A Method for Extracting Clinical Diagnosis and Treatment Knowledge for Traditional Chinese Medicine Literature

Kehan Zhen, Dan Xie, and Yu Peng(✉)

College of Information Engineering, Hubei University of Chinese Medicine, Wuhan 430065, China
{dinaxie,pengyu}@hbtcm.edu.cn

Abstract. To realize automatic extraction of rich clinical diagnosis and treatment knowledge from traditional Chinese medicine (TCM) literature, a method using information extraction technology to achieve automatic acquisition of clinical diagnosis and treatment knowledge in TCM is proposed. First, the entity types of TCM clinical diagnosis and treatment were defined, and the corpus base of TCM diagnosis and treatment was constructed. Then, the literature sample set was labelled based on the corpus, several commonly used information extraction models were selected to train the sample set, and the most appropriate models and training parameters were selected to extract clinical diagnosis and treatment knowledge. Finally, a case study of breast cancer was carried out, and the implementation scheme of diagnosis and treatment knowledge extraction based on the literature is described in detail. The experimental results show that the UIE model has the best information extraction effect, and its F1 value is 89.69%. Data mining was carried out on the extracted diagnosis and treatment knowledge, and it was found that the basic principles of drug use were qi and blood circulation and liver and spleen strengthening, and the main drugs were tonifying deficiency and clearing heat. This method provides an effective technical route for the automatic acquisition of TCM clinical diagnosis and treatment knowledge, which can help clinical researchers quickly and accurately obtain diagnosis and treatment knowledge in the literature and has practical value for TCM clinical research.

Keywords: information extraction · TCM literature · rules of diagnosis and treatment

1 Introduction

As an important part of traditional Chinese culture, traditional Chinese medicine (TCM) has a profound historical accumulation and contains a wealth of diagnostic and therapeutic knowledge. In the process of inheritance and development of TCM culture, literature, as an important carrier, has recorded a large amount of valuable medical knowledge and wisdom [1]. Mining the knowledge contained in massive literature based on data mining has become a new scientific research method in the era of big data. Discovering the

M. Zhang et al. (Eds.): CCF NCCA 2023, CCIS 1959, pp. 235–244, 2024.
https://doi.org/10.1007/978-981-99-8764-1_18

knowledge contained in these documents can not only deepen the knowledge of the origin of TCM and its theoretical system but also explore latent therapeutic ideas and methods and apply them to today's clinical practice. However, most of this knowledge exists in unstructured text, and its information is often difficult to obtain and understand quickly and accurately [2]. Therefore, using information extraction techniques to semantically analyse and extract information from TCM literature and transform it into structured data that can be processed by modern computers can greatly improve the utilization of information and research efficiency and provide more valuable data to support TCM research and application [3].

As the influence of TCM continues to grow worldwide, the exploration and utilization of the knowledge contained in TCM literature has attracted increasing attention. Currently, scholars in related fields are deeply exploring the valuable information contained in TCM literature to provide theoretical support and scientific guidance for the modernization and internationalization of traditional medicine. Lu Yongmei [4] et al. designed a deep learning-based sequence annotator for identifying text fragments describing clinical experiences in ancient documents. Gao Su [2] et al. used a joint event extraction model to achieve the initial extraction of sentence-level events from TCM literature. Ma Jie [5] et al. mined the implicit knowledge in TCM medical cases based on the CART algorithm to explore the relationship between the disease and each attribute of the patient's symptoms. In this study, we proposed and established a systematic and rigorous automated knowledge mining process for Chinese medical literature and used the information extraction model to automatically extract the medical knowledge in Chinese medical literature under the guidance of the process. The feasibility and effectiveness of the process in clinical practice were verified.

The main contributions of this paper are as follows:

(1) An automatic extraction method of TCM literature diagnosis and treatment knowledge was proposed, and an empirical study was carried out with breast cancer as an example, which has strong practical significance.
(2) The knowledge base constructed is based on national and industry standards, including "Classification and codes of diseases and patterns of TCM" and "Pharmacopoeia of the People's Republic of China" (2020), as well as classic professional textbooks, such as "Chinese Medicinal Formulas" and "Chinese Materia Medica", with strong authority.

