Named entity recognition method of Chinese EMR based on BERT-BiLSTM-CRF

Wenchao Gao1*, Xiaohui Zheng2, Shanshan Zhao3

1,2,3 China University of Mining and Technology-Beijing, University, Beijing, Beijing, 100089, China

*Corresponding author’s e-mail: gaowc@cumtb.edu.cn

Abstract. With the widespread use of Chinese electronic medical records, the extraction of medical named entities and the resolution of the polysemous term of entities are of great significance to the analysis and processing of patient information and disease diagnosis. This paper uses a BERT Chinese pre-training vector that does not rely on manual feature selection, combines BiLSTM and CRF Chinese named entity recognition algorithm model, and applies it to the processing of the CCKS2020 electronic medical record data set. This paper conducts experimental tests and comparisons between the BERT-BiLSTM-CRF model and other models. The results show that the results of this model are better than other models, and it can accurately identify multiple entity categories and has a good application prospect.

1. Introduction

With the wide application of big data in the medical field, electronic medical records (EMR, Electronic Medical Records) are gradually replacing traditional paper medical records with their high efficiency, ease of statistics and processing. EMR is generally stored in a semi-structured or unstructured form. Extracting information from it can obtain specific information and form structured data to assist people in sorting and analyzing this information [1]. However, its internal medical vocabulary is highly professional and there are a large number of phenomena of polysemy or entity co-referencing which poses a challenge to the information extraction effect of EMR.

NER (Named Entity Recognition) refers to identification of the specific entities in the text, such as the names of persons, places, organization, which is the main content information extraction. The named entity recognition of EMR which is in the form of natural language as the main existing form can mine a large amount of medical knowledge to serve medical decision-making and other scenarios and make full use of medical record information. Therefore, improving the recognition accuracy and precision of medical named entities in EMR is an urgent problem to be solved.

However, there are still some problems in the research of EMR named entity recognition, such as: the recognition accuracy of professional medical vocabulary is not high enough; the number of new medical entities is increasing rapidly, so a large number of unregistered words need to be recognized; existing named entity recognition methods cannot meet the requirements of cross-domain and multiple types of entity recognition. In response to the above problems, this paper uses the pre-trained model BERT with stronger text feature representation capabilities as the feature representation layer, combined with BiLSTM and CRF model, extracts global and local features of the text. In this paper, the CCKS 2020 Chinese EMR dataset is tested and compared with another model to verify the practical advantages of this model.
2. Related Research

Named entity recognition is generally regarded as a sequence labeling task in machine learning, which can be divided into rule-based and dictionary-based methods, statistical machine learning methods, and deep learning methods.

Based on rules and dictionaries, experts construct rule templates or dictionaries, and use matching methods to process text. The limitations of this method is that it consumes a lot of manpower, and is not conducive to expanding to data sets in other fields, and it cannot adapt to data updates.

Machine learning method include: HMM (Hidden Markov Model \([2]\)), SVM (Support the Vector Machine \([3]\)), CRF (Conditional Random Field \([4]\)) and so on. Among them, CRF considers the global distribution of data when normalizing, solves the problem of label offset, and makes full use of internal and contextual feature information.

The deep learning method uses a large amount of data to automatically learn features. During the learning process, the parameters of the training model are optimized to improve learning efficiency and result accuracy, avoiding subjectivity and contingency when manually selecting features. One of the more commonly used is the LSTM (Long-Short Term Memory) proposed by Hochreiter et al. \([5]\), which uses the "gate" mechanism and the concept of cells to solve the problem of long-term dependence of RNN. In 2018, some excellent models appeared in the field of natural language processing, such as BERT \([6]\), ELMo \([7]\), OpenAI-GPT. Among them, the BERT (Bidirectional Encoder Representations from Transformers) pre-trained language model achieved the best results in some subtasks of NLP Performance, updating the performance 11 downstream tasks of natural language processing in 2018.

