Symptom Extraction of Internal Medicine Diseases of Traditional Chinese Medicine Based on BERT-BiLSTM-CRF Model

Hanqing ZHAO *, Yuehan Li , Shuai Zhang

Posted Date: 19 February 2024

doi: 10.20944/preprints202402.0957.v1

Keywords: Named entity recognition; Corpus; Information extraction; BERT-BiLSTM-CRF

 doi:10.20944/preprints202402.0957.v1

Article

Symptom Extraction of Internal Medicine Diseases of Traditional Chinese Medicine Based on BERT-BiLSTM-CRF Model

Zhao Hanqing *, Li Yuehan and Zhang Shuai

College of Traditional Chinese Medicine, Hebei University, Baoding 071000, P.R.China
* Correspondence: zhaohq@hbu.edu.cn

Abstract: This study focuses on the reasoning of symptoms of TCM internal diseases. Taking cough as an example, the BERT-BiLSTM-CRF model is used for entity recognition. Experiments show that the model has the best entity recognition effect on three types of texts, including teaching cases, clinical cases and literature data, and the F1 value is up to 0.967. It can effectively identify symptoms, course of disease, tongue condition and pulse condition in TCM diagnosis and treatment texts, which lays a solid foundation for intelligent assisted syndrome differentiation and treatment of TCM. By constructing four types of entity corpora and using three mainstream entity recognition models for experiments, it is found that the BERT-BiLSTM-CRF model has a good entity recognition effect in three types of data: teaching cases, clinical cases and literature data. Experimental results show that the F1 values of the BERT-bilSTM-CRF model in teaching cases, clinical cases and literature data are 0.967, 0.82 and 0.91, respectively, which provides an effective method for information extraction in the field of TCM syndrome differentiation diagnosis and treatment, and lays a foundation for further research on knowledge reasoning.

Keywords: named entity recognition; corpus; information extraction; BERT-BiLSTM-CRF

1. Introduction

Through the clinical experience of the wise ancestors and continuous exploration, the science of traditional Chinese medicine has developed more mature, with a complete theoretical basis and clinical diagnosis and treatment system. Syndrome differentiation and treatment is the main method of traditional Chinese medicine (TCM) diagnosis and treatment, of which the key link is symptom description, and this part of information mostly exists in the form of text in TCM literature resources. In order to carry out knowledge reasoning from these text information, we first need to apply information extraction technology. The development of entity recognition and extraction technology has roughly gone through three stages[1]: entity extraction based on pattern matching, entity extraction based on machine learning and entity extraction. Accurate identification and extraction of the elements of TCM syndrome differentiation is the basis for realizing intelligent differentiation of TCM. On this basis, the use of knowledge graph technology methods can more deeply mine TCM knowledge, and provide more powerful support for clinical auxiliary diagnosis and treatment applications. At present, there have been many related studies on the use of entity recognition models[2–5] for TCM text entity recognition. In recent years, there has been progress in entity recognition in the field of traditional Chinese medicine. However, the diversity, complexity and irregularity of data require more accurate annotation methods. Different models need to be compared to select more accurate and efficient entity recognition to support research and clinical application. In this study, TCM symptom information is annotated and processed, and entity recognition technology is selected for automatic recognition to improve the accuracy and provide support for TCM information extraction and mining.

 doi:10.20944/preprints202402.0957.v1

2. Data and Methods

2.1. Data Sources

The data sources of this study were mainly from three aspects. The data of teaching cases were from the seventh edition of Internal Medicine of Traditional Chinese Medicine textbooks and related teaching reference materials, the data of clinical cases were mainly from the Essence of Modern Chinese famous medical Cases, and the data of literature were mainly from the medical cases in relevant papers published on CNKI. The cough was selected in this study, and all data were collected by manual entry and review. The required text was input into excel sheet for storage, and 100 cases of each data were collected.

2.2. Model Methods

2.2.1. Named entity recognition

Named entity recognition (NER) refers to the extraction of words with specific meanings from the specified free text. In general texts, entities usually refer to organization names, place names, and personal names. However, in the field of traditional Chinese medicine, entities mainly refer to symptoms, tongue images, pulse conditions, disease names, drugs, parts, times, and diagnosis and treatment sites. Named entity recognition (NER) technology has gone through three stages. In the early research, it is generally based on manually built rules and dictionaries[6]. However, it can only be applied to specific scenarios, and cannot be promoted and applied in a large area. Later, traditional machine learning methods represented by CRF[7] and the combination of rules and dictionary methods have reduced the dependence on manual work in entity recognition, but there are some problems such as low efficiency of feature extraction and long training time of model. In recent years, deep learning has made great progress. More text feature information can be captured, which greatly improves the training efficiency of the model, and the automatic extraction function greatly reduces the dependence on manual work.