2 Method

2.1 Technical Process

To realize the automatic extraction of treatment knowledge in TCM literature and mining its treatment laws, a standardized process is constructed in this study (see Fig. 1). First, a corpus base of TCM diagnosis and treatment is constructed. Then, TCM-specific literature is retrieved from the literature database, screened and saved. Some texts are manually annotated with the knowledge base, and the annotated sample set is used for model training. The trained model is selected for information. The extracted entities are aligned, and the data are analysed.

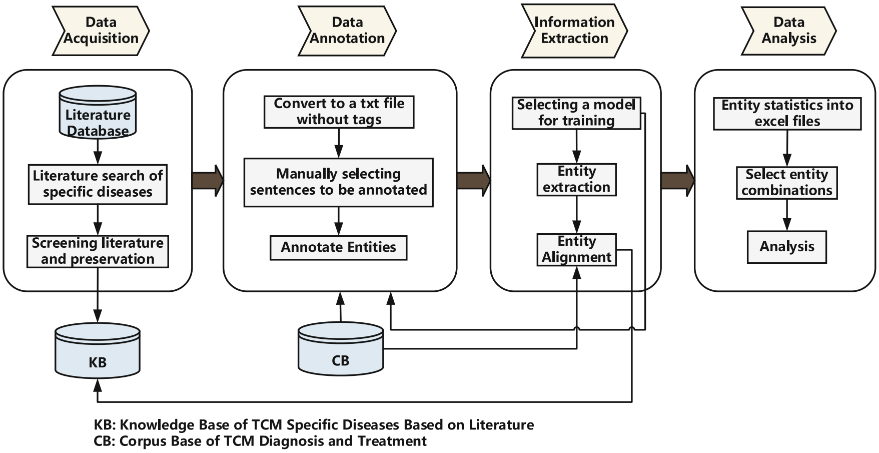

Data Acquisition
Data Annotation
Information Extraction
Data Analysis
Literature Database
Literature search of specific diseases
Screening literature and preservation
Convert to a txt file without tags
Manually selecting sentences to be annotated
Annotate Entities
Selecting a model for training
Entity extraction
Entity Alignment
Entity statistics into excel files
Select entity combinations
Analysis
KB
CB
KB: Knowledge Base of TCM Specific Diseases Based on Literature
CB: Corpus Base of TCM Diagnosis and Treatment

Fig. 1. Overall flow chart of extracting clinical diagnosis and treatment knowledge from the TCM literature.

2.2 Research Objects

The literature data were obtained from Chinese and English databases such as the China National Knowledge Internet (CNKI), China Online Journals (WANFANG DATA), China Science and Technology Journal Database (VIP database), PubMed and Web of Science. Taking the CNKI database as an example, the literature search formula was as follows:

$$SU = Disease\,name\,AND\,(SU = TCM\ OR\,SU = Chinese\,medicine) \qquad (1)$$

The search years were from 2010 to 2022, and the literature was screened according to the inclusion and exclusion criteria, read in full and then screened again, and saved in HTML format.

2.3 Knowledge Base of TCM Diagnosis and Treatment

The knowledge base of TCM diagnosis and treatment includes the corpus base of TCM diagnosis and treatment and the knowledge base of TCM-specific diseases based on the literature (see Fig. 2). The construction of the corpus base of TCM diagnosis and treatment refers to a series of authoritative materials, such as “Classification and codes of diseases and patterns of traditional Chinese medicine” [6], “Pharmacopoeia of the People’s Republic of China” (2020 Edition) [7], “Chinese Medicinal Formulas” [8] and “Chinese Materia Medica” [9], as well as the content provisions and requirements of policies and regulations. The knowledge base of TCM-specific diseases based on the literature mainly contains literature-related information, including the literature title, abstract, author, key words, publication name, publication time, original text and annotated text of the literature.