2.1. BERT pre-trained language model

In text classification, the text needs to be represented by a vector first, that is, the Word Embedding. Mikolov et al. \([8]\) first proposed the Word2vec model for word vector training, which is widely used in multiple tasks of natural language processing. But Word2vec can only use a unique word vector to represent the multiple semantics of a word, and cannot express the polysemous situation of a word. The BERT model proposed later uses transformer structure in two directions to encode a word, which can represent the specific semantics of the word in the context according to the semantic relationship of the context.

The input of the BERT model is composed of the addition of three parts of embedding vectors: Token embeddings, Segment embeddings and Position embeddings. Token Embeddings represents word vectors and also uses several special signs: [CLS] sign indicates the beginning of a sentence, [SEP] sign indicates the separation between two sentences, and [UNK] sign is used to indicate unknown characters. In the classification task with multiple sentences as input, Segment Embeddings is used to distinguish different parts or types of sentences. Position Embeddings represents character position information. The input representation structure of BERT is shown in figure 1.

The BERT model pre-trains a large amount of corpus, which can make words better learn the feature representation. The BERT model has two pre-training tasks: MLM (Masked Language Model) and NSP (Next Sentence Prediction), which can capture the word-level and sentence-level vector representations, and then optimize the model through joint training. MLM refers to using some words to randomly replace some words in a sentence and predicting the original words that are covered based on context information. NSP will randomly replace some sentences and predict whether the next sentence is
continuous with the previous sentence, so as to learn the continuous relationship between adjacent sentences.

The working mechanism of BERT is: first obtain the text sequence to be trained through word segmentation, randomly mask some words of the sequence, and add special tags such as [CLS] and [SEP] to the sequence. Input the obtained three-part Embeddings vector into the two-way transformer to extract the sequence features, and obtain the sequence vector with rich semantic features. The BERT model structure is shown in figure 2.

Google provides two basic models of BERT: BERT-Base and BERT-Large. The feedforward size of the two versions is set to 4 layers, and the detailed parameters are shown in table 1. Among them, Layer, hidden, and heads respectively represent the number of layers (i.e., two-way transformers encoder module), the hidden size of hidden layer neurons covered, and the number of self-attention heads.

The BERT-Base Chinese pre-training model proposed by Google uses the BERT-Base version of the framework, and due to too many BERT-Large parameters, the pre-training cost is too high, so most experiments use the BERT-Base Chinese version of the model structure.

2.2. model of BiLSTM
In sequence labelling tasks, the RNN model can dynamically capture sequence data information and store the information in memory, but it is easy to cause problems such as gradient disappearance.

Compared with the RNN model, LSTM adds a memory unit to the hidden layer, which solves the problems of gradient disappearance or gradient dispersion caused by long sequence information. In addition, it also adds a threshold mechanism to selectively store and discard the required Therefore, it is
widely used in named entity recognition tasks. The network structure of LSTM can be expressed as follows:

\[ i_t = \sigma(x_t w_{ih}^i + h_{t-1} w_{hh}^i + b_i) \]  
\[ f_t = \sigma(x_t w_{ih}^f + h_{t-1} w_{hh}^f + b_f) \]  
\[ o_t = \sigma(x_t w_{ih}^o + h_{t-1} w_{hh}^o + b_o) \]  
\[ c_t = \tanh(x_t w_{ih}^c + h_{t-1} w_{hh}^c + b_c) \]  
\[ c_t = i_t \odot c_t + f_t \odot c_{t-1} \]  
\[ h_t = o_t \odot \tanh(c_t) \]  

The traditional one-way LSTM model cannot process contextual information at the same time, so Graves A et al. [9] used the basic memory unit of LSTM to construct the BiLSTM model, using a forward and a backward LSTM module, respectively, connected to the same output layer. Two different hidden layer representations are obtained by calculating in order (from left to right according to the sentence direction) and reverse order (from right to left according to the reverse of the sentence) for each sentence, and the final hidden layer representation is obtained through vector splicing. BiLSTM can better capture semantic dependencies, learn more comprehensive contextual and semantic co-occurrence information than LSTM, therefore effectively use the context information of text sequences [10].

