TCM Named Entity Recognition Based On Character Vector With Bidirectional LSTM-CRF

Jigen LUO, Jianqiang DU*, Bin NIE, Wangping XIONG, Jia HE, Yanyun YANG
School of Computer
Jiangxi University of Traditional Chinese Medicine
Nanchang, China
E-mail: jianqiang_du@163.com

***Abstract*—In order to better solve the problem of low accuracy caused by inaccurate word segmentation in the task of extracting Chinese medicine named entities, the extraction technology relies heavily on the characteristics of manual development, and needs guidance of domain knowledge. This paper proposes a named entity recognition in the field of Traditional Chinese Medicine (TCM) Based on character vectors for Bidirectional Long Short Term Memory with a Conditional Random Field (Bidirectional LSTM-CRF). First of all, the model uses the word2vec tool to convert the corpus into a character vector, which can avoid the influence of inaccurate word segmentation in Chinese medicine field on entity recognition; then use Bidirectional LSTM neural network to extract deep features of sentence level, and reduce the workload of manual feature setting in the traditional method; Finally access to the CRF layer, and the Viterbi algorithm is used to dynamically plan the most reasonable tag output of the sentence, and the correlation between the output tags is considered. We use different models to conduct experiments on the TCM corpus. The results show that the model proposed in this paper has a good effect. The F value of the evaluation index on the three types of Chinese medicine, prescription and syndrome type has reached 90%.**

Keywords-Named Entity Recognition; Character vector; Bidirectional LSTM-CRF; Chinese Medicine Informatics.

I. INTRODUCTION

The knowledge of TCM diagnosis is a historical treasure left by the Chinese nation for thousands of years and has a strong guiding role in the clinical treatment. In the field of traditional Chinese medicine, the forms of expression between syndromes, prescriptions and traditional Chinese medicines are diverse and complex, and they are used throughout the treatment process of Chinese medicine [1][2]. In order to construct a knowledge graph of the Chinese medicine field and form a structured knowledge, the extraction of three types of entities such as syndrome type, prescription and traditional Chinese medicine is particularly important and a very meaningful step.

Named entity recognition is a type of problem of sequence labeling. It is the basic work of information extraction, information retrieval, machine translation and other tasks [3]. The current mainstream methods for named entity recognition are based on statistical and rule-based language model methods [4]. The statistical based method automatically extracts the composition law of the named entity directly from the text, and identifies the entity through the model-trained language model [5]. The statistical method for the named entity recognition has a hidden Markov model [6] and the conditional random field Model [7]. The rule-based approach [8] uses the rational knowledge of linguists to identify named entities through rules written by linguists. Traditional named entity recognition methods require a wealth of domain expert knowledge to extract a large number of artificial features [8].

With the rapid development of deep learning, the advantages of neural networks in dealing with natural language processing problems have gradually emerged. The core technology of deep learning is to express text features in the form of word vectors. Zheng et al. [9] proposed using neural networks for named entity recognition tasks, and using perceptron algorithm for accelerated training models; later, Recurrent Neural Network applications appeared in sequence labeling problems. Chowdhury et al. [10] The Recurrent Neural Network is used for sequence tasks, achieving better results and avoiding feature engineering. However, it is difficult for recurrent neural networks to obtain long distance dependent information, and long distance information is lost. In order to solve the problem of long distance information dependence. Xu et al.[11] used a Bidirectional LSTM neural network and a conditional random field method to identify medical names. This method takes advantage of LSTM's feature extraction and considers the relationship between output tags in combination with CRF, making Long Short-Term Memory with a Conditional Random Field (LSTM-CRF) model is widely used in Chinese named entity recognition tasks [12][13]. However, due to the lack of obvious space between Chinese characters, word segmentation processing is required, but the current word segmentation software is aimed at the general field and rarely involves special fields. In order to solve the erroneous influence of inaccurate word segmentation on entity recognition, Dong et al. [14] used character-level vectors as input for deep learning to perform entity recognition, but these are all considered in the general field, without considering the particularity of the Chinese medicine field.