2.2.2. BiLSTM-CRF

The bilSTM-CRF model mainly includes a four-layer structure of word embedding layer, BiLSTM layer, CRF layer and output layer, as shown in Figure 1. (1) Word embedding layer: The external text input sequence is converted into the word vector sequence of the corresponding word, and the word vector is concatenated according to the window size. (2) BiLSTM layer: it models the semantic information of the word vector sequence, extracts the text feature expression, and outputs to the hidden layer before and after splicing. (3) CRF layer: The transition probability matrix is obtained by multiplying the output sequence of the hidden layer with the parameter matrix, and then the maximum likelihood estimation method is used to solve the model parameters. In the prediction process, the dimension bit algorithm is used to dynamically decode the label sequence that maximizes the objective function. (4) Output layer: directly output the entity prediction label.

2.2.3. BERT-BiLSTM-CRF

The BERT-bilSTM-CRF model framework is shown in Figure 2, which is composed of four modules. The BERT layer generates dynamic word vectors by pre-training the external input text data, and the word vector information becomes the input of the BiLSTM layer to complete the bidirectional training, so as to further extract text features[8]. The attention mechanism mainly extracts the feature information that plays a decisive role in entity recognition from the output results of the BiLSTM layer, assigns the weights, and uses the weight check to directly evaluate which embeddings are the preferred embeddings for specific downstream tasks. Finally, the CRF layer can effectively constrain the dependence between the predicted labels and model the label sequence. In this way, the global optimal sequence [9] can be obtained.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 February 2024 doi:10.20944/preprints202402.0957.v1

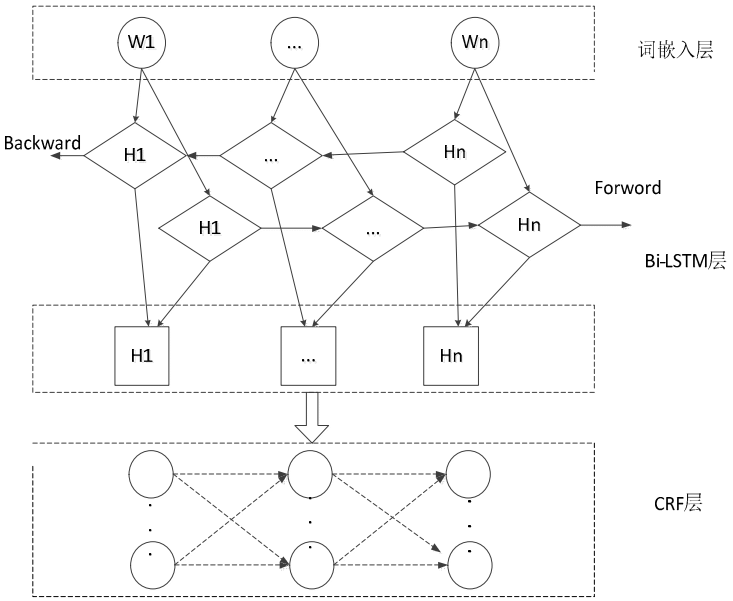

Figure 1. Framework diagram of BiLSTM-CRF model.

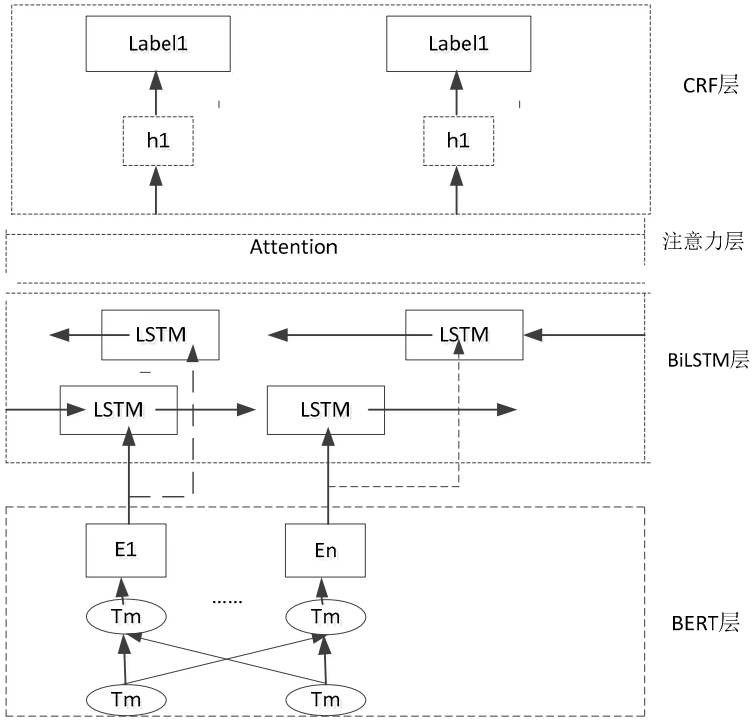

Figure 2. Framework diagram of BERT-BiLSTM-CRF model.

2.2.4. BERT

Bert is built on top of Transformers with powerful language representation and feature extraction capabilities. [10]Achieving state of the art on several NLP benchmark tasks, the BERT model used in this study is partially consistent with the BERT model in 2.2.3.

3. Experimental Procedure

3.1. Data Preprocessing

Firstly, data preprocessing was carried out, including eliminating duplicate data, correcting typos, and removing redundant Spaces. The samples of the three types of data after preprocessing

 doi:10.20944/preprints202402.0957.v1

are shown in Table 1. In the experiment, all kinds of data were divided into training set, validation set and test set in the form of 6:2:2.

Table 1. Data samples.

Data types	Sample raw text
Teaching Cases	Fatigue, nasal congestion and runny nose, cough for more than a month, aggravated with low-grade fever for three days. The tongue is pale, the tongue coat is thin and white, and the pulse is floating
Literature	For more than three years, there was no abnormality in the stomach, abdominal distension after eating, and loose stool. Normal limbs were tired and weak, and sweating was heavy when exercising a little, and it was easy to get cold. One month ago, I felt uncomfortable on the second day after cleaning and sweating, with headache and sore body, nasal congestion and clear nasal discharge, coughing and spitting, and fever of 37.8℃.
Clinical medical records	Liang, male, 56 years old, worker. Coughing with massive sputum production for more than half a month. The symptoms are as follows: Mental distress, sleepiness, fear of cold, pale face, shortness of breath, frequent cough, cough a lot of phlegm, come to the doctor with a disposable cup, phlegm has fast spit full, phlegm is yellow and white with thick phlegm, smell fishy, poor appetite, poor sleep, no sweat, no thirst, no bitterness, two stool tone, dark purple tongue, fat tongue body, edge with tooth stains, white thick and slippery slightly yellow greasy, pulse heavy fine number.

3.2. Construction of corpus

3.2.1. Data annotation

A corpus is a collection of corpora (texts, etc.) collected for special processing of natural languages. It is characterized by representativity, structure, certain scale and detectability. It is a structured, representative and large-scale corpus text collection[11] specially collected for one or more application goals. The use of deep learning technology for information extraction requires a large number of manually labeled corpora. TCM clinical syndrome differentiation information usually includes symptoms, symptom duration, tongue diagnosis, pulse diagnosis, etc[12]. In this study, five entity corpora including symptoms, course of disease, tongue texture, tongue coating and pulse condition were constructed according to the published standards of TCM Clinical Basic Symptom Information Classification and Code, Tongue Diagnosis classification and Code Standard, and pulse diagnosis classification and Code standard.

3.2.2. Training annotation

In the work of named entity recognition, each word needs to be labeled. This study uses BIO labeling method, B-begin represents the beginning of the entity, I-inside represents the middle or end of the entity, and O-outside represents the entity that does not belong to the entity.

3.3. Experimental environment

Python 3.8 was used as the main programming language in this study, and the model was built based on PyTorch 2.2 architecture. The computer operating system was Ubuntu20.04, the memory was 64GB, and the NVIDIA A100 80G GPU computing card was equipped.

 doi:10.20944/preprints202402.0957.v1

3.4. Model parameters

In this experiment, the learning rate, batch-size and epoch were mainly adjusted. The learning rate is set as 0.1, 0.01, 0.001 and 0.0001, the batch size is set as 2, 4, 8, 16, 32 and 64, and the iteration number is set as 10, 20, 33, 40, 40, 50, 60, 70, 80, 90 and 100. By modifying these three parameters, the experiment is carried out continuously. Finally, the most appropriate information recognition model and parameter Settings for each type of TCM diagnosis and treatment data were selected.

3.5. Evaluation Metrics

In this study, precision, recall and F1 value were used to evaluate the model effect. TP represents the number of entities correctly identified by the model, FP represents the number of non-relevant entities identified by the model, and FN represents the number of relevant entities not detected by the model.

$$\mathrm{P} = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times \mathrm{P} \times \mathrm{R}}{\mathrm{P} + \mathrm{R}}$$

4. Results

4.1. Entity recognition results

In the entity recognition experiments of symptoms, tongue, tongue coat, pulse condition, disease course and other entities in three types of TCM diagnosis and treatment data, the entity recognition effect of BERT-BiLSTM-CRF model is better than the other two models. The BERT-BiLSTM-CRF model is used for entity recognition of the test data, and the obtained information recognition results are shown in Table 2.

Table 2. Entity recognition examples of three types of TCM text data using BERT-BiLSTM-CRF model.

Data types	Automatic identification results
TCM electronic medical records	{Symptom: fatigue}, {symptom: nasal congestion and runny nose}, {symptom: cough} more than one month, aggravated with {symptom: low fever}{duration: three days}. {Tongue: light tongue}, {tongue coating: thin white tongue coating}, {pulse condition: floating pulse}
Chinese medicine literature	After eating, {symptom: abdominal distension}, {symptom: loose stool} has been {course of disease: more than three years}, and no abnormality was found in multiple physical examinations. Normal limbs {symptoms: fatigue and weakness}, a little exercise, sweating, easy to get cold, {course: 1 month ago} because of cleaning and sweating {course: the second day} that felt uncomfortable, {symptom: headache} body sour, {symptom: nasal congestion} runny nose, {symptom: cough} spit clear thin, fever 37.8℃.
TCM medical records	Liang, male, 56, a worker. {Symptoms: cough} with massive expectoration {course of disease: more than half a month}. Diagnosis: {Symptom: mental distress, {symptom: drowsing and drowsing}, always {symptom: fear of cold}, {symptom: pale}, {symptom: wheezing {symptom: shortness of breath}, {symptom: cough} frequent, cough a lot of phlegm, come to the doctor with a disposable cup, phlegm is full, phlegm is yellow and white {symptom: thick phlegm}, smell fishy, {symptom: anorexia}, {Symptom: poor sleep}, {symptom: no sweat}, {symptom: no thirst}, not bitter, {symptom: two stool tone}, {tongue: tongue purple dark}, {tongue: fat tongue body, edge with tooth mark, {tongue coating: moss white thick slippery slightly yellow greasy}, {pulse condition: pulse heavy fine number}.

 doi:10.20944/preprints202402.0957.v1

4.2. Model identification results

Aiming at the relevant annotation information in the text data of cough, a common disease in traditional Chinese medicine, three models are used for entity recognition. In the experiment, the values of learning rate, batch-size, and epoch were modified to compare the recognition effect of the model. The experimental results are shown in Table 3.

Table 3. The recognition effects of different models on three types of TCM text data.

Data	Model	Lr	Batch-size	Epoch	Time(min)	Precision	Recall	F1
Teaching Cases	BiLSTM-CRF	0.01	4	100	0.52	0.911	0.921	0.91
Teaching Cases	BERT-BiLSTM-CRF	0.01	32	90	9.63	0.966	0.97	**0.967**
Teaching Cases	BERT	0.001	2	80	8.96	0.932	0.924	0.926
Literature	BiLSTM-CRF	0.01	8	100	0.74	0.78	0.788	0.784
Literature Information	BERT-BiLSTM-CRF	0.01	8	50	10.02	0.905	0.922	**0.91**
Literature	BERT	0.01	16	70	9.67	0.864	0.872	0.863
Clinical records	BiLSTM-CRF	0.01	2	100	1.35	0.797	0.763	0.78
Clinical records	BERT-BiLSTM-CRF	0.01	16	50	26.34	0.824	0.819	**0.82**
Clinical records	BERT	0.001	2	80	22.59	0.792	0.801	0.796

Through experiments, it is found that the BERT-BiLSTM-CRF model has the best effect on entity recognition in traditional Chinese medicine teaching cases, and the F1 value is 0.967. Secondly, the BERT model has an F1 value of 0.926, which achieves good recognition effect. The BERT-BiLSTM-CRF model has the best effect on entity recognition in traditional Chinese medicine literature, with an F1 value of 0.91. The second is the BERT model, and the difference is not very large. The BERT-BiLSTM-CRF model still has the best effect on entity recognition in TCM clinical medical cases, with an F1 score of 0.82, which shows that this method has good robustness in different corpus environments.

5. Discussion

During the experiment, it is observed that introducing the BERT pre-training layer into the BiLSTM-CRF model will greatly increase the training time, but can significantly improve the entity recognition effect. When using the BERT-BiLSTM-CRF model for entity recognition, it can achieve good recognition results for different types of traditional Chinese Medicine text data. The study also shows that the performance of the model is greatly affected by the learning rate and the number of batch samples. Different learning rate Settings will significantly affect the entity recognition effect, and a smaller batch size will reduce the number of samples used in one iteration, but it will also significantly increase the training time of the model. At the same time, by performing entity recognition experiments on the same traditional Chinese medicine text data, it can be observed that the entity recognition effect of the BERT model is slightly lower than that of the BERT-bilSTM-CRF model, but the model training time is significantly shorter. Considering the time factor, the BiLSTM-CRF model shows great advantages when dealing with TCM clinical diagnosis and treatment text data.

6. Conclusions

In this study, three mainstream entity recognition models are used to conduct entity recognition experiments on three types of common texts in the field of TCM syndrome differentiation diagnosis and treatment. A corpus of five types of entities, including symptoms, course of disease, tongue

 doi:10.20944/preprints202402.0957.v1

texture, tongue coating and pulse condition, is constructed for entity recognition experiments. In the experiment, the model parameters suitable for a certain type of data are selected for different models. Through the comparative experiments, it can be found that the degree of text structure and the quality of corpus have an important impact on the effect of entity recognition. In the future research, the entity corpus will be expanded and modified to solve the problems of entity nesting and negative prefixes, and constantly improve the corpus. At the same time, more fusion research methods can be explored, and the TCM knowledge graph technology can be combined to lay a foundation for the next step of knowledge reasoning research.

Funding: This work was supported by National Natural Science Foundation of China (No.82004503) and Science and Technology Project of Hebei Education Department(BJK2024108).

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] Kong Jingjing, Yu Qi, Li Jinghua et al. A review of entity Extraction and its application in the field of Traditional Chinese Medicine [J]. World Science and Technology-Traditional Chinese Medicine Modernization, 2012,24(08):2957-2963.

[2] Li C, Xie D. Research on automatic extraction method of admission record information of traditional Chinese medicine electronic medical record [J]. World Science and Technology-Modernization of Traditional Chinese Medicine,2023,25(5):1615-1622.

[3] Xu Lina, Li Yan, Zhong Xinyu, Chen Yueyue, Shuai Yaqi. Named Entity Recognition of Traditional Chinese Medicine Prescription Text Based on Bert [J]. Medical Information,2023,36(04):32-37.

[4] Gao Jiayi, Yang Tao, DONG Haiyan et al. Chinese Journal of Traditional Chinese Medicine Information,2021,28(05):20-24.]

[5] Liu Andong, Peng Lin, Ye Qing, et al. Research progress of Named entity recognition in electronic medical records [J]. Computer Engineering and Application,2023,59(21):39-51. (in Chinese) DOI:10.3778/j.issn.1002-8331.2303-0237.

[6] Wang Z, Liu L, Yao K, et al. TCM-SAS: A Semantic Annotation System and Knowledgebase of Traditional Chinese Medicine[C]//2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2022: 3727-3732.

[7] Wu Zong-you, Bai Kun-long, Yang Lin-rui et al. Review on Text Mining of Electronic Medical Records [J]. Journal of Computer Research and Development,2021,58(03):513-527.

[8] Shuai Yaqi, Li Yan, Chen Yueyue et al. Entity recognition of chronic bronchitis Traditional Chinese Medicine Cases Based on BERT-BILSTM-CRF [J]. Modern Information Technology,2023,7(05):145-148+152.

[9] Xie Teng, Yang Junan, Liu Hui. BERT-BiLSTM-CRF Model for Chinese Entity Recognition [J]. Applications of Computer Systems, 2020,29 (07):48-55.

[10] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[11] Deng N, Fu H, Chen X. Named entity recognition of traditional Chinese medicine patents based on BiLSTM-CRF[J]. Wireless Communications and Mobile Computing, 2021, 2021: 1-12.

[12] Sun S H. Research on key technologies of information extraction in acupuncture and moxibustion field of traditional Chinese Medicine [D]. Dalian University of Technology,2019.

[13] Yi Junhui, ZHA Qinglin. Review of TCM Symptom Information Extraction [J/OL]. Computer Engineering and Application :1-15[2023-05-09].

[14] Cao S, Wu Q. MC-TCMNER: A Multi-modal Fusion Model Combining Contrast Learning Method for Traditional Chinese Medicine NER[C]//International Conference on Multimedia Modeling. Cham: Springer Nature Switzerland, 2024: 341-354.

[15] Smith,L.N.(2017).Cyclical learning rates for training neural networks.In 2017 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp.464-472).