Chinese named entity recognition model based on BERT

Hongshuai Liu[1], *Ge* Jun[1,*] , and *Yuanyuan* Zheng[1]

[1]Modern Post College, Nanjing University of Posts and Telecommunications, Nanjing, China

Abstract. Nowadays, most deep learning models ignore Chinese habits and global information when processing Chinese tasks. To solve this problem, we constructed the BERT-BiLSTM-Attention-CRF model. In the model, we embeded the BERT pre-training language model that adopts the Whole Word Mask strategy, and added a document-level attention. Experimental results show that our method achieves good results in the MSRA corpus, and F1 reaches 95.00%.

1 Introduction

In 1991, a paper on company name recognition was published at the IEEE Conference on Artificial Intelligence [1]. Since then, a branch of named entity recognition has appeared in the field of natural language processing research. The initial named entity recognition was mainly based on dictionary [2] and rule-based matching [3]. Then it is mainly based on machine learning methods, these methods include Hidden Markov Model, Maximum Entropy Model, Support Vector Machine and Conditional Random Field.

Nowadays, the performance of computers has been continuously improved. Therefore, named entity recognition based on deep learning has gradually become a research hotspot in the field of natural language processing.

In recent years, related research work has improved the named entity recognition model. Collobert et al. [4] proposed the use of convolutional neural networks to realize named entity recognition. Huang et al. [5] proposed the use of bi-directional long and short-term memory network combined with artificially designed features to realize named entity recognition, and the F1 reached 84.83% on the CoNLL2003 data set. Ma et al. [6] added a convolutional neural network for extracting character-level representations of words in the model, and the F1 reached 91.21%. Luo et al. [7] used the BiLSTM-CRF model combined with the attention, and the F1 on the BioCreativeIV dataset reached 91.14%. Wu et al. [8] proposed joint word segmentation training with the CNN-BiLSTM-CRF model, and at the same time processed samples with the help of pseudo-labels, which further improved the performance of entity recognition. Peters et al. [9] used BiLSTM to generate contextual representations through pre-trained language character models and achieved good results. Jana Straková et al. [10] applied the BERT pre-processing model to named entity recognition and F1 reached 93.38% on the CoNLL-2003English data set.

* Corresponding author: gej@njupt.edu.cn

2 Our frame

Our model framework is BERT-BiLSTM-Attention-CRF, which consists of a BERT embedding layer, a bi-directional long short-term memory network layer, an attention layer and a conditional random field layer. The structure of the model is shown in Figure 1.

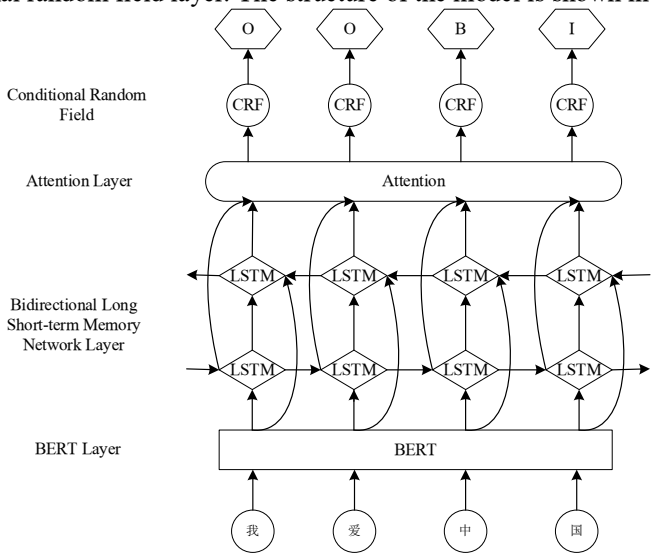

Fig. 1. Our frame.

2.1 BERT model

BERT is a general language pre-training model proposed by the Google artificial intelligence team in 2018[11]. It can represent words in vector form to obtain the similarity between words. BERT uses a bidirectional Transformer neural network as an encoder, so that the prediction of each word can refer to contextual information. The model also proposes a "mask language model" and a "next sentence prediction task model" to capture word-level and sentence-level feature representations. The Mask language model masks 15% of the information in the corpus to maximize the representation of each word in the model. The BERT model is shown in Figure 2.

