the recall of the BERT-BiLSTM-CRF-a model is 87.4%, 8.2% higher than the traditional named entity recognition model BiLSTM-CRF, and the F1 value is also 4.8% higher. Therefore, the model we proposed can be effectively used in the task of named entity recognition in thangka field.

Keywords: thangka; named entity recognition; BERT; attention mechanism; LSTM

1. Introduction

Thangka is a special art form of painting and an important part of Tibetan culture. And thangka is closely related to Tibetan Buddhism which refers to the branch of Buddhism that has been introduced into Tibet of China. Because Tibetan Buddhism is jointly influenced by Han Buddhism, Indian Buddhism and Tibetan Bonism, so it has a large system and includes many denominations. The deities of Tibetan Buddhism have complex origins, including Buddhas, Bodhisattvas, Arhats, Dhammapalas, local deities heavens and various kinds of ghosts and elves. A considerable part of the Chinese names of deities are derived from transliteration or free translation of Sanskrit, Pali and Tibetan. Therefore,

[1] Associate Prof., Dept. of Mathematics and Computer Science, Northwest Minzu University, China, e-mail: guoxr1982@163.com

[2] M.S. candidate, Dept. of National Languages Information Technology, Northwest Minzu University, China, e-mail: chengsujie10742@163.com

[3] Prof., Dept. of Key Laboratory of China's Ethnic Languages and Information Technology, Northwest Minzu University, China, corresponding author e-mail: wangweilan@xbmu.edu.cn

these Tibetan names, Sanskrit names, transliterated names, free translation names, abbreviations, full names, reputations, common names, etc. have no unified patterns and fixed lengths, so it is difficult to identify named entities from Chinese texts.

Named entity recognition is an important task in natural language processing (NLP) and the basis for entity relationship extraction, question answering system, and knowledge graph construction. It aims to extract entity names with specific meanings from unstructured texts. In the general field, named entities generally include person names, place names, organization names, time, and numeric expressions [1]. Yet specific fields are not limited to this, they mainly include the identification of specific entities and their categories. For example, in the bio-medicine field, the names of genes, proteins, diseases, drugs, etc. are entities to be extracted [2-3]; in the military field, troops, weapons, and foreign military targets are important entities [4-5]; in the field of finance and economics, it is necessary to extract important information such as company names, stock names, and stock codes [6-7].

Named entity recognition uses the method based on rules and dictionaries at the earliest. Linguistics experts manually construct rule templates to classify and recognize them by matching patterns and character strings. Such methods rely heavily on high-quality knowledge bases and perfect dictionaries, and their generalization ability is not strong. Since then, statistical methods based on large-scale corpora have gradually replaced rules. Common machine learning models include: Hidden Markov Model (HMM), Maximum Entropy Model (ME), Support Vector Machine (SVM), Conditional Random Field Model (CRF) and so on. Relatively speaking, the CRF model not only overcomes the problem of label bias, but also has a stronger feature fusion ability[8]. Adding rules on it can further improve the recognition effect[9,10]. However, it requires artificial features and expert knowledge which makes its cost relatively high.

In recent years, the method of deep learning has become the mainstream, and deep neural networks can automatically extract features for modeling, avoiding manual participation [11]. Bidirectional Long-and-Short Term Memory (BiLSTM) is an improved Recurrent Neural Network (RNN), which has the ability of long distance dependence in processing sequence data. Huang Z et al applied BiLSTM and CRF model for sequence labeling task for the first time [12]. Wang Lulu introduced the BiLSTM network in the recognition of Uighur named entities, which increased the value of F1-measure by 2.6% compared with that of the pure CRF model [13]. The entity recognition framework combined Convolutional Neural Network (CNN) with CRF's has also achieved good results [14,15].