There are deficiencies in Chinese word segmentation in the field of Chinese medicine. Many Chinese medical terms are misclassified. For example, “Yin and Yang deficiency syndrome” is a word, and the commonly used word

 ISBN: 978-1-61208-688-0

segmentation tool will divide it into: “Yin and Yang”, “Two” and "Deficiency syndrome", this wrong participle will seriously affect the entity recognition effect.

This paper provides a Bidirectional LSTM-CRF named entity recognition method based on character vector, which lays a foundation for the knowledge graph construction in the field of traditional Chinese medicine. By continuously adjusting the parameters of the model, until the neural network parameter combination suitable for entity recognition in the Chinese medicine field is found, the effect of the entity recognition is optimized.

Section II introduces the model proposed in this paper; the model is based on character vector with Bidirectional LSTM entity recognition method in the field of traditional Chinese medicine. In Section III, the contrast experiment and adjustment model parameters are given. It is proved that the proposed model in this paper has certain advantages. The parameters of LSTM neural network model suitable for traditional Chinese medicine are found. Finally, the conclusion and future work of this research are found in Section IV.

II. BIDIRECTIONAL LSTM-CRF MODEL BASED ON CHARACTER VECTOR

For the identification of entities in the field of traditional Chinese medicine, this paper identifies the three types of entities: syndrome type, prescription and traditional Chinese medicine, which are marked by two words. B-XX indicates the first type of entity, while I-XX indicates other words of the entity. The specific label forms of the three types of entities are shown in TABLE I.

As shown in Figure 1, the schematic diagram of the character vector based Bidirectional LSTM-CRF TCM domain named entity recognition model designed for this paper. The model is divided into three parts: the character vector layer, the Bidirectional LSTM neural network layer, and the CRF tag inference layer. The character vector layer converts the input text into a recognizable numerical form of the neural network, and then calculates the output feature vector h_t of each character through the bidirectional LSTM neural network. The CRF inference layer finds the most suitable output tag sequence of the sentence through dynamic programming to make up Subsequent to the defect that the Softmax output tags are independent of each other.

TABLE I. ENTITY TWO WORD POSITION LABEL REPRESENTATION

Sign	Significance
B-SYN	The first Chinese character of the syndrome type entity
I-SYN	All Chinese characters remaining in the first characters of the syndrome type entity
B-PRE	The first Chinese character of the prescription type entity
I-PRE	All Chinese characters remaining in the first characters of the prescription type entity
B-MED	The first Chinese character of the Chinese medicine type entity
I-MED	All Chinese characters remaining in the first characters of the Chinese medicine type entity
O	Irrelevant words

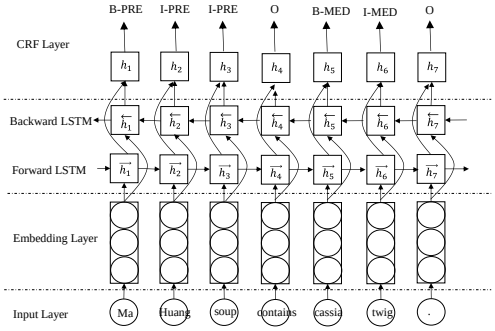

Figure 1. The model structure of BLSTM-CRF

A. Characters Vector Layer

The structure of the Bidirectional LSTM-CRF model is shown in Figure 1. The bottom end of the input is the sequence to be labeled "Ma Huang soup contains cassia twig", which is the input layer in the figure. Since the neural network algorithm cannot directly process Chinese characters, it is necessary to convert the text into a vector form. There are two ways to vector the text: one-hot and distributed representation. Since the distributed vector representation can reduce the dimension of the vector and can effectively represent the association between semantics. This paper adopts the distributed representation method for text vector. The character embedding operation is required before entering the model. This article use Google's opening source tool word2vec to convert the text into a character vector. Suppose the input sentence is S , and the set of characters contained is $W(w_1, w_2, w_3, ..., w_m)$, m is the length of the sentence, where the $t-th$ character vector is $w_t^* \in R^d$, where the dimension of the word vector is in the above formula, the input text is expressed as:

$$S = \left[w_1^*, w_2^*, ..., w_m^* \right] \in R^{T \times d} \quad (1)$$

B. Bidirectional LSTM Neural Network Layer

The LSTM neural network is an improvement to the common RNN neural network. Its main purpose is to solve the problem of gradient disappearance or gradient explosion. The biggest difference from the RNN neural network is the neural unit. The neural unit of the LSTM neural network joins the gate structure, including the input gate, the output gate and the forgetting gate. The neural unit of the LSTM neural network is shown in Figure 2.

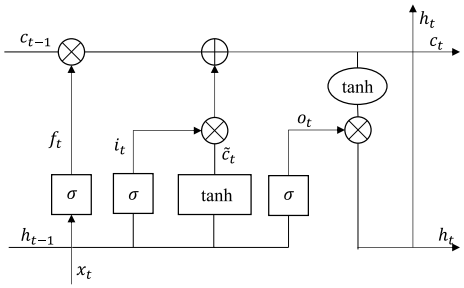

Figure 2. The neural unit of the LSTM neural network

At t time, the LSTM unit components are updated as follows.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (2)$$

$$i_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_i) \quad (3)$$

$$\tilde{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (4)$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \tilde{c}_t \quad (5)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (6)$$

$$h_t = o_t \circ \tanh(c_t) \quad (7)$$

Among them, represents the sigmoid activation function, which is the element multiplication， x_t is the input vector of LSTM at t time, h_t represents the hidden state; W_f ,W_i , W_c ,W_o represents the weight matrix of the forgotten gate, input gate, memory cell, and output gate; b_f , b_i , b_c , b_o represents bias of the forgotten gates, input gates, memory cells, output gate. f_t , i_t , c_t , o_t represents forgotten gate, input gate, memory cell status, and output gate.

The output feature vector calculation method of unidirectional LSTM neural network is introduced above. In order to make full use of context information and mine more hidden features, and effectively solve the problem of new word discovery in TCM named entity recognition task, this paper adds on top of LSTM neural network. A layer of reverse LSTM neural network structure forms a bidirectional LSTM neural network. The assumption $\overrightarrow{h}_t$ is that the output of the forward LSTM neural network at the moment, $\overleftarrow{h}_t$ is the output of the backward LSTM unit at the moment, and the output of the time t is the splicing of the preceding and succeeding vectors, that is $h_t = [\overrightarrow{h}_t, \overleftarrow{h}_t]$.

C. CRF Layer

The CRF algorithm is a machine learning model specifically designed for sequence labeling tasks. Suppose the input sequence of the model is $W = \{w_1, w_2, w_3, ..., w_m\}$, which w_i represents the $i-th$ Character in input sentence, output sequence is $Y = \{y_1, y_2, y_3, ..., y_m\}$, and y_i is the output sequence label of w_i . In this paper, P is the scoring matrix of the bidirectional LSTM layer, and P is a $m \cdot k$ matrix, where m is the length of sentences and k is the number of custom tags. P_{ij} Represents the $j-th$ label score value of the $i-th$ Character in the sentence. For example, P_{12} refers to the probability that the first Character in a sentence is labeled as the second label in the label set. In order to better calculate the sentence path, you need to add the start tag at the beginning of the sentence and the end tag at the end of the sentence. Therefore, the parameter of the CRF layer is a state transition matrix of $(k+2)*(k+2)$, then the whole model is marked as $Y = \{y_1, y_2, y_3, ..., y_m\}$ for the input sequence $W = \{w_1, w_2, w_3, ..., w_m\}$ .The final score is calculated as follows.

$$score(W, Y) = \sum_{i=0}^{m} A_{y_i, y_{i+1}} + \sum_{i=1}^{m} P_{i, y_i} \quad (8)$$