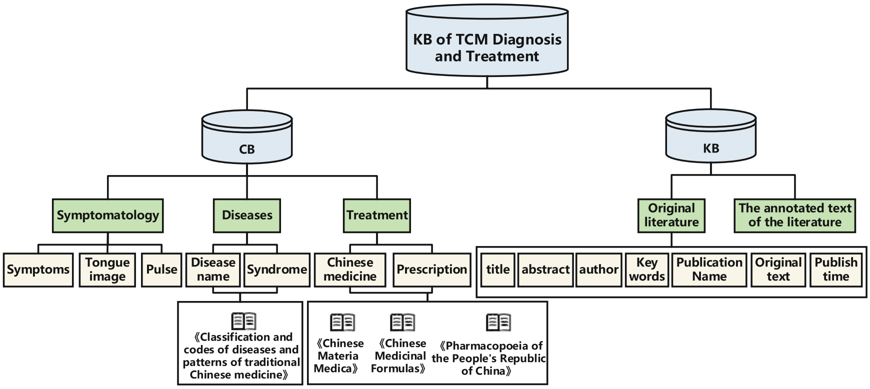

KB of TCM Diagnosis and Treatment
CB
KB
Symptomatology
Diseases
Treatment
Original literature
The annotated text of the literature
Symptoms
Tongue image
Pulse
Disease name
Syndrome
Chinese medicine
Prescription
title
abstract
author
Key words
Publication Name
Original text
Publish time
《Classification and codes of diseases and patterns of traditional Chinese medicine》
《Chinese Materia Medica》
《Chinese Medicinal Formulas》
《Pharmacopoeia of the People's Republic of China》

Fig. 2. Structure of the Knowledge Base of TCM Diagnosis and Treatment.

2.4 Information Extraction Models

In the current field of information extraction, there are a variety of models to choose from. Here, we introduce UIE, BERT, BiLSTM-CRF and BERT-BiLSTM-CRF, which are widely used at present. UIE [10] is a unified modelling model jointly published by the Chinese Academy of Sciences and Baidu in ACL 2022, which realizes tasks such as entity extraction, relationship extraction, event extraction, and emotion analysis and enables good migration and generalization ability among different tasks. BERT[11] is a bidirectional encoder representation based on Transformer proposed by Google AI Research Institute, which aims to pretrain deep bidirectional representation based on left and right contexts of all layers and more fully describe character-level, word-level, sentence level and even intersentence relationship characteristics. BiLSTM-CRF [12] was proposed by Baidu Research Institute in 2015 to conduct text annotation by using LSTM, a two-way long short-term memory network, and CRF, a conditional random field. BERT-BiLSTM-CRF [13] applies the language pretraining model BERT to Chinese entity recognition. It takes the pretraining results as the input of the downstream task BiLSTM-CRF, which not only reduces the workload of the downstream task but also obtains better results.

2.5 Data Analysis Methods

The extracted entities are subjected to entity alignment methods such as synonym replacement and incorrect value deletion, and then the data are analysed. The data analysis methods used were: (1) Frequency statistics were performed using Excel to discover the distribution between commonly used Chinese medicines, prescriptions and different evidence and symptoms in order to grasp the basic statistical characteristics of the data and grasp the overall distribution pattern of the data; (2) Association rule analysis was performed using the Apriori algorithm in IBM SPSS Modeller 18.0 to discover the drug-evidence, drug-symptom, prescription-evidence, prescription-symptom, and the association relationships between evidence-syndromes can help reveal the potential connections and characteristic factors contained in the healing process of various diseases in TCM clinics [14]; (3) Cluster analysis using systematic clustering in IBM SPSS Statistic 25.0 to aggregate valuable homogeneous elements for a more intuitive understanding of their common characteristics [15].

3 Experiment

Following the technical process set in Sect. 2.1, an empirical study was conducted with breast cancer as an example, and the experiment was divided into four steps: data acquisition, data annotation, information extraction and data analysis.

3.1 Data Acquisition

The literature related to the treatment of breast cancer by Chinese medicine was retrieved from the CNKI database, and the search period was from January 2010 to December 2022. A total of 3689 related studies were retrieved, and 2104 were finally included after screening.

Inclusion criteria: (1) The prescription is clear and contains specific TCM components; (2) Treatment is mainly based on traditional Chinese medicine, supplemented by radiotherapy, chemotherapy and other adjuvant therapies; (3) Summary of the experience of famous old Chinese medicine in diagnosis and treatment of diseases; (4) The patient's disease is clearly diagnosed; (5) Adopt the universal diagnostic and treatment evaluation standards recognized by international and domestic counterparts.