In addition, the researchers further improved the recognition effect by using the attention mechanism [16] to highlight keywords in the model [17]. For

example, Rei M et al used attention mechanism to obtain the contextual representation of words in the full texts, which improved F1 value of entity recognition [18]. The Transformer model based on the attention mechanism has higher accuracy, recall rate and F value when used for named entity recognition of Chinese electronic medical records [19]. Another improvement is about the word embedding. The word embedding trained on the language model using large-scale corpus contains more semantic informations than the statistical features. It can also solve the problem of data sparseness in the high-latitude embedding space, of which the famous word embedding model is Word2Vec. However, its network has only three layers. The word embeddings obtained through this shallow network training have a better understanding of synonyms, but they cannot solve the problem of polysemy. In 2018, some pre-trained language methods are proposed such as BERT [20]. The full name of BERT is Bidirectional Encoder Representations from Transformers, which can dynamically generate word embeddings based on context information, and significantly improved the effect of multiple NLP tasks.

In this paper, we propose a BERT-BiLSTM-CRF-a model fused with a multi-layer neural network to identify and extract the names of deities in thangka description texts. The main work is as follows: (1)we use the BERT pre-training model in replace of shallow network to train the word embeddings to enhance the dynamic representation ability; (2)we use the BiLSTM network to extract the context information features for semantic expansion, and finally use CRF to decode and get the best recognition result; (3)we introduce the weighting method from attention mechanism to weight the forward LSTM hidden layer vectors and the backward LSTM hidden layer vectors before concatenating to further improve the effective utilization of context features. Experiments show that the final recall rate of the model on the test sets is 87.4%, which is 8.2% higher than the traditional sequence labeling model BiLSTM-CRF, and the F1 value is also 4.8% higher.

The rest of the paper is organized as follows. In Section 2, a multi-neural network fusion named entity recognition model based on the BERT, BiLSTM and CRF for thangka field is proposed. In Section 3, we present and analysis of the experiment. Conclusion and future work are given in Section 4.

2. Named entity recognition model

The overall structure of the named entity recognition model is shown in Fig. 1, which is mainly consists of three parts: the BERT layer, the BiLSTM layer and the CRF layer. The actual input of the model is Chinese sentences.

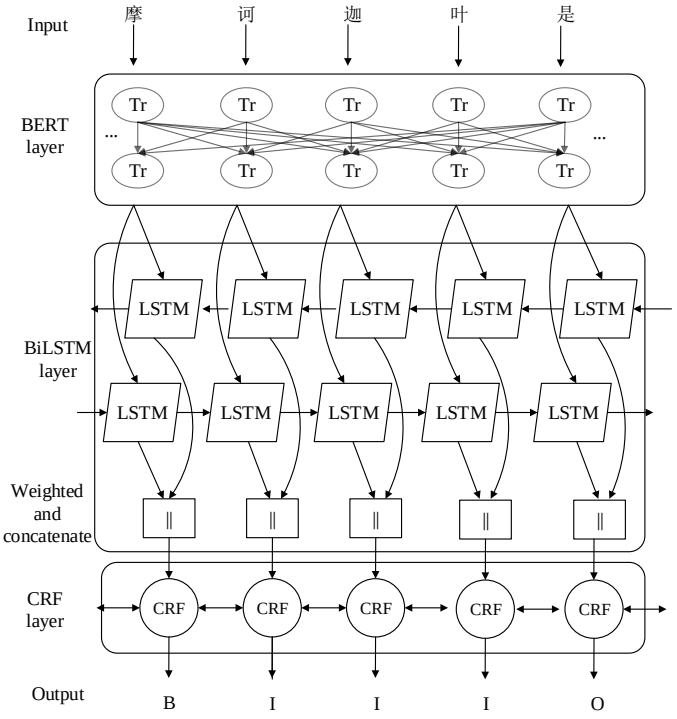

Fig.1. BERT-BiLSTM-CRF-a model

First, the embedding of the input texts are trained by the BERT pre-training language model. Through the look-up table operationis, the input data sequence of character embeddings will be sent to the BiLSTM layer, and the context informations are used for bidirectional semantic encoding. Then a weighting method from attention mechanism is introduced to calculate the weight coefficient of the forward hidden layer vector and the backward hidden layer vector before concatenating them. Finally, the semantic vector containing the context information is input into the CRF layer for decoding, and the label sequence with the highest probability is output, thereby obtaining the label category of each character. Among them, "B" indicates the beginning of the entity name, "I" indicates the other word of entity name, and "O" indicates the non-entity word.