The final score of the sentence can be divided into two parts, where P_{i,y_i} is the scoring matrix of the bidirectional LSTM neural network model and A is the transfer scoring matrix between adjacent tags. In order to get the sentence probability, the score matrix s needs to be normalized. The probability formula is as follows:

$$p(Y \mid W) = \frac{\exp(score(W, Y))}{\sum_{y} \exp(score(W, y))} \quad (9)$$

In order to get the final prediction label, it can be obtained by the Viterbi decoding algorithm. The dynamic programming can find the best path. The calculation formula is as follows.

$$Y^* = \underset{Y}{argmax}\, score(W, Y) \quad (10)$$

III. Experiment And Analysis

In this section, we mainly verify the validity of the application of the Bidirectional LSTM neural network based on character vector on the entity extraction task in the field of traditional Chinese medicine.

A. Experiment Data And Evaluation Indicators

The corpus used in the experiment is the Chinese medicine diagnosis text and classic Chinese medicine books provided by Jiangxi University of Traditional Chinese Medicine, such as the TCM syndrome differential diagnosis book, a total of 32,700 sentences related to syndromes and prescriptions, and Chinese medicine, totaling more than 2 million character. After many years of clinical experience, the corpus has been comprehensively covered with information on syndromes, prescriptions, Chinese medicine and other related entities.

In order to display the experimental results comprehensively and intuitively, this paper adopts the Precision rate (P), Recall rate (R) and F value as model evaluation indicators.

B. Experimental Design And Results

Some parameter settings of the LSTM neural network are shown in TABLE II.

TABLE II. Initial setting of partial parameters of bidirectional LSTM neural network

Hyper parameter	Initial value
Learning rate	0.001
Dropout	0.5
Gradient clipping	5.0
Embedding-dim	300
Optimizer	Adam
Batch-size	64
Hidden-dim	300
Epoch	30

 ISBN: 978-1-61208-688-0

In order to verify the effect of the proposed character vector based Bidirectional LSTM-CRF on TCM entity recognition, this paper sets up three sets of comparative experiments.

Experiment 1: The first set of experiments was to verify the impact of different algorithms on entity recognition in the Chinese medicine field. TABLE III shows the experimental results of the CRF model, the Character vector based bidirectional LSTM-Softmax model (Char-BLSTM-Softmax) and the Character vector based Bidirectional LSTM-CRF model (Char-BLSTM-CRF).

As we saw from the test results in TABLE III, The Char-BLSTM-CRF model has the best experimental results, which is more than one percentage point higher than the F-value of the Char-BLSTM-Softmax model. Due to the Softmax classifier in the Char-BLSTM-Softmax model does not consider the dependencies between the outputs tags, CRF classifiers use Viterbi algorithm for dynamic programming, the output label path is the most scientific. Therefore, the Char-BLSTM-CRF model has a certain improvement compared to Char-BLSTM-Softmax. The CRF model has the worst experimental results because the model requires manual extraction of high quality features. This indicates that the character vector based on BLSTM-CRF model proposed in this paper has a good effect in entity recognition in the field of Chinese medicine.

Experiment 2: This paper constructs the embedding vector with character and words. Experiment to verify the effect of character and word vector on Bidirectional LSTM-CRF model. It is further proved that the Bidirectional LSTM-CRF entity recognition model based on character vector can avoid the influence of inaccurate word segmentation on entity recognition in TCM domain. The Bidirectional LSTM-CRF entity recognition model based on word vector is abbreviated as Word-BLSTM-CRF. The experimental results are shown in TABLE IV.

The experimental performance of the Char-BLSTM-CRF model is 6 percentage points higher than the Word-BLSTM-CRF in F value. Because of the Word-BLSTM-CRF model needs to segment the corpus before labeling, the existing word segmentation tools are inaccurate for the terminology of the Chinese medicine field, resulting in poor experimental results. However, Char-BLSTM-CRF model does not need to word segmentation. It labels each character to avoid errors caused by inaccurate word segmentation.