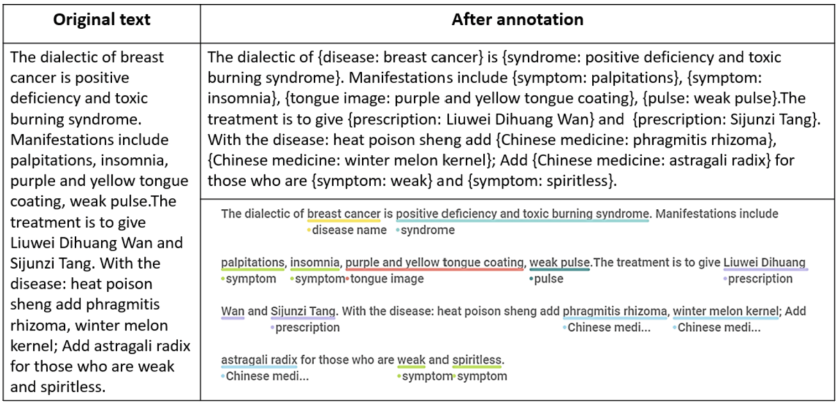
Exclusion criteria: (1) The main treatments were acupuncture and moxibustion, traditional Chinese medicine emotional nursing literature; (2) The research objects were animal and cell literature; (3) Recurring literature.

3.2 Data Annotation

The annotated data came from the literature on diseases treated by traditional Chinese medicine (diseases in 11 directions). Sentences in the text of some studies were randomly selected to form a sample set with 500 data points in total and then annotated using text annotation and the open-source annotation system doccano. Based on TCM theory and clinical practice, seven types of entity labels were established based on the text content in the literature, including Chinese medicine, prescription, tongue image, pulse, syndrome, symptom and disease name. Examples of annotations are shown in Fig. 3. The upper part is text annotation, and the lower part is doccano annotation.

3.3 Information Extraction

The BiLSTM-CRF, BERT-BiLSTM-CRF and UIE models were selected for information extraction, and the labelled sample set was used to train the models. The modified model parameters included learning-rate, batch-size and epoch. The value of the learning rate ranges from 0.0001–0.1, the batch size ranges from 2–64, the number of epochs ranges from 1–100, the precision, recall and F1 score are taken as the evaluation indexes of the model, and the model with the best training results is selected for extraction. The



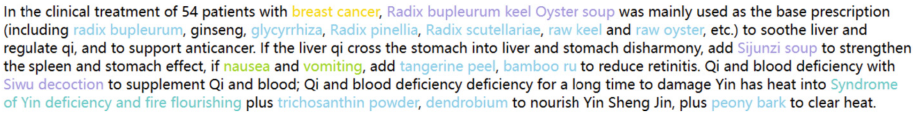
Original text	After annotation
The dialectic of breast cancer is positive deficiency and toxic burning syndrome. Manifestations include palpitations, insomnia, purple and yellow tongue coating, weak pulse.The treatment is to give Liuwei Dihuang Wan and Sijunzi Tang. With the disease: heat poison sheng add phragmitis rhizoma, winter melon kernel; Add astragali radix for those who are weak and spiritless.	The dialectic of {disease: breast cancer} is {syndrome: positive deficiency and toxic burning syndrome}. Manifestations include {symptom: palpitations}, {symptom: insomnia}, {tongue image: purple and yellow tongue coating}, {pulse: weak pulse}.The treatment is to give {prescription: Liuwei Dihuang Wan} and {prescription: Sijunzi Tang}. With the disease: heat poison sheng add {Chinese medicine: phragmitis rhizoma}, {Chinese medicine: winter melon kernel}; Add {Chinese medicine: astragali radix} for those who are {symptom: weak} and {symptom: spiritless}. The dialectic of breast cancer is positive deficiency and toxic burning syndrome. Manifestations include palpitations, insomnia, purple and yellow tongue coating, weak pulse.The treatment is to give Liuwei Dihuang Wan and Sijunzi Tang. With the disease: heat poison sheng add phragmitis rhizoma, winter melon kernel; Add astragali radix for those who are weak and spiritless.

Fig. 3. Annotation example.

results are shown in Table 1. The training effect of the UIE model is relatively good. The precision is 91.78%, the recall is 87.70%, and the F1 score is 89.69%. At this time, the learning rate of the model is 0.0001, the epoch is 100, and the batch size is 16. The trained UIE model was used for entity recognition of 7 types of entities in the TCM literature. An example of the entity recognition results is shown in Fig. 4.