2.1 BERT pre-training language model

BERT is a deep bidirectional language model based on Transformer encoder. The Transformer encoder is formed by stacking several identical units. Each encoder unit has two main sub-layers: a multi-head self-attention layer and a feed-forward neural network layer. Residual connection and data normalization

are used in the two sub-layers. As shown in Fig. 2, the core of its architecture is the self-attention mechanism.

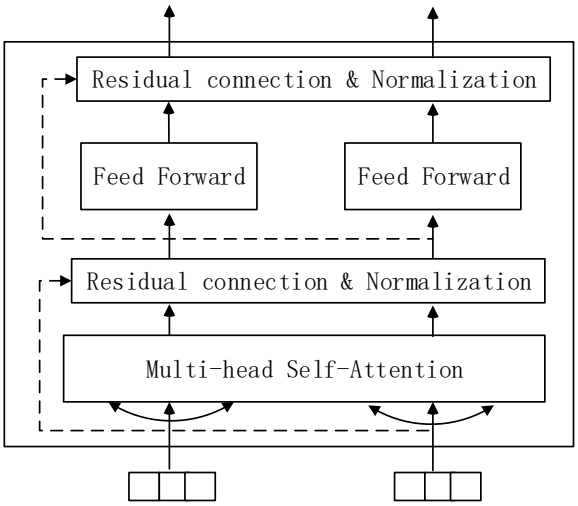

Fig. 2. Transformer encoder unit

BERT is a deep bidirectional language model based on Transformer encoder. The Transformer encoder is formed by stacking several identical units. Each encoder unit has two main sub-layers: a multi-head self-attention layer and a feed-forward neural network layer. Residual connection and data normalization are used in the two sub-layers. As shown in Fig. 2, the core of its architecture is the self-attention mechanism.

Attention mechanism is essentially an attention resource allocation model. At a specific moment, the main attention is focused on a certain key point of the thing and ignores other non-key points. In natural language processing, the texts that carry key and important information will be biased with higher weight values.

The self-attention mechanism calculates the probability of the correlation between words within a sentence. Suppose that the input is a matrix $A \in R^{l \times d}$, l is the length of the sequence and d is the dimension of the input. Through three different weight matrices W^q, W^k, W^v, A will be mapped to different spaces Q, K, V, as shown in formula (1). The dimension of the weight matrix is $R^{d \times d_k}$. The scaled dot product attention can be calculated by (2):

$$Q, K, V = AW^q, AW^k, AW^v \tag{1}$$

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2}$$

Where d_k is the dimension of self-attention layer, which can prevent QK^T being too large and play a role of expansion and contraction. After encoding based on the self-attention mechanism, the vectors contains not only the information of the word itself, but also the correlation information of the words in other before and after positions, so it is more abundant in semantic expression.

Transformer uses multi-head self-attention mechanism to expand the model's ability to focus on different positions and different semantic subspaces. Through self-attention multiple times without sharing parameters (W^q, W^k, W^v), and finally concatenate the results. It is calculated as formula (3) and (4), Where i is the number of self-attention heads, and the value range is $[1, h]$.

$$head_i = Attention(QW_i^q, KW_i^k, VW_i^v) \tag{3}$$

$$Multihead(Q, K, V) = Concat(head_1, \ldots, head_h) \tag{4}$$

After multi-head self-attention calculation, the next processing is residual connection and data normalization, which can solve the degradation problem of deep network training and speed up the training process. Then the output will be sent into the feed-forward network (FFN), which can be calculated by formula (5).

$$FFN(x) = \max\left(0, xW_1 + b_1\right)W_2 + b_2 \tag{5}$$

Where, W_1, W_2, b_1, b_2 are the learnable parameters in FFN. There is also a residual connection and normalization layer after FFN.