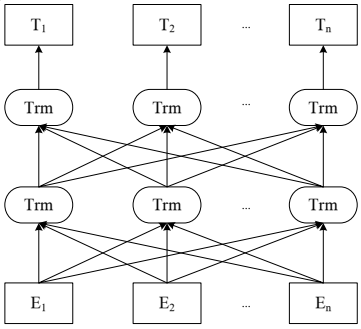

Fig. 2. BERT model.

Different from the BERT in the Single Word Mask (SWM)[12], we use the Whole Word Mask(WWM)[13] in accordance with the Chinese habit when pre-processing the sentence.

In the Whole Word Mask, if the part of a word is masked, the other parts that belong to the word will be masked, as shown in Table 1.

Table 1. The Whole Word Mask.

Original Text	Tomorrow is a good weather
Single Word MASK	[MASK] Tomorrow is a good [MASK]weather
Whole Word MASK	[MASK][MASK] is a good [MASK][MASK]

2.2 BiLSTM model

LSTM (Long-Short Term Memory) is an improved model of Recurrent Neural Network (RNN). It solves the problem of gradient explosion or gradient disappearance that occurs when RNN processes long sequence information. LSTM adds a memory unit, forgetting gates, input gates and output gates [14], alleviating the problem of long sequence forgetting. The structure is shown in Figure 3.

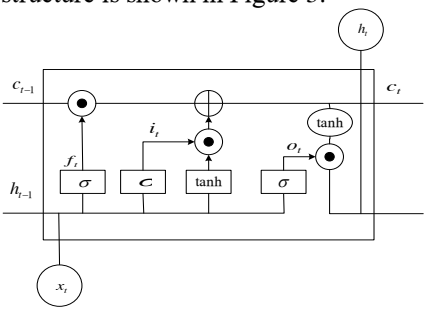

Fig. 3. LSTM structure.

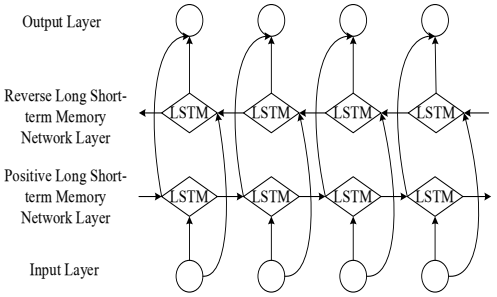

Fig. 4. BiLSTM model.

The LSTM calculation formula is as follows:

$$f_t = \sigma\left(W_f h_{t-1} + U_f x_t + b_f\right) \tag{1}$$

$$i_t = \sigma(W_i h_{t-1} + U_i x_t + b_i) \tag{2}$$

$$o_t = \sigma(W_o h_{t-1} + U_o x_t + b_o) \tag{3}$$

$$\tilde{c}_t = tanh(W_c h_{t-1} + U_c x_t + b_c) \tag{4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{5}$$

$$h_t = o_t \odot tanh(c_t) \tag{6}$$

In which, f_t represents the forget gate, the function is to decide what information to discard from the current cell state; i_t represents the input gate, the function is to decide how much to choose from the newly acquired information to update the state; o_t represents the output gate, which determines how much information generates the hidden layer State variables; U_f , U_i and U_o are their adjustable parameters; W_f , W_i and W_o are their weights; b_f , b_i and b_o are their biases; σ is the sigmoid activation function; tanh is the hyperbolic tangent activation function.

Since the one-way LSTM model can only learn the above information, it cannot use the following information, which limits the effect of entity recognition. The BiLSTM (Bi-directional Long-Short Term Memory) proposed by GravesA et al. [15] consists of a forward LSTM and a backward LSTM. The basic idea is to take each word sequence separately

forward propagation and backward propagation, and then concatenate the output at the same time. The structure is shown in Figure 4.