Experiment 3: In experiment 1 and experiment 2, we fixed the parameters of the LSTM neural network model. In this experiment, by adjusting the parameters, the best parameters suitable for the identification of named entities in the field of Chinese medicine were found. Parameters that can be adjusted include Dropout values, optimizer and learning rate.

TABLE III. COMPARISON OF DIFFERENT MODELS' RECOGNITION OF TCM ENTITIES

Model	P（%）	R（%）	F（%）
CRF	79.26	78.56	78.91
Char-BLSTM-Softmax	89.92	90.70	90.31
Char-BLSTM-CRF	**91.47**	**91.37**	**91.42**

TABLE IV. COMPARISON OF CHARACTER VECTOR AND WORD VECTOR RECOGNITION

Model	P（%）	R（%）	F（%）
Word-BLSTM-CRF	84.05	86.19	85.11
Char-BLSTM-CRF	**91.47**	**91.37**	**91.42**

The Dropout parameter adjustment experiment means that the fixed optimizer is Adam; the learning rate is 0.001; and the Dropout value ranges from 0.1 to 0.5, and is incremented by an integral multiple of 0.1. The experimental results are shown in TABLE V. It can be clearly seen that when the Dropout value is 0.5, the model has the best effect and the F value is the highest.

The optimizer parameter adjustment experiment refers to the fixed Dropout of 0.5, the learning rate is 0.001, and the experimental results of the optimizer can select Adam, Adagrad, SGD, and Momentum. The experimental results are shown in TABLE VI. It can be clearly seen that the fixed Dropout and learning rate, the Adam optimizer have the fastest convergence and the best results.

The experimental results of the learning rate parameter adjustment are shown in TABLE VII. At this time, the fixed Dropout is 0.5, the optimizer is Adam, and the learning rate is set to 0.01, 0.001 and 0.0001. According to the experimental results, when the optimizer and Dropout are fixed, learning rate is 0.001, the effect of the model is optimal at this time.

According to the results of the above three sets of parameter experiments, the parameter combination of the Bidirectional LSTM-CRF model which is most suitable for TCM entity identification is Dropout=0.5, the learning rate is 0.001, and the optimizer is Adam. At this time, the comprehensive performance F value of the model reaches 90% above, the identification of named entities in the field of Chinese medicine has a good effect.

TABLE V. THE DROPOUT PARAMETER ADJUSTMENT EXPERIMENT

Dropout, optimizer=Adam, learning rate=0.001	P（%）	R（%）	F（%）
0.5	**91.47**	**91.37**	**91.42**
0.4	91.31	91.18	91.25
0.3	90.12	90.59	90.35
0.2	88.50	89.26	88.88
0.1	85.48	87.14	86.30

TABLE VI. THE OPTIMIZER PARAMETER ADJUSTMENT EXPERIMENT

Dropout=0.5, optimizer, learning rate=0.001	P（%）	R（%）	F（%）
sgd	81.71	76.45	78.99
Momentum	87.23	82.50	84.80
Adam	**91.47**	**91.37**	**91.42**
Adagrad	77.91	77.32	77.61

TABLE VII. THE LEARNING RATE PARAMETER ADJUSTMENT EXPERIMENT

Dropout=0.5, optimizer=Adam, learning rate	P（%）	R（%）	F（%）
0.0001	87.02	87.44	87.23
0.001	**91.47**	**91.37**	**91.42**
0.01	80.07	87.83	83.77

IV. CONCLUSION AND FUTURE WORK

The identification of TCM entities is a basic work in the construction of knowledge graph in this field. The main work of the article is to focus on this topic. This paper proposes a Bidirectional LSTM-CRF model based on character vector, which uses character vector as input to replace the word vector in traditional deep learning, to avoid the influence of inaccurate TCM segmentation on entity recognition. The bidirectional LSTM with context information is used as the input. The hidden layer of the neural network solves the problem of dependence on long text input and mitigates the gradient explosion. Finally, it accesses the CRF tag inference layer to solve the dependency problem between output tags. In this paper, a lot of experiments have been done on the TCM entity corpus. The results show that the Bidirectional LSTM-CRF model based on character vector is better than other algorithms, and the parameters that are most suitable for TCM entity recognition are found through experiments.

However, from the results, the F value of the best experimental results in the article is only 91.42%, and there is still much room for improvement. In order to seek a bigger breakthrough, the future work will focus on the establishment of a high-quality corpus. In this process, more scholars in the field of Chinese medicine will be introduced to participate in the establishment of the corpus, so that the work has a better effect.

ACKNOWLEDGMENT

The research was financially support by three National Natural Science Foundations (61562045 & 61762051). The Postgraduate innovation fund of Jiangxi Province (YC2017-S349). Jiangxi Province key research and development program key projects (20171ACE50021, 20171BBG70108).

REFERENCES

[1] N.Sager, C.Friedman and M.S.Lyman . “Review of medical language processing: computer management of narrative data”. Computational Linguistics, vol.15, pp.195–198, 1989.

[2] P.Lu, K.H.Lin and C.L.Zhou . “Research of TCM face diagnosis and symptom factor based on NN”. Application Research of Computers, vol.25, pp.2655-2657, 2008.

[3] S.Qiu et al. “Chinese Named Entity Recognition Based on Part of Speech Feature with Edges”. Computer Engineering, vol.38, pp.128-130, 2012.

[4] J.P.Ju, W.W.Zhang, J.J.Ning and G.D.Zhou. “Geospatial Named Entities Recognition Using Combination of CRF and Rules”. Computer Engineering, vol.37, pp.210-212, 2011.

[5] Z.H.Wang. “Name entity recognition using language models”. IEEE Workshop on Automatic Speech Recognition & Understanding. IEEE, pp. 554-559 , 2003.

[6] G.D.Zheng and J.Su. “Named entity recognition using an HMM-based chunk tagger”. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 209–219, 2002.

[7] Y.J.Zhang, Z.T.Xu and T.Zhang. “Fusion of Multiple Features for Chinese Named Entity Recognition Based on CRF Model”. Journal of Computer Research & Development, vol.45, pp.1004-1010, 2008.

[8] L.Chiticariu, F.Reiss,Y.Y.Li,R.Krishnamurthy and S. Vaithyanathan. “Domain Adaptation of Rule-based Annotators for Named-Entity Recognition Tasks”. EMNLP, pp.1002-1012, 2010.

[9] X.Q.Zheng, H.Y.Chen and T.Y.Xu. “Deep learning for Chinese word segmentation and POS tagging”. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp.647-657, 2013.

[10] S.Chowdhury et al. “A multitask bi-directional RNN model for named entity recognition on Chinese electronic medical records”. BMC Bioinformatics, vol.19, pp.75-84, 2018.

[11] K. Xu, Z.F.Zhou,T.Y.Hao and W.Y.Liu. “A Bidirectional LSTM and Conditional Random Fields Approach to Medical Named Entity Recognition”.International Conference on Advanced Intelligent Systems & Informatics, pp.355-365, 2017.

[12] L.Sun,Y.Rao,Y.Lu and X.Li. “A method of Chinese named entity recognition based on CNN-BILSTM-CRF model”. Communications in Computer and Information Science, vol.902 , pp.161-175, 2017.

[13] Z.Q.Xu , S. Li and W.H. Deng . “Learning temporal features using LSTM-CNN architecture for face anti-spoofing”.2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR). IEEE,pp.141-145, 2015.

[14] C.H.Dong,J.J.Zhang, C.Q.Zong, M.Hattori and H.Di. “Character-Based LSTM-CRF with Radical-Level Features for Chinese Named Entity Recognition”. Natural Language Understanding and Intelligent Applications. Springer International Publishing, vol.10102, pp.239-250, 2016.

 ISBN: 978-1-61208-688-0