In order to learn the context information at the same time to achieve bidirectional feature extraction, BERT uses two unsupervised prediction tasks in the pre-training stage, namely the Masked LM and the Next Sentence Prediction task. Masked LM randomly masks 15% of the words in the corpus, of which 80% of the masked words are replaced with a mask [Mask], 10% remain unchanged, and 10% are replaced with random words. The masked words are predicted based on the remaining before and after vocabulary during training. Next Sentence Prediction trains sentence-level representation, it randomly replaces some sentences in the corpus, and uses a binary classification model to determine whether the second sentence B is the true subsequent sentence of the first sentence A. Use "[CLS] + Sentence A + [SEP] + Sentence B + [SEP]" to complete the splicing and bring it into the training model. [CLS] is the beginning of the sentence, and [SEP] is the separator between the sentences.

Since self-attention cannot capture the order of words in the text, add position coding to Transformer to indicate sequence information. So, the input of the BERT model is a linear sequence. The vector of each word is superimposed

by three parts, namely, the word embedding E_C, the text embedding E_S, and the position encoding E_P. As shown in Fig. 3.

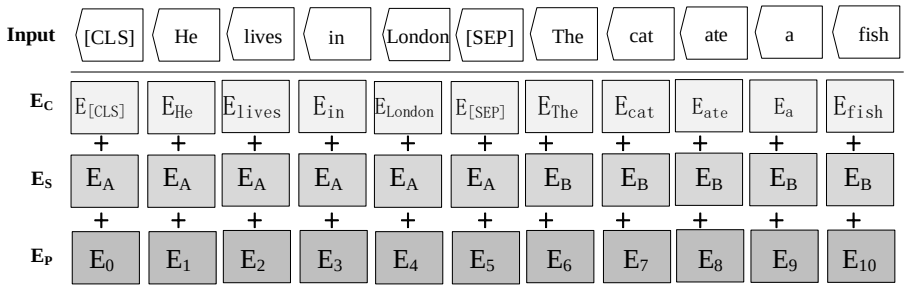

Fig. 3. Input vectors of BERT model

2.2 BiLSTM layer

LSTM overcomes the gradient disappearance and gradient explosion problems of traditional RNNs. Each unit has three gating units and a memory cell. The core idea is to use the gating mechanism to control the transfer and preservation of gradient information, as shown in Fig. 4.

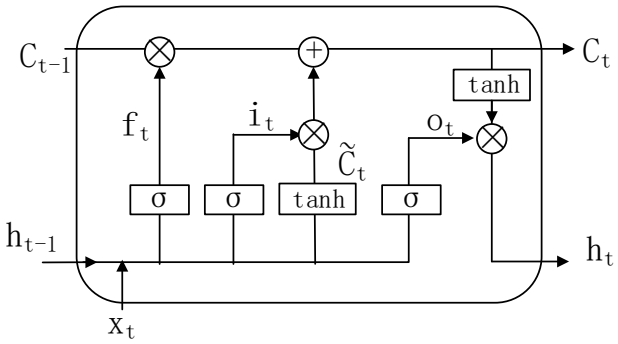

Fig. 4. LSTM unit

At a certain time t, the input gate i_t decides whether to record new information x_t into the cell. The forget gate f_t controls the degree of forgetting the information in the cell at the previous moment. The output gate o_t controls what information needs to be output. These three gates are all independently linearly calculated based on the output result h_{t-1} of the LSTM unit at the previous time and the input x_t at the current time, and use the sigmoid function for nonlinear activation, the calculation is as follows:

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right) \tag{6}$$

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right) \tag{7}$$

$$o_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right) \tag{8}$$

Among them: $W_i, W_f, W_o, b_i, b_f, b_o$ represent the weight matrix and offset term of input gate, forget gate and output gate respectively. σ represents the sigmoid function.

The memory cell C_t records historical information up to the current time and $\tilde{C}_t$ represents the candidate status obtained by the nonlinear function. The state value C_t of the current memory unit is updated jointly by its own state C_{t-1} at the previous moment and the current candidate value $\tilde{C}_t$. At the same time, the forget gate and the input gate are added to adjust the two parts. $\odot$ represents the dot product of the vector elements.

$$\tilde{C}_t = tanh\left(W_C \cdot \left[h_{t-1}, x_t\right] + b_C\right) \tag{9}$$

$$C_t = f_t \square C_{t-1} + i_t \square \tilde{C}_t \tag{10}$$

Finally, the hidden layer h_t of LSTM at the current moment is determined jointly by the output gate and the state of the memory cell, as shown in the following formula:

$$h_t = o_t \square \tanh\left(C_t\right) \tag{11}$$

In this paper, bidirectional LSTM is used to extract the context information to obtain bidirectional semantic coding. The input text sequence $(x_1, x_2, ..., x_n)$ processed by BERT is correspondingly converted into a sequence of word vectors and input into this layer model. A unit of the forward LSTM layer calculates the current word x_t and the above-mentioned representation $\overrightarrow{h_t}$ on the left; a unit of the backward LSTM layer calculates the current word x_t and the below-mentioned representation $\overleftarrow{h_t}$ on the right.

Considering that the encoding representations in different directions contribute differently to the semantics of the text, this paper introduced a weighting method from attention mechanism to give different weight values to the forward and backward encodings. The forward encoding and backward encoding are weighted and then concatenated together to obtain a more reasonable semantic vector representation, as shown in formula (12) and (13), where $\square$ represents concatenate. The weight coefficient a is obtained through two-layer neural

network training, W_a^1, W_a^2, W_a^3 represent the weight matrix of the network respectively.

$$h_t = a \cdot \overrightarrow{h_t} \oplus (1-a) \cdot \overleftarrow{h_t} \tag{12}$$

$$a = \sigma(W_a^3 tanh(W_a^1 \overrightarrow{h_t} + W_a^2 \overleftarrow{h_t}) \tag{13}$$

2.3 CRF layer

CRF is a discriminant model based on conditional probability. It models the target sequence based on the given observation sequence and focuses on solving the problem of serialization annotation. In the identification of named entities, there is a strong constraint between the tags. CRF can consider the relationship between context tags and obtain a globally optimal label sequence.

The output sequence $(h_1, h_2, ..., h_n)$ of the BiLSTM layer converts into the probability $(r_1, r_2, ..., r_n)$ of each label through a softmax layer and then is sent to the CRF layer. Assuming that one of the possible prediction sequences is $(y_1, y_2, ..., y_n)$, the evaluation score can be calculated by formula (14):

$$s(R, \mathrm{y}) = \sum_{i=0}^{n} T_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i} \tag{14}$$

Where T is the transition matrix, $T_{i,j}$ indicates the transition probability from label i to label j, and P_{i,y_i} represents the score of the y_i label of the character. The probability of sequence y appearing in all possible prediction sequences is:

$$p(\mathrm{y} \mid \mathrm{R}) = \frac{e^{s(R,y)}}{\sum_{\tilde{y} \in Y_r} e^{s(R,\tilde{y})}} \tag{15}$$

During CRF training, a maximum likelihood estimation method is introduced to define the loss function:

$$L = s(R, \mathrm{y}) - \log\left(\sum_{\tilde{y} \in Y_r} e^{s(R,\tilde{y})} \right) \tag{16}$$

The training process is performed on the training set $\{(h_i, y_i)\}$, and the candidate label sequence with the highest probability is selected as the final result of entity recognition.

3. Experiment

3.1 Experimental corpus

The source of the experimental corpus consists of two parts. First, we selected 527 entries related to Tibetan Buddhist deities from the "Buddhist Dictionary". Second, we selected 896 description texts corresponding to the thangka images. After data processing, 2436 sentences were obtained as experimental corpus, containing 5515 deity name entities.

In order to avoid the negative impact of word segmentation errors, we splited all texts into single Chinese character and marks them according to the BIO labeling strategy. That is, the starting word label of the deities named entity is "B", and the other word in the entity is "I", and the non-entity word is "O". Examples of corpus and label formats are shown in the table 1. Due to the small size of the corpus, we divided the training set, validation set and test set according to the ratio of 6: 2: 2 in this paper. The distribution of Chinese names of deities in the data set according to length is shown in Table 2:

Table 1

Examples of corpus and label formats

Tokens	药	师	王	佛	,	主	管	东	方	琉	璃	世
Labels	B	I	I	I	O	O	O	O	O	O	O	O
Tokens	界	,	两	位	胁	侍	是	日	光	普	照	菩
Labels	O	O	O	O	O	O	O	B	I	I	I	I
Tokens	萨	和	月	光	普	照	菩	萨	。			
Labels	I	O	B	I	I	I	I	I	O			

Table 2

Distribution of Chinese names of deities in the data set

Data set	Deity name length / Quantity										total
	2	3	4	5	6	7	8	9	10	>10	
training set	276	886	1384	543	199	81	45	10	17	10	3451
validation set	61	289	405	145	54	17	17	4	12	4	1008
test set	55	271	476	188	37	14	6	2	5	2	1056

3.2 Evaluation criteria

The experimental results use recognition accuracy rate (P), recall rate (R) and F1-measure as evaluation indicators. P refers to the percentage of the number of entities correctly identified by the model in the total number of entities identified, and R refers to the percentage of correctly identified entities by the model in all entities in the test set. F1 is the harmonic mean value of P and R. Considering the model performance comprehensively, the calculation is as follows:

$$F1 = \frac{2PR}{P+R} \times 100\% \tag{17}$$

3.3 Experimental process

In order to verify the effectiveness of the model we proposed, four sets of comparative experiments were carried out. The models used were as follows:

(1) BiLSTM-CRF model. This model is a classic model to solve the problem of sequence labeling. In the experiment, Word2Vec model is used to train word vectors on unlabeled corpus where each word vector has a dimension of 256×1, and then input them into BiLSTM-CRF model to train;

(2) BiLSTM-CRF-a model. This model introduce weighting method on the basis of model (1), weight the forward and backward hidden layer vectors processed by BiLSTM and then concatenate;

(3) BERT-BiLSTM-CRF model. This model use BERT pre-training language model instead of Word2Vec in model (1) to train word vectors;

(4) BERT-BiLSTM-CRF-a model, This is the model proposed in this paper, see section 2.

Google provides two Bert pre-training language models, one is BERT-Base and the other is BERT-Large. BERT-Base is used in the experiment, which is composed of 12 layers of bidirectional Transformer encoder, with a hidden layer size of 768, 12 self-attention heads, and a total of 110M parameters. The maximum sequence length of the training corpus is 100, the size of each batch is 16, and the learning rate is 5×10^{-5}. The LSTM hidden layer size is set as 256, and the number of layers is 2. In order to prevent the network from overfitting, random dropout is performed on the LSTM circulation layer and the fully connected layer, with a value of 0.5. The parameter of the fully connected layer after passing through the CRF layer is 3, that is, it is divided into 3 categories, and each category is a 3×1 dimension one-hot vector, and the maximum value is output as the labeling result. Models (1) and (2) reach the maximum value of F1 when training 30 epochs. Models (3) and (4) reach the maximum F1 after 10 epochs.

4. Results

The recognition results of named entities on the test set of different models are shown in Table 3.

Table 3

Named entity recognition results of different models

Models	P(%)	R(%)	F1(%)
BiLSTM-CRF	87.5	79.2	83.1
BiLSTM-CRF-a	88.0	79.5	83.5
BERT-BiLSTM-CRF	87.6	86.0	86.8

BERT-BiLSTM-CRF-a	88.4	87.4	87.9

The identification of different lengths of the name of the deity by BERT-BiLSTM-CRF-a model on the test set is shown in Table 4.

Table 4

The recognition results of BERT-BiLSTM-CRF-a Model on the test set

	Deity name length					Total (quantity)
	1	2	3~5	6~9	>=10	
Test set	0	55	935	64	2	1056
Model Identification	7	75	909	51	2	1044
Correct identification	0	39	836	48	0	923
Identification accuracy	-	52.0 %	91.9%	94.1%	0%	88.4%

Comparing the BERT-BiLSTM-CRF model and the BiLSTM-CRF model, the BERT pre-training language model has a 0.1% increase in the precision rate, but a 6.8% increase in the recall rate, and a 3.7% increase in the F1 value. It shows that the deep neural network in BERT can better express semantic information than the word vector trained by the shallow network Word2Vec. The training method of BERT's Random Mask can enhance the bidirectional representation ability and migration ability, and it can also obtain better recognition results on small-scale corpus.

At the same time, comparing BiLSTM-CRF-a model and BiLSTM-CRF model, as well as BERT-BiLSTM-CRF-a model and BERT-BiLSTM-CRF model, weighting forward LSTM vector and backward LSTM vector before concatenating in this paper is effective. That is, to distinguish the contribution of the above and below information of the current word in the model, and give the more important parts higher weights, thereby increasing the accuracy rate by about 1% - 2%.

Here are some examples of identification results by different models.

Sentence 1: "燃灯古佛，音译提和竭罗、提洹竭。"

BERT-BiLSTM-CRF-a model recognition results {"燃灯古佛", " 提和竭罗", "提洹竭"}

BiLSTM-CRF model recognition results {"燃灯古佛", " 提 ", "竭罗", "提洹竭"}

Among them, the BiLSTM-CRF model cannot correctly identify "提和竭罗", the main reason is that the Chinese character "和" in sentence is often used as the conjunction connecting the names of two entities, and it is relatively rare to appear in entity names. In the BERT-BiLSTM-CRF-a model, the semantic vector of the Chinese character "和" after BERT training is context-

sensitive. The context of the sentence indicates that the Chinese character "和" is the name of the entity name, not the connecting word.

Sentence 2: "大日如来亦现菩萨形"

BERT-BiLSTM-CRF-a model recognition result {"大日如来"}

BiLSTM-CRF model recognition results {"大日如来", "现菩萨"}

The two Chinese characters "菩萨" often appears as the suffix of the entity name. The BiLSTM-CRF model mistakenly used "现菩萨" as the recognition result, while the BERT-BiLSTM-CRF-a model did not commit such mistake.

5. Conclusions

In this paper, based on the BiLSTM-CRF sequence labeling model, the BERT pre-training language model and attention mechanism are introduced to improve the effect. On the one hand, the word embedding's representation ability is improved, and on the other hand, the effective utilization rate of the context semantic vector is improved. The experimental results show that the proposed BERT-BiLSTM-CRF-a model fused with multi-layer neural network has greater improvement in recall rate and F1value, and it also has a better recognition effect on entities with longer names.

Acknowledgment

This work was supported by the National Civil Affairs Commission Innovation Team Plan under Grant No.2018(98); Central University Young Teachers Innovation Project under Grant No. 31920200067.

R E F E R E N C E S

[1]. *Chinchor N*, "MUC-6 named entity task definition (version2.1)", in proceedings of the 6th Message Understanding Conference, Columbia, Maryland, Nov., 1995. Stroudsburg: ACL, 1995, pp. 317-332

[2]. *Bin Ji,Shasha Li,Jie Yu*, "Research on Chinese medical named entity recognition based on collaborative cooperation of multiple neural network models", in Journal of Biomedical Informatics, 2020, 104(103395).

[3]. *Indra Neil Sarkar,Andrew Georgiou,Paulo Mazzoncini de Azevedo Marques,Yonghui Wu,Min Jiang,Jianbo Lei,Hua Xu*, "Named Entity Recognition in Chinese Clinical Text Using Deep Neural Network", in Studies in Health Technology and Informatics, 2015, pp. 624-628.

[4]. *Xu Shukui, Cao Jinran.* "Military target entity recognition method based on hierarchical bilateral long term and short term memory neural network conditional random field model", in Computer Science, **vol. 45**, no. 6, 2019, pp. 18-22,46

[5]. *Han Xinxin, Ben Kerong, Zhang Xia*, "Research on named entity recognition technology in the field of military software testing", in Computer Science and Exploration, doi: 10.3778/j.issn.1673-9418.1906031 .

[6]. *Yu Chuanming, Huang Jianqiu, Guo Fei*, "Identifying named entities from customer reviews implementation based on maximum entropy model", in Data Analysis and Knowledge Discovery, **vol. 27**, no. 5, 2011, pp. 77-82

[7]. *Chen Chao, Zhu Hongbo, Wang Yaqiang*, "Automatic recognition of company name abbreviations in Chinese financial texts", in Journal of Sichuan University(Natural Science Edition), **vol. 48**, no. 2, 2011, pp. 308-314

[8]. *Wang Pengyuan, Ji Donghong*, "Disease name extraction based on multi-label CRF", in Application Research of Computers, **vol. 34**, no. 1, 2017, pp. 118-122

[9]. *Li Wei, Zhao Dazhe, Li Bo, et al*, "Combining CRF and rule based medical named entity recognition", in Application Research of Computers, **vol. 32**, no. 4, 2015, pp. 1082-1086.

[10]. *Konkol M, Konopik M*, "CRF-based Czech named entity recognizer and consolidation of Czech NER research", in Proceedings of the 2014 International Conference on Text Speech and Dialogue, Brno, Czech Republic, Sept., 2014.

[11]. *Liu Liu, Wang Dongbo*, "A review of research on named entity recognition", in Journal of Information, **vol. 37**, no. 3, 2018, pp. 329-340.

[12]. *Huang Z, Xu W, Yu K*, "Bidirectional LSTM-CRF models for sequence tagging", arXiv preprint arXiv:1508.01991, 2015.

[13]. *Wang Lulu*, "Research on Uyghur named entity recognition based on deep neural network", in Journal of Chinese Information Processing, **vol. 33**, no. 3, 2019, pp. 64-70.

[14]. *Ma X, Hovy E*, "End-to-end sequence labeling via Bidirectional LSTM-CNNs-CRF", in Meeting of the Association for Computational Linguistics, Berlin, Germany, Aug., 2016.

[15]. *Cao Yiyi, Zhou Yinghua, Shen Faha*, "Recognition of named entities for Chinese electronic medical records based on CNN-CRF", in Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), **vol. 31**, no. 6, 2019, pp. 869-874.

[16]. *Vaswani A , Shazeer N , Parmar N*, "Attention Is All You Need", arXiv preprint arXiv:1706.03762, 2017.

[17]. *Yang Pei, Yang Zhihao, Luo Ling*, "Chemical entity named entity recognition based on attention mechanism", in Computer Research and Development, **vol. 55**, no. 7, 2018, pp. 194-202.

[18]. *Rei M, Crichton G K O, Pyysalo S*, "Attending to Characters in Neural Sequence Labeling Models", arXiv preprint arXiv: 1611. 04361, 2016.

[19]. *Li Bo, Kang Xiaodong, Zhang Huali*, "Named entity recognition of Chinese electronic medical recordsusing Transformer-CRF",in Computer Engineering and Applications, **vol. 56**, no. 5, 2019, pp. 153-159.

[20]. *Devlin J, Chang M W, Lee K*, "Bert: Pre-training of deep bidirectional transformers for language understanding", arXiv preprint arXiv: 1810. 04805, 2018.