2.3 Attention model

Aiming at the characteristics of entity naming methods in Chinese texts and uneven distribution of entities, we introduce a document-level attention mechanism to focus on the global information of the document, while increasing the similarity evaluation of the cosine distance score. We use $D = (S_1, S_2, \dots, S_n)$ to represent the document, which contains n sentences S. Each sentence $S = (W_1, W_2, \dots, W_m)$ contains m words. The output sequence of BiLSTM is input to the A matrix, so as to obtain the correlation degree between the current character W_i and all words and the global feature representation g_i of the target word. The calculation formula is:

$$\mathrm{g_i} = \sum_{\mathrm{j=1}}^{\mathrm{N}} \mathrm{A_{i,j}} \mathrm{h_j} \tag{7}$$

$$\mathrm{A_{i,j}} = \frac{\exp\left(\mathrm{score}\left(\mathrm{w_i}, \mathrm{w_j}\right)\right)}{\sum_{\mathrm{k=1}}^{\mathrm{m}} \exp\left(\mathrm{score}\left(\mathrm{w_i}, \mathrm{w_k}\right)\right)} \tag{8}$$

$$\mathrm{score} = \left(\mathrm{w_i}, \mathrm{w_j}\right) = \frac{\mathrm{W_a}\left(\mathrm{w_i}, \mathrm{w_j}\right)}{|\mathrm{w_i}||\mathrm{w_j}|} \tag{9}$$

$$\mathrm{c_i} = \tanh\left(\mathrm{W_g}[\mathrm{g_i}, \mathrm{h_i}]\right) \tag{10}$$

where $A_{i,j}$ represents the attention weight of the character w_i and the other character w_j in the document, h_j represents the output of the BiLSTM layer, $score$ represents the similarity score of the two characters using cosine distance, $W_a$and$W_g$represents the parameter matrix learned during the training process, and c_i is the output of attention layer.

2.4 CRF model

In the task of named entity recognition, the dependency between adjacent tags is also one of the factors that cannot be ignored. CRF can obtain an optimal prediction sequence through the relationship between adjacent tags, which can make up for the shortcomings of BiLSTM. For any sequence $X = (x_1, x_2, \dots, x_3)$, assume that $\mathbf{P}$ is the output matrix of BiLSTM, and the size of $\mathbf{P}$ is $n \times k$, where n is the number of words, k is the number of tags, and $\mathbf{P}_{i,j}$ is the score of the jth tag of the ith word. For the prediction sequence $Y = (y_1, y_2, \dots, y_n)$, The calculation formula is as follows:

$$s(X, Y) = \sum_{i=0}^{n} \boldsymbol{A}_{y_i, y_{i+1}} + \sum_{i=1}^{n} \boldsymbol{P}_{i, y_i} \tag{11}$$

$$p(Y|X) = \frac{e^{s(X,Y)}}{\sum_{\tilde{Y} \in Y_X} s\left(X, \tilde{Y}\right)} \tag{12}$$

$$\ln\big(p(Y|X)\big) = s(X,Y) - \ln\left(\sum_{\tilde{Y}\in Y_X} s\big(X,\tilde{Y}\big)\right) \tag{13}$$

$$Y^* = \underset{\tilde{Y}\in Y_X}{\operatorname{argmax}}\, s\big(X,\tilde{Y}\big) \tag{14}$$

where $\mathbf{A}$ represents the transition score matrix, $\mathbf{A}_{i,j}$ represents the score of label i transitioning to label j, and the size of $\mathbf{A}$ is $k+2$. $\tilde{Y}$represents the true label sequence, Y_X represents all possible labels sequence. $s(X,Y)$ is the score function of the predicted sequence Y. $p(Y|X)$ is the probability of the occurrence of the predicted sequence Y. $\ln\big(p(Y|X)\big)$ is the likelihood function of Y. Y^* is the output sequence of the maximum score.

3 Experiments

3.1 Experimental data set

The commonly used labeling modes for named entity recognition are BIO mode, BIOE mode and BIOES mode. We adopt the BIO mode, which has 7 tags, namely "O", "B-PER", "I-PER", "B- ORG", "I-ORG", "B-LOC", "I-LOC", where O is a non-named entity, B is the first word of a named entity, I is a non-first word of a named entity, PER is a person's name, ORG is an organization First name, LOC is geographic name.

We use MSRA corpus for model experiments. This data set is issued by Microsoft Research Asia and is a public Chinese data set in China. The MSRA data set contains 7 types of labels in three categories: geographic name, organization name, and person name.

Our experiment mainly identifies and evaluates people's names, geographic names, and organizations. The specific data set in the MSRA corpus is set to 46364 sentences in the training set and 4365 sentences in the test set.

3.2 Evaluation rule

The experiment in our paper uses precision rate, recall rate and F1 value as indicators to judge the accuracy of the model. The calculation formula is as follows:

$$P = \frac{TP}{TP+FP} \times 100\% \tag{15}$$

$$R = \frac{TP}{TP+FN} \times 100\% \tag{16}$$

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

In which, TP represents the number of correctly identified named entities; FP represents the number of incorrectly identified named entities; FN represents the number of unidentified named entities; P is the accuracy rate; R is the recall rate.

3.3 Experiments environment

We use Tensorflow1.14.0 to build the experimental model. Our computer memory is 32GB, the graphics card is NVIDIA GeForceRTX2070 and the python version is 3.6.9.

In Our experiment, Adam optimizer is used, the maximum input sequence length is 128, LSTM_dim is set to 200, batch_size is 16, and learning_rate is 5e-5. To prevent over-fitting problems, set drop_out_rate to 0.5.

3.4 Analysis of results

To prove the effectiveness of the model, We compare BERT-BiLSTM-Attention-CRF (Whole Word Mask) with the previous model The experimental results are shown in Table 2.

Table 2. Comparison of experimental results (%).

Model	*P*	*R*	*F1*
LSTM-CRF	83.44	80.75	82.07
BiLSTM-CRF	86.75	84.50	85.61
BERT-BiLSTM-CRF(SWM)	94.35	94.05	94.20
BERT-BiLSTM-CRF(WWM)	94.38	94.83	94.61
BERT-BiLSTM-Attention-CRF(WWM)	94.89	95.10	95.00

From the Table 3, we can see that the accuracy, recall, and F1 of our model on the MSRA data set have achieved the best results, and the F1 of the method in our paper reaches 95.00%. First of all, the performance on the MSRA data set shows that the F1 of BiLSTM-CRF is 3.54% higher than that of LSTM-CRF. It can be seen that the bi-directional structure of BiLSTM has a stronger ability to acquire context sequence features than the one-way structure. Comparing the BERT-BiLSTM-CRF (SWM) model and the BiLSTM-CRF model, it shows that the BERT pre-training language model has significantly improved named entity recognition, and its accuracy has increased by 8.59%. When the mask strategy of BERT is changed to the Whole Word Mask in the BERT-BiLSTM-CRF model, its F1 on the MASR data set increases by 0.41%, indicating that its feature extraction ability is stronger. The model in this paper introduces a document-level attention on the basis of BERT(WWM)-BiLSTM-CRF, and the F1 reaches 95.00%, indicating that the attention can enhance the feature extraction ability of the model under global information.

The F1 of the experiment in Our paper changes in the first 20 epochs as shown in Figure 5.

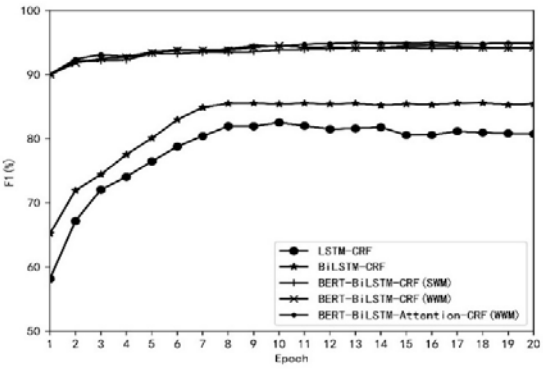

Fig. 5. F1 update.

As shown in the figure above, the neural network models BiLSTM-CRF and LSTM-CRF, which do not use BERT, have a low F1 at the beginning of training, and reach a stable level

after many iterations, but they are still lower than the three models using BERT. The three models of BERT-BiLSTM-CRF (SWM), BERT-BiLSTM-CRF (WWM), and BERT-BiLSTM-Attention-CRF (WWM) can reach a higher level after one round of training, the F1 reaches about 90%, and as the number of training rounds increases, the F1 continues to increase, eventually reaching a high and stable level.

We also compare the performance of the BERT-BiLSTM-Attention-CRF (WWM) model and other existing models on the MSRA corpus, as shown in Table 3.

Table 3. Comparison of other models (%).

Model	*P*	*R*	*F1*
CNN-BiLSTM-CRF[16]	91.63	90.56	91.09
DC-BiLSTM-CRF[17]	92.14	90.96	91.54
Lattice-LSTM-CRF[18]	93.57	92.79	93.18
BERT-IDCNN-CRF[19]	94.86	93.97	94.41
Our Model	94.89	95.10	95.00

As shown in the above table, the CNN-BiLSTM-CRF model uses convolutional neural networks and bidirectional LSTM to extract character feature sequences. The DC-BiLSTM-CRF model uses DC-LSTM to learn sentence features, combined with self-attention mechanism for entity recognition. Lattice-LSTM-CRF uses mesh LSTM for character feature extraction to complete entity recognition. The BERT-IDCNN-CRF model uses BERT (SWM) for word embedding, combined with iterative expansion convolution to extract sentence features, and its F1 reaches 94.41%, which is still far from the performance of our model. So we can see that the performance of the BERT-BiLSTM-Attention-CRF (WWM) model is better.

4 Conclusion

In our paper, the word vector is obtained by the BERT (WWM)language pre-processing model, and then the word vector is input into the BiLSTM-CRF and attention layer to construct the BERT-BiLSTM-Attention-CRF model. By verifying on the MSRA corpus, compared with other existing models, the BERT-BiLSTM-Attention-CRF model of the F1 reaches 95.00% on the MSRA corpus, which has the best performance. The BERT model using the Whole Word Mask strategy and document-level attention are the biggest advantages of our model. This makes the sequence features extracted by Our model conform to the Chinese habits, and can learn the word-level structural features and contextual semantic information. In the future, we consider applying it to entity recognition in the professional field.

Our work was sponsored by School-level scientific research fund of Nanjing University of Posts and Telecommunications (Grant No. NY220063).

References

1. Liu L, Dongbo W.A Survey of Named Entity Recognition Research[J].Journal of Information,2018,**37**(3):329–340.

2. Carol F , Alderson P O , Austin J H M , et al. A general natural-language text processor for clinical radiology.[J]. Journal of the American Medical Informatics Association, 1994, **1**(2):161-174.
3. R.Gaizauskas, G.Demetriou, and K.Humphreys. Term recognition and classification in biological science journal articles.In Computational Terminology for Medical &Biological Applications Work shop of the 2nd International Conference on NLP[C],2000,pp.37–44.
4. Collobert R , Weston J , Bottou, Léon, et al. Natural Language Processing (almost) from Scratch[J]. Journal of Machine Learning Research, 2011, **12**(1):2493-2537.
5. Huang Z , Xu W , Yu K . Bidirectional LSTM-CRF Models for Sequence Tagging[J]. Computer ence, 2015.
6. Ma X , Hovy E . End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF[J]. 2016.
7. Ling L , Zhihao Y , Pei Y , et al. An Attention-based BiLSTM-CRF Approach to Document-level Chemical Named Entity Recognition[J]. .Bioinformatics(8):**8**.2019.
8. Wu FZ, Liu JX, Wu CH, et al. Neural Chinese named entity recognition via CNN-LSTM-CRF and joint training with word segmentation. The World Wide Web Conference[C]. New York, NY, USA. 2019. 3342–3348.
9. Peters M E , Ammar W , Bhagavatula C , et al. Semi-supervised sequence tagging with bidirectional language models[J]. 2019.
10. Jana Straková, Straka M , Haji J . Neural Architectures for Nested NER through Linearization[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.
11. Devlin J , Chang M W , Lee K , et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.
12. Souza, Fábio, Nogueira R , Lotufo R . Portuguese Named Entity Recognition using BERT-CRF[J]. 2020.'
13. Cui Y , Che W , Liu T , et al. Pre-Training with Whole Word Masking for Chinese BERT[J]. 2019.
14. Hochreiter S , Schmidhuber J . Long Short-Term Memory[J]. Neural Computation, 1997, **9**(8):1735-1780.
15. Graves A , Jürgen Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks, 2005, **18**(5–6):602-610.
16. Jia Y Z, Xu X B. Chinese named entity recognition based on CNN-BiLSTM-CRF.2018 IEEE 9th International Conference on Software Engineering and Service Science [C].Beijing, China. 2019. 1–4.
17. Xiaojun L , Lichuan G, Xianzhang S. Named entity recognition based on Bi-LSTM and attention mechanism. Journal of Luoyang Institute of Technology (Natural Science Edition)[J], 2019,**29**(1): 65-70,77.
18. Zhang Y , Yang J . Chinese NER Using Lattice LSTM[J]. 2018.
19. Li N, Huanmei G, Piao Y, et al. Chinese named entity recognition method based on BERT-IDCNN-CRF. Journal of Shandong University (Science Edition)[J],2020, **55**(1):102–109.