Table 1. Training results of the BiLSTM-CRF, BERT-BiLSTM-CRF, and UIE models.

Model	Precision	Recall	F1 score
BiLSTM-CRF	79.41	67.50	72.97
BERT-BiLSTM-CRF	88.37	90.48	89.41
UIE	91.78	87.70	**89.69**



In the clinical treatment of 54 patients with breast cancer, Radix bupleurum keel Oyster soup was mainly used as the base prescription (including radix bupleurum, ginseng, glycyrrhiza, Radix pinellia, Radix scutellariae, raw keel and raw oyster, etc.) to soothe liver and regulate qi, and to support anticancer. If the liver qi cross the stomach into liver and stomach disharmony, add Sijunzi soup to strengthen the spleen and stomach effect, if nausea and vomiting, add tangerine peel, bamboo ru to reduce retinitis. Qi and blood deficiency with Siwu decoction to supplement Qi and blood; Qi and blood deficiency deficiency for a long time to damage Yin has heat into Syndrome of Yin deficiency and fire flourishing plus trichosanthin powder, dendrobium to nourish Yin Sheng Jin, plus peony bark to clear heat.

Fig. 4. Example of entity recognition results using the UIE model.

3.4 Data Analysis

Frequency Statistics

Formulary Statistics

The prescriptions that appeared 1% (21 times) or more of the total literature were counted. Fifty-nine prescriptions were counted, and the top 5 prescriptions were ranked by their frequency of occurrence, namely, Free Wanderer Powder, Rhinoceros

Bezoar Pill, Bupleurum Liver-Coursing Power, Harmonious Yang Decoction, and Four Gentlemen Decoction.

Statistics of Chinese Medicines

Chinese herbal medicines with frequencies of 1% or more of the total literature were counted, and the statistics revealed that there were 308 eligible Chinese herbal medicines with a total frequency of 58271 times. The 20 drugs with the highest frequency were listed in descending order, namely, astragalus, poria, tangkuei, ovate atractylodes, bupleurum, white peony, codonopsis, licorice, hedyotis, zedoary, bearded scutellaria, tangerine peel, shancigu, cooked rehmannia, ligusticum, curcuma, epimedium, lycium, dioscorea, and coix.

In this study, 308 Chinese herbal medicines were analysed for four qi, and the results showed that cold (33.77%) and warm (33.44%) were predominant; the five flavours were analysed, and the results showed that sweet (32.57%) was the most prevalent, as shown in Fig. 5; the top three categories of meridians were liver (21.12%), spleen (15.67%), and lung (14.61%), as shown in Fig. 6.

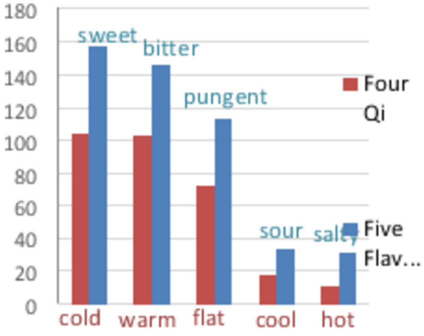


Fig. 5. Histogram of the Distribution of Four Qi and Five Flavours.

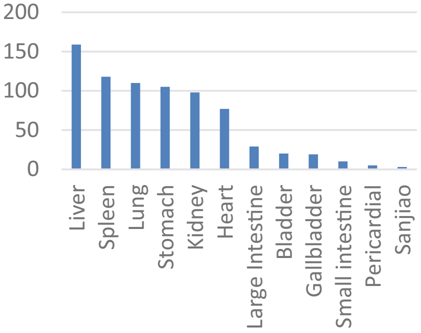


Fig. 6. Histogram of Meridian Distribution.

Association Rule Analysis

The support degree was set to 20%, the confidence degree was set to 80%, the maximum number of previous terms was set to 1 and gradually increased until no new frequent term set was generated, and the association rules were analysed for the top 20 high-frequency Chinese medicines. The association rules were extracted from the pairs and groups of Chinese medicines, among which the highest confidence level was poria-ovate atractylodes (81.752%) for 3 pairs of pairs, 89.565% for poria-white peony-ovate atractylodes (89.565%) for 10 groups of 3 groups, and 90.625% for poria-white peony-tangkuei-ovate atractylodes (90.625%) for 10 groups of 4. The network diagram of association rules is shown in Fig. 7.

