Chinese named entity recognition based on MacBERT and Joint Learning

Dangguo Shao, Kun Huang*, Lei Ma, Sanli Yi

Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650504, China
1298335567@qq.com

2023 IEEE 9th International Conference on Cloud Computing and Intelligent Systems (CCIS) | 979-8-3503-0442-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/CCIS59572.2023.10262869

Abstract: Extracting medical entities from Chinese medical texts is of great significance to the establishment and application of medical information systems. Chinese text lacks spacers to segment words, which makes it difficult to recognize entities. Therefore, this paper proposes a named entity recognition method combining MacBERT and Joint Learning. The MLM as correction BERT (MacBERT) pretrained language model, which alleviates the differences between pre-training and fine-tuning stages, obtains dynamic word vector feature expression, and then enters into a framework that integrates the named entity recognition task and the word segmentation task for joint learning, so as to improve the feature capture ability of the model. Experiments show that the model can effectively obtain the important entities in the instruction of traditional Chinese medicine. On the Chinese medicine instruction dataset, fusing MacBERT with Joint Learning model achieved the best result with 68.43% F1, which increased by 1.6% compared with the two-layer BiLSTM non-joint model, and increased by 1.07% compared with the combined BERT with Joint Learning model. Experimental results show that the recognition performance is better than that of the non-joint model and other pre-trained models.

Key words: Named Entity Recognition; MacBERT; Joint Learning; Chinese Medical Text

1 Introduction

With the steady development of information construction in the medical field, the modern medical information system has accumulated a large amount of medical data, including electronic medical records, medical reports, Chinese medicine instructions, etc. [1]. In the massive biomedical information data, there are a lot of valuable academic resource information. In this epidemic situation, Chinese medicine plays an important role, which makes us realize that it is extremely important to inherit and innovate the important information of Chinese medicine. The construction of knowledge map of rational drug use of Traditional Chinese Medicine（TCM）.

Deep learning is a method that has been widely used and mature at present. Pre-trained language models have achieved good results in Natural Language Processing (NLP), such as BERT [2], ELMo [3], etc. Methods for Named Entity Recognition (NER) are constantly being proposed, most of which are designed for NER of English texts, and these studies are also aimed at the characteristics of English itself. Compared with the NER of English text, Chinese NER started late. Due to the language characteristics of Chinese itself, Chinese NER is more complicated [4]. First, the Chinese text lacks the obvious characteristic indication of the entity name that exists in the English text, such as capitalization, italics, etc. Second, Chinese entity names are more context-dependent. And the same entity in different texts, it belongs to different categories. Third, most of the entities in Chinese text are composed of single Chinese characters, and there is no natural separator for separation like English text, which makes it more difficult to identify the boundaries of entities.

It has been shown that combining NER model training with related tasks can improve NER performance. For example, Huang et al. [5] found that entity boundary is an important factor affecting Chinese NER. This paper combined Chinese Word Segmentation (CWS) task with NER model training as an auxiliary task, so as to improve the ability to recognize entity boundary. In the pre-training stage, BERT model cannot obtain Chinese word-level semantic features, resulting in the model failing to fully learn the feature representation of Chinese text. Cui et al. [6] proposed BERT's error-correcting mask pre-training model, namely MLM as correction BERT(MacBERT), in 2020. The main framework of the two models is completely consistent, but MacBERT alleviates the inconsistency between pre-training and downstream tasks.

2 System framework

This paper proposes Chinese named entity recognition based on MacBERT and Joint Learning, the main structure of which is shown in Figure 1.

979-8-3503-0442-8/23/$31.00 ©2023 IEEE

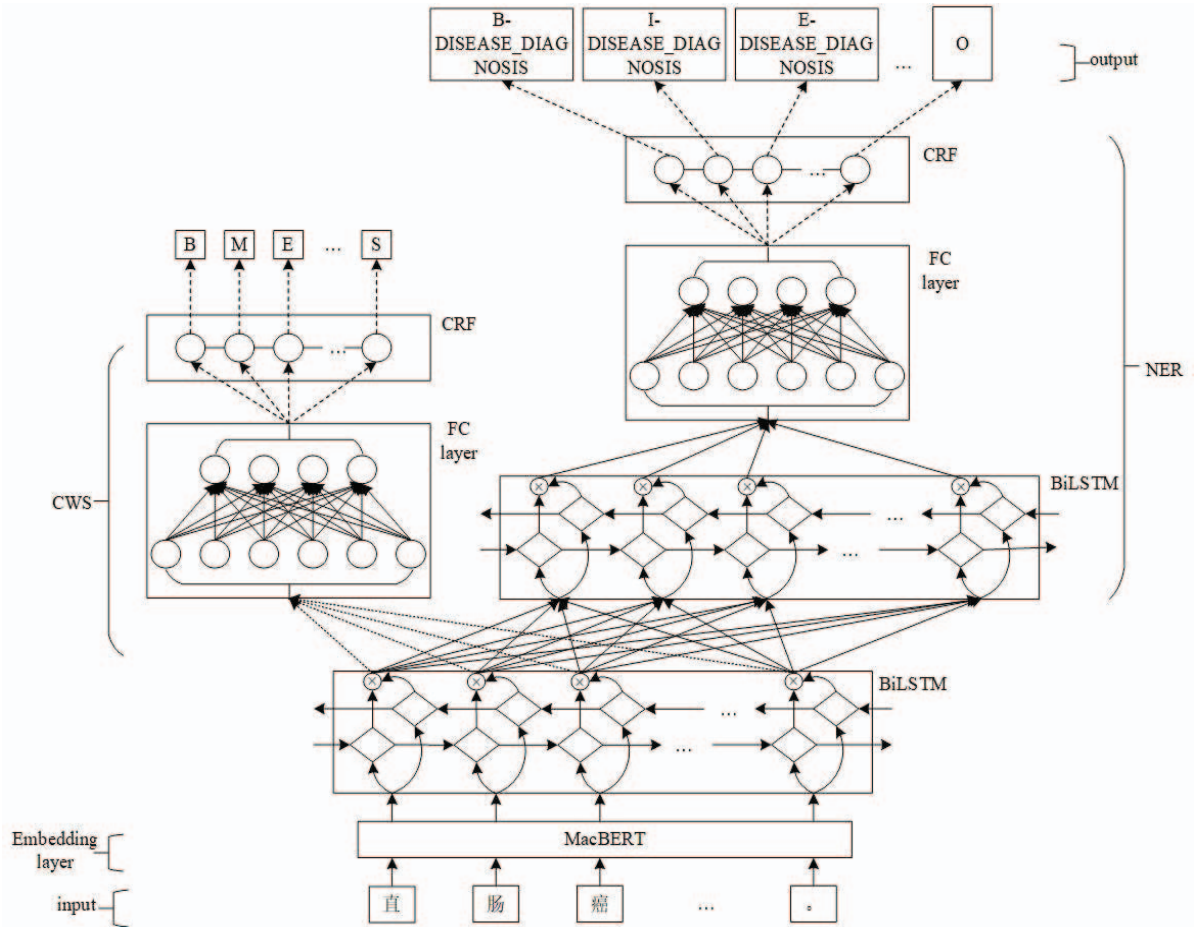

Figure 1 Frame Diagram

The model is mainly divided into three layers, sharing the MacBERT pre-training layer, and the CWS model including a one-layer BiLSTM-CRF and a two-layer BiLSTM-CRF Chinese named entity recognition model. The CWS model and NER model shares the MacBERT pre-training layer and the first BiLSTM network layer. In order to dynamically extract word vector features according to context information, MacBERT pre-training model is introduced into our model. NER part and CWS part are jointly trained, so that the model can obtain Chinese character features. At the same time, the word-level features of the text in the CWS task are obtained, so as to improve the ability to identify the entity boundary.

Define the input sentence sequence as $S=\{c_0,c_1,\cdots,c_n\}$, where c_i represents the i_{th} word in the sentence. After the input sentence is processed by the MacBERT model, the output word-level feature sequence is $\boldsymbol{e}=\{e_0,e_1,\cdots,e_n\}$, where n is the length of the input sequence. $e\in R^{n\times d_e}$ represents the set of character vectors corresponding to the input data, which d_e is the word embedding dimension.

2.1 MacBERT layer

BERT (Bidirectional Encoder Representations from Transformers) is an unsupervised deep bidirectional language representation model, which is often used to pretrain semantic representation in named entity recognition tasks [7]. BERT model in pretrained by MLM (Masked Language Model) and NSP (Next Sentence Prediction). The MLM task is designed to replace the characters randomly selected from the input sequence with [MASK] in a certain proportion, and then predict the characters replaced by [MASK] based on the context. The NSP task is designed to determine whether two sentences, A and B, in the original text are connected. In 2019, Cui et al. [8] proposed the technique of Whole Word Masking (WWM) to optimize the original masking in MLM tasks. In this setting, instead of randomly selecting word blocks for [MASK], [MASK] corresponds to all characters of the whole word block at one time.

ALBERT(A Lite BERT，ALBERT) [9] is a lite BERT model. The NSP task of BERT model is aimed at judging whether two sentences are adjacent sentences. The SOP (Sentence Order Prediction) task of ALBERT model is designed to determine the coherence between sentences.

MacBERT is an improved version of BERT. In order to adapt the model to Chinese text, combined with the characteristics of Chinese text, MacBERT model uses the improved BERT whole word mask (BERT-WWM) and N-gram strategy to select candidate characters for [MASK]. At the same time, N-gram strategy is used to determine the ratio between a character and the context character of a character, and then the context character of a character is replaced by [MASK]. In order to alleviate the problem that the [MASK] mark appearing in the pre-training stage

will not appear in the fine-tuning stage, MacBERT introduces MLM as correction (Mac), using similar words to replace [MASK]. When no similar words can be replaced, random words are used. In order to determine whether two sentences are contiguous and sequential in the original text, SOP task in ALBERT model was introduced to replace NSP task of BERT, which achieved good results in upstream and downstream tasks [9]. Different Masking strategies are shown in Table I.

Table I Different Masking strategies

Different Masking strategies	Sample
Original Sentence	使用语言模型来预测下一个词的概率。
Original Masking	使用语言[M]型来[M]测下一个词的**概**率。
+WWM	使用语言[M] [M]来[M] [M]下一个词的**概**率。
++N-gram Masking	使用[M] [M] [M] [M] [M] [M] [M]来[M] [M]下一个词的**概**率。
+++Mac Masking	使用语法模型来预见下一个词的几率。

2.2 BiLSTM-CRF named entity recognition part

The output vector of MacBERT enters the first layer BiLSTM network, which is shared with the CWS model. Output is as follows:

$$\overrightarrow{h^{(1)}} = \overrightarrow{LSTM}(e) \tag{1}$$

$$\overleftarrow{h^{(1)}} = \overleftarrow{LSTM}(e) \tag{2}$$

$$h_i^{(1)} = \left[\overrightarrow{h_i^{(1)}}, \overleftarrow{h_i^{(1)}}\right], i = 1,2,\cdots,n \tag{3}$$

$$h^{(1)} = \left[h_1^{(1)}, h_2^{(1)}, \cdots, h_n^{(1)}\right] \tag{4}$$

Where $\overrightarrow{h^{(1)}}$ and $\overleftarrow{h^{(1)}}$ are the outputs of the first layer BiLSTM forward layer and backward layer respectively; $h^{(1)} \in R^{n \times 2d_h}$ represents the synthesis of output in two directions, and d_h represents the dimension of LSTM hidden layer.

Similar to the first layer, the output of the second layer BiLSTM is:

$$\overrightarrow{h^{(2)}} = \overrightarrow{LSTM}\left(h^{(1)}\right) \tag{5}$$

$$\overleftarrow{h^{(2)}} = \overleftarrow{LSTM}\left(h^{(1)}\right) \tag{6}$$

$$h_i^{(2)} = \left[\overrightarrow{h_i^{(2)}}, \overleftarrow{h_i^{(2)}}\right], i = 1,2,\cdots,n \tag{7}$$

$$h^{(2)} = \left[h_1^{(2)}, h_2^{(2)}, \cdots, h_n^{(2)}\right] \tag{8}$$

The fully connected layer classifies the output information of the BiLSTM layer to obtain the label prediction probability of the neural network:

$$o_j^{NER} = \sigma\left(W^{NER} h_j + b^{NER}\right), j = 1,2,\cdots,n \tag{9}$$

$$\boldsymbol{o} = \left\{o_1^{NER}, o_2^{NER}, \cdots, o_n^{NER}\right\} \tag{10}$$

Where W^{NER} 、 b^{NER} denotes the parameter of the full connection layer, $o_i \in R^{tags}$ denotes the prediction of each label corresponding to the i_{th} character entered.

The BiLSTM layer learn contextual information, but does not take into account the dependencies between successive labels. So this paper uses CRF to predict the final label.

The input text sequence is represented as $x = \{x_1, x_2, \cdots, x_n\}$, and the prediction sequence label $y = (y_1, y_2, \cdots, y_n)$ is obtained. The fraction function of the label corresponding y to the text is obtained by adding the transfer matrix $\boldsymbol{A}$ and the emission matrix $\boldsymbol{P}$, and its formula is as follows:

$$S(X, y) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i} \tag{11}$$

Here $\boldsymbol{A}_{y_i, y_{i+1}}$ is the transfer fraction from label y_i to y_{i+1}; $\boldsymbol{P}_{i, y_i}$ denotes the emission matrix. The probability of the predicted sequence y generation is:

$$P(y, X) = \frac{\exp\left(\text{score}(X, y)\right)}{\sum_{y'} \exp\left(\text{score}\left(X, y'\right)\right)} \tag{12}$$

Where y' denotes the set of all possible state sequences; y denotes the real sequence.

When the decoding was carried out, the Veterbi algorithm was adopted to find the sequence with the largest score among all the state sequences, which is the final labeling sequence of the CRF layer, so as to obtain the global optimal labeling sequence.

In this paper, we use the Negative log-likelihood function as the loss function of our method, as follows:

$$L_{NER} = \sum_{s \in S} \log\left(P\left(\boldsymbol{y}_s^{NER} \mid \boldsymbol{o}_s; \theta^{NER}\right)\right) \tag{13}$$

Where S is the set of training sets; $\boldsymbol{y}_s^{NER}$ and $\boldsymbol{o}_s$ represents the actual label sequence of corresponding training data and the output of the neural network layer, respectively. θ^{NER} represents the parameter of the CRF layer of the named entity model.

2.3 BiLSTM-CRF joint training word segmentation part

The CWS task identifies the boundaries of words[10]. In this paper, CWS and NER are jointly trained to enhance the ability of named entity recognition to recognize entity boundaries.

Since CWS and NER are similar, most of them are modeled as sequence annotation tasks. CRF model also achieves good results in word segmentation tasks. When

the word segmentation model is trained jointly with the NERmodel as an auxiliary task, it shares the MacBERT pre-training layer and the first BiLSTM neural network layer, and the output formula of the layer BiLSTM is (4). In the word segmentation model, the output is also used as the input of the fully connected layer. In the word segmentation model, the label prediction layer CRF layer is also used to predict the label of the word segmentation, and the output formula of the CRF layer is (12). The output and loss functions of the fully connected layer of the joint training word segmentation model are as follows:

$$o_i^{CWS} = \sigma\left(W^{CWS} h_i + b^{CWS}\right), i = 1, 2, \cdots n \tag{14}$$

$$\boldsymbol{o} = \left\{o_1^{CWS}, o_2^{CWS}, \cdots o_n^{CWS}\right\} \tag{15}$$

$$L_{CWS} = \sum_{s \in S} \log\left(P\left(\boldsymbol{y}_s^{CWS} \mid \boldsymbol{o}_s; \theta^{CWS}\right)\right) \tag{16}$$

Where W^{CWS} and b^{CWS} are the parameters of the fully connected layer of the CWS model. y_s^{CWS} denotes the word segmentation label sequence corresponding to the training data. θ^{CWS} denotes the parameter of the CRF layer of the word segmentation model.

2.4 Loss function

In order to facilitate model training, we combined the loss functions of the two models in this paper, and the formula is as follows:

$$L = \lambda L_{NER} + (1 - \lambda) L_{CWS} \tag{17}$$

Where $\lambda \in (0,1]$ denotes the weight of NER, which is used to control the relative importance of the loss of the NER model in the total loss.

3 Experiments

3.1 Parameter setting

In this paper, MacBERT-base model released by Harbin Institute of Technology - iFLYTEK Joint Laboratory were used. MacBERT-base had 12layers, 768dimensions of hidden layer, 12-headmodel and 102M parameters in total. The learning rate of MacBERT model is 2e-5, and that of other modules is 2e-3. Dropout is set to 0.5. Adam is used as the optimizer in the experiment, and the batch size is 32.

3.2 Dataset

This paper adopts the training set provided by "' Wan Chuang Cup 'TCM Tian chi Big Data Competition -- TCM Instruction Manual Entity Recognition Challenge". Some of the original data in the data set had multiple spaces, strings, garbled characters and other noises. In order to use more reliable data sets for entity recognition tasks, the data were cleaned and preprocessed. This dataset includes 1,000 drug instructions and corresponding entity labeling, in which the data set defines 13 types of entities such as drug ingredients, diseases and symptoms.

3.3 Evaluation metrics

Precision(P) 、 Recall(R) 、 and $F1$ score are used as evaluation indexes of model recognition effect .These are the most used NER metrics.

$$P = \frac{A}{B} \tag{18}$$

$$R = \frac{A}{C} \tag{19}$$

$$F_1 = 2 * \frac{P * R}{P + R} \tag{20}$$

A denotes with precision extracted entities， B denotes numbers of known entities， and C denotes the cumulative total of extracted entities with precision in the dataset.

3.4 Analysis of experimental results

In order to verify the effect of this model, this paper will compare it with several mainstream NER models, which are as follows: BiLSTM-crf (here referred to as LC), BiLSTM-CRF *(using a two-layer BiLSTM model, here referred to as LLC), LC+Joint(joint model using one layer of BiLSTM), LLC+Joint(joint model using two layers of BiLSTM), and BERT+LLC+Joint. Table II shows the experimental results on different models:

Table II Results of Different Model

Models	P	R	F1
BiLSTM-CRF	65.64	66.59	66.11
BiLSTM-CRF*	67.77	65.91	66.83
LC+Joint	71.43	63.47	67.21
LLC+Joint	72.89	63.47	67.85
BERT+LLC+ Joint	64.82	70.12	67.36
Our	65.56	71.56	68.43

Compared with the non-joint learning models BiLSTM-CRF and BiLSTM-CRF*, the Joint learning models LC+Joint and LLC+Joint increased the model value by 1.1% and 1.02% respectively. The reason may be the introduction of the word segmentation task for joint training. Compared with the value of the one-layer BiLSTM structure, the value of the two-layer BiLSTM structure increased by 0.64%. The reason may be that the shared BiLSTM layer can make the model better capture the context information. Compared with the combined BERT joint learning model, the value of the combined MacBERT joint learning model is increased by 1.07%. MacBERT pre-training model's Mac pre-training task strategy can reduce the difference between pre-training stage and fine-tuning stage, which can improve the ability of entity recognition. Compared with the combination of MacBERT joint learning model and LLC+Joint model, the value is increased by 0.58%, indicating that the addition of MacBERT model can more dynamically capture the syntacticity and semantic level information of context, and has better semantic representation ability. The results

of runtime are shown in Figure 2.

entity	precision	recall	f1-score
DRUG	0.5962	0.4306	0.5000
DRUG_INGREDIENT	0.4731	0.7213	0.5714
DISEASE	0.5847	0.6544	0.6176
SYMPTOM	0.6619	0.7107	0.6855
SYNDROME	0.4936	0.4427	0.4668
DISEASE_GROUP	0.4848	0.5333	0.5079
FOOD	0.0000	0.0000	0.0000
FOOD_GROUP	0.5040	0.5780	0.5385
PERSON_GROUP	0.6800	0.6939	0.6869
DRUG_GROUP	0.0000	0.0000	0.0000
DRUG_DOSAGE	0.5584	0.7051	0.6232
DRUG_TASTE	0.9016	0.9283	0.9148
DRUG_EFFICACY	0.7477	0.9113	0.8214
ALL	0.6556	0.7156	0.6843

Figure 2 The Results of runtime

In this paper, parameters are introduced into the loss function to control the weight of NER model loss in the whole joint training model. We conduct experiments with different values to explore the influence of these parameters on the final experimental results. The experimental results corresponding to different values are shown in Figure 3.

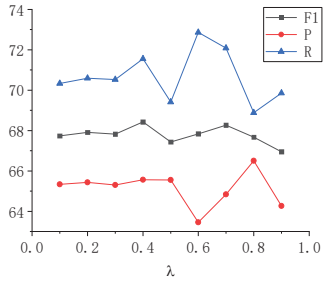

Figure 3 The Results of different λ

As can be seen from the picture above the value of λ has a certain influence on the experimental results. When it is too small, the segmentation model of auxiliary joint training is overemphasized, while the loss of NER model is not paid enough attention, so the experimental results are not good. When the value is 0.4, the experiment gets the best effect. When the value exceeds this value, the more it approaches 1, the information of the segmentation model auxiliary to joint training is not fully utilized, so the effect of the overall model will become worse.

4 Conclusions

Based on the characteristics of Chinese text and the inconsistency between the pre-training language model and the fine-tuning stage, MacBERT pre-training model is first introduced. MacBERT model improves the masked language model (MLM) task as a language correction method, which effectively alleviates the inconsistencies between the pre-training stage and the fine-tuning stage of the model. Then, the auxiliary CWS task and NER task are introduced for joint training, and the intrinsic relationship between the two tasks is used to improve the ability of Chinese NER model to recognize entity boundaries. The comparison with non-joint learning model also proves the advantages of the joint learning model. The experimental results show that the model designed in this paper can effectively identify the named entity of traditional Chinese medicine, which has a certain improvement compared with other models. Since the dataset is domain-specific, the next step is to consider leveraging other features of the text.

Acknowledgements

This work is supported by the National Science Foundation of China 62266025.

References

[1] XU Guo-hai. Research on Named Entity Recognition for Chinese Medical Texts[D]. East China Normal University, 2019. rained to enhance the ability of named entity recognition to recognize entity boundaries.

[2] Devlin, J., Chang, M. W., Lee, K., et al. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Jun 2-7, 2019. Stroudsburg: ACL, 2019: 4171-4186.

[3] Peters, M., Neumann, M., Iyyer, M., et al. Deep Contextualized Word Representation[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, Jun 1-6, 2018. Stroudsburg: ACL, 2018: 227-2237.

[4] LI Dong-mei, LUO Si-si, ZHANG Xiao-ping, et al. A Review on Named Entity Recognition [J]. Journal of Frontiers of Computer Science and Technology, 2022, 16(9): 1954-1968.

[5] HUANG Xiao-hui, QIAO Li-sheng, YU Wen-tao, et al. Joint learning of Chinese word segmentation and named entity recognition[J]. Journal of National University of Defense Technology, 2021, 43(1): 86-94.

[6] Cui, Y., Che, W., Liu, T., et al. Revisiting Pre-trained Models for Chinese Natural Language Processing[J]. arXiv preprint arXiv:2004.13922, 2020.

[7] ZHANG Fang-cong, QIN Qiu-li, JIANG Yong, et al. Named Entity Recognition for Chinese EMR with RoBERTa-WWM- BiLSTM-CRF[J]. Data Analysis and for Knowledge Discovery, 2022, 6(Z1):251-262.

[8] Cui, Y., Che, W., Liu, T., et al. Pre-Training with Whole Word Masking for Chinese BERT[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504-3514.

[9] Lan, Z., Chen, M., Goodman, S., et al. AlBERT: A Lite BERT for Self-supervised Learning of Language Representations[J]. arXiv preprint arXiv: 1909.11942, 2019.

[10] Wu, F., Liu, J., Wu, C., et al. Neural Chinese Named Entity Recognition via CNN-LSTM-CRF and Joint Training with Word Segmentation[C]//The World Wide Web Conference. 2019: 3342-3348.