2.3. model of CRF

Although the BiLSTM model can identify entity boundaries, it does not consider whether the relationship between the entity sequences is correct, while the CRF model can obtain the global optimal tag sequence by considering the dependency relationship between adjacent tags, so it is often applied in tasks such as part-of-speech tagging and named entity recognition.

CRF is a sequence labelling algorithm proposed on the basis of the EM model and the HMM model. It can solve the label bias problem by considering the global information of the label sequence, and better predict the label.

The basic principle of CRF is to calculate the conditional probability distribution of the output random variable with a given random variable as input, usually using the Viterbi algorithm for decoding. The CRF model used in named entity recognition is to use the word sequence in the input sentence as the observation sequence, and the labelling process is to infer the most likely label sequence based on the known word sequence.

Therefore, by combining CRF with BiLSTM neural network and reprocessing the output of BiLSTM, the output result of the BiLSTM is processed and revised again to obtain the best entity annotation.

3. Model design and Implementation

The BERT-BiLSTM-CRF model structure of this article is divided into three parts: BERT pre-training language model, BiLSTM word vector processing module and CRF decoding module. First, use the BERT model to obtain the word vector corresponding to each input character in the labelled corpus, and then input the output word vector sequence into the BiLSTM layer for semantic encoding to obtain the global sequence feature, and finally decode the output result of the BiLSTM layer through the CRF layer, output the corresponding predicted label sequence, then extract and classify the entities of the label sequence to obtain the final recognition result.

For example, the model first embeds the input text content on the BERT layer, obtains the vector representation of each word, and generates an input sequence: \{T_1, T_2, ..., T_N\}, and then the coding sequence of the input vector is returned through the bidirectional LSTM structure of the BiLSTM layer,
and the preliminary classification and recognition results are obtained through the softmax layer. Finally, the CRF layer corrects the preliminary recognition results to obtain the best classification label. The BERT-BiLSTM-CRF model structure is shown in figure 3.

The advantage of this model is that the BERT language preprocessing model does not need to train the word vectors in advance. It only needs to directly input the sequence, then the model will automatically extract multiple features such as rich word-level features, grammatical structure features and semantic features in the sequence.

Studies have pointed out that the characteristics learned by each layer of the BERT model are different. The bottom layer of the BERT model is mainly to obtain feature information at the phrase level, the middle layer is mainly to learn feature information such as syntactic structure, and the top layer is to capture the semantic information of the entire sentence, therefore, BERT can obtain word vectors with rich semantics, and its internal self-attention mechanism makes it effective for processing long-distance information-dependent sentences. The traditional word vector model mainly focuses on the acquisition of word or character-level features, and rarely involves syntactic structure and other semantic information. It can be seen that the feature extraction ability of the BERT model is stronger than that of the traditional model.

In addition, this model does not use traditional methods such as dictionaries and templates, and does not rely on artificially selected semantic features, which can reduce manpower consumption and help strengthen the understanding of text information, thereby improving the generalization ability of the model.

4. Experimental process and result analysis

4.1. Experimental data and evaluation

This paper mainly uses the publicly released data set of the EMR named entity recognition evaluation task in the CCKS2020 competition. The training data includes a total of 1500 labeled data, 1000
unlabeled data, and 6292 entity vocabularies in 6 categories. The training data set statistics are shown in table 2:

**Table 2. Entity statistics of CCKS 2020 labeled dataset**

| category          | Disease and Diagnosis | Anatomy | Image Inspection | Laboratory Inspection | Drug | Surgery |
|-------------------|-----------------------|---------|-------------------|-----------------------|------|---------|
| number            | 6211                  | 12660   | 1490              | 1885                  | 2841 | 1327    |

Each medical record text in the training data is divided into two parts: original text and entities entity description part. The original text and its corresponding entity description are stored in a key-value pair dictionary format. The original text is the original text of the Chinese electronic medical record provided by the hospital. The entities entity description part is an explanation of the starting position, ending position and entity type of the entity words appearing in the original text.

In order to test the performance of the model more intuitively, this paper adopts the precision rate P, the recall rate R and the F1 value as the evaluation criteria. The evaluation criteria are defined as:

\[
P = \frac{TP}{TP + FP} \quad (7)
\]

\[
R = \frac{TP}{TP + FN} \quad (8)
\]

\[
F1 = \frac{2 \times P \times R}{P + R} \quad (9)
\]

Among them, P represents the ratio of the number of correctly predicted named entities to the number of all identified named entities, R represents the ratio of the number of correctly predicted named entities to the number of named entities in the standard result, and the F1 value is the harmonic average of P and R.

### 4.2. Data preprocessing

First, the original medical record data is pre-segmented, and the entity is corresponding to its category according to its entity description part, and then divided into training set and validation set according to the ratio of 8:2, which are used for model training and verification respectively.

In the data tagging stage, this paper uses the BIO tagging system, that is, the named entities of each category are labeled as B-X, I-X and O, where B-X indicates that the entity word belongs to category X and the labeled element is located at the beginning of the entity word. I-X indicates that the entity word belongs to category X and the labeled element is located in the middle of the entity word to which it belongs, and O indicates that it does not belong to any predefined category.

### 4.3. Model building and parameter setting

The model in this article uses the Keras framework for experimentation, and specific experimental environment is shown in table 3. In the training process, the Adam optimizer is used, the learning rate is set to \(2 \times 10^{-5}\), the maximum input dimension is set to 500, and the batch_size is 32. In addition, dropout is used in the BiLSTM layer to suppress the overfitting problem, and the value is 0.2.

**Table 3. Experimental environment**

| Experiment Environment | Parameter |
|------------------------|-----------|
| CPU                    | Intel® Core™ i5-4210U CPU @ 1.70GHz 2.39GHz |
| GPU                    | Tesla K80 |
| Python                 | 3.7       |
| Tensorflow             | 1.13.1    |
Different from other models, the feature extraction strategy used in the training of this model is to fix the BERT parameters and only update the BiLSTM-CRF parameters. Although this training idea will inevitably lose a certain degree of accuracy, it can greatly reduce the total amount of actual training parameters, effectively shorten training time, and improve training efficiency.

4.4. Analysis of experimental results
The performance of the Word2vec-BiLSTM-CRF model and the BERT-BiLSTM-CRF model are compared and analyzed on the official CCKS data set. The precision rate, recall rate and F1-score results of the two models are shown in Table 4 and Table 5, respectively.

| Category                | Word2vec-BiLSTM-CRF | Precision | Recall | F1 value |
|-------------------------|---------------------|-----------|--------|----------|
| Disease and Diagnosis   | 61.84%              | 70.45%    | 65.87% |
| Anatomy                 | 77.30%              | 69.95%    | 73.44% |
| Image Inspection        | 68.75%              | 78.57%    | 73.33% |
| Laboratory Inspection   | 70.73%              | 65.90%    | 68.24% |
| Drug                    | 94.43%              | 91.67%    | 93.03% |
| Surgery                 | 86.64%              | 81.39%    | 83.93% |

| Category                | BERT-BiLSTM-CRF     | Precision | Recall | F1 value |
|-------------------------|---------------------|-----------|--------|----------|
| Disease and Diagnosis   | 85.71%              | 90.55%    | 88.06% |
| Anatomy                 | 92.55%              | 95.57%    | 94.04% |
| Image Inspection        | 89.03%              | 98.03%    | 93.31% |
| Laboratory Inspection   | 86.15%              | 96.42%    | 90.99% |
| Drug                    | 94.36%              | 99.42%    | 96.82% |
| Surgery                 | 85.26%              | 93.47%    | 86.94% |

It can be seen from the table that the precision rate, recall rate and F1 value of the BERT-BiLSTM-CRF model are higher than Word2vec-BiLSTM-CRF model, but the precision rate in the two categories of drugs and surgery is slightly lower and no better recognition effect is obtained. This shows that the BERT pre-training language model dynamically generates the contextual semantic representation of characters through the two-way transformer structure, which can better capture the contextual semantic information of characters than traditional word embedding vectors, and has a better effect in the recognition of Chinese electronic medical records, but there are also some shortcomings.

From a specific point of view of the data, the precision, recall and F1 value of this model on the first four types of entities are increased by 15.25% to 23.87%, 19.46% to 30.52%, and 19.98% to 22.75% compared with the Word2vec-BiLSTM-CRF model, but the recognition precision rates on two types of labels for drugs and surgery dropped by 0.07% and 1.38%, respectively.

There are two main reasons for this result: the first reason is that there are a large number of complex and nested vocabulary for drug and surgical entities. Such entities are highly specialized and difficult to identify. Generally, there are two or more entity words are spliced or nested and an entity contains other entity categories. It is difficult for the model to distinguish and identify the exact entity boundary, so the recognition accuracy cannot be guaranteed. In addition, the proportion of data in each category is not balanced. For example, the anatomical parts have the most labeled entities, as many as 12,660, while there are only 1,327 labeled entities for surgery. Unbalanced data can also easily lead to over-fitting or
under-fitting in the training process. In particular, the domains of drugs and surgery and the dependence on professional knowledge are one of reasons for the low scores of drugs and surgery.

Another reason is that Chinese electronic medical records are basically derived from manual input by doctors, and there is inevitably a small amount of human error. And with the lack of professional dictionaries in the medical field for correction, which can easily lead to problems such as recognition misalignment or blurred category boundaries. This model does not use professional dictionaries as support to assist the model to perform named entity recognition tasks. Although it avoids manual costs, it also ignores the limitations of deep learning models in training corpus.

5. Concluding remarks
In order to facilitate the processing and use of medical electronic medical record information, this article applies BERT-BiLSTM-CRF deep learning model to the task of named named entities in medical electronic medical records, which is used to identify and extract professional medical terms, such as drugs, anatomical parts.

The main contribution of this paper is mainly reflected in the use of a BERT Chinese pre-training vector that does not rely on artificial feature selection, fusion of the well-performing BiLSTM and CRF models, and application of it to the processing of the CCKS2020 electronic medical record data set. The overall performance of this model is better than the Word2vec-BiLSTM-CRF model. It has good application prospects in natural language processing tasks in the medical field, for example, it can be applied to the information processing of patients' electronic medical records and disease diagnosis.

In the next step, the model can be optimized in combination with domain dictionary and other professional knowledge, and the method of feature engineering can be combined to improve the accuracy of the recognition results; this method can also be extended to text processing in other professional fields to test the applicability of this model in multi-domain named entity recognition; in view of the low accuracy of nested named entity recognition, the next step can be to use graph neural network model to analyze the structural relationship between multiple entities; in addition, because the BERT model is too large, the training will be slow and the training cost will be high. So we can consider reducing some parameter calculations to optimize the model structure and improve the training speed of the BERT model.

Acknowledgments
This work had supported by the Special Funds for Fundamental Research Funds for Central Universities (2020YQJD15) and National University Student Innovation Training Project Funding (C202004828).

References
[1] Wang, T., Xuan, P., Liu, Z., & Zhang, T. (2020). Assistant diagnosis with chinese electronic medical records based on cnn and bilstm with phrase-level and word-level attentions. BMC Bioinformatics, 21(1).
[2] Rabiner, & L., R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2), 257-286.
[3] Roberts A, Gaizauskas R, Hepple M. Extracting Clinical Relationships from Patient Narratives [C]. Stroudsburg: Proceedings of the 2008 Workshop on Current Trends in Biomedical Natural Language Processing, 2008:10-18.
[4] Lafferty, J., McCallum, A., & Pereira, F. C. N.. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proc. 18th International Conf. on Machine Learning.
[5] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780.
[6] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/pdf/1810.04805.pdf
[7] Peters M, Neumann M, Iyyer M, et al. Deep contextualized word representations. In Proceedings
of NAACL, 2018: 2227-2237

[8] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. Computer Science.

[9] Graves, A., & Jürgen Schmidhuber. (2005). Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural Networks, 18(5–6), 602-610.

[10] Chiu, J. P. C., & Nichols, E. (2015). Named entity recognition with bidirectional lstm-cnns. Computer Science.