Clustering Analysis

Using Pearson's correlation coefficient in systematic clustering, the top 20 Chinese medicines in terms of frequency were clustered and analysed to generate a tree cluster diagram (see Fig. 8). The diagram shows the homogeneity among Chinese medicines, and the shorter the distance corresponding to the horizontal axis, the higher the homogeneity. The Chinese medicines with distances not exceeding 22 were grouped into one category, and five groups of Chinese medicine clustering combinations were obtained by screening.

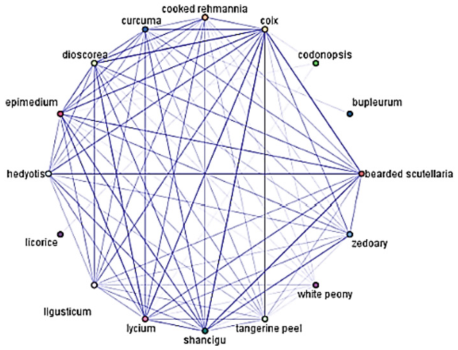


Fig. 7. Association rule network diagram.

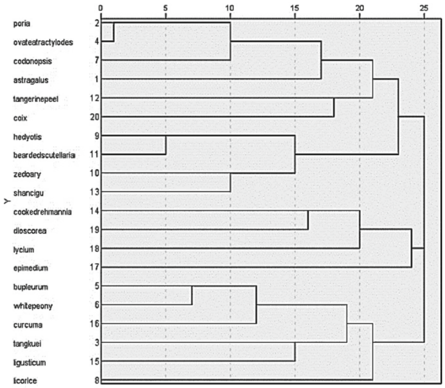


Fig. 8. Tree clustering diagram.

The statistical results of the formulas show that the top 5 formulas with the highest frequency of occurrence have effects related to several categories: reconciliation, regulating qi, tonifying, and treating carbuncles and abscesses. According to the statistics of traditional Chinese medicine, among the top 20 drugs with the highest frequency of use, the main ones are tonifying deficiency drugs, followed by promoting blood circulation and resolving blood stasis drugs and clearing heat drugs, along with the use of diuretic and dampness regulating drugs, qi regulating drugs, and surface relieving drugs. The results of association rule analysis show that the confidence level of the combination of poria and ovate atractylodes is the highest, and they can be traced back to the “Shennong Materia Medica Classic”. This medicine has the effect of strengthening the spleen and dispelling dampness [16]. The results of cluster analysis showed that the main drugs of traditional Chinese medicine for breast cancer were soothing the liver, promoting qi, relieving depression, clearing away heat and detoxification.

4 Conclusion

This study establishes a method for mining treatment laws from TCM literature based on an information extraction model, and this method contains several steps, such as data acquisition, data annotation, information extraction, and data analysis. Data acquisition is the foundation of the whole process, which requires collecting a large amount of TCM literature data. Data annotation and information extraction are also important. Establishing an accurate information extraction model can better extract and analyse the disease information in the TCM literature, which also needs to overcome problems such as entity recognition accuracy. In terms of mining methodology, this study uses a variety of data mining techniques, including association rule analysis and clustering analysis. Through these techniques, the treatment rules in the Chinese medical literature can be better mined.

The information extraction model-based mining method can help TCM clinicians formulate treatment plans and optimize prescriptions more scientifically and rationally. At the same time, this method can also be applied to the promotion of TCM modernization and the research and development of new drugs in TCM. The present study is mainly based on the analysis and mining of existing TCM literature, which inevitably has limitations such as language and cultural background and needs to be validated and revised in the context of actual TCM clinical practice. In addition, as the number of TCM studies is huge and covers a wide range of fields, how to classify and filter them is also a problem that needs to be considered in the follow-up of this study.

Acknowledgements. This research is supported by the Hubei Provincial Administration of Traditional Chinese Medicine (Grant No. ZY2023M072).

References

1. Guangyao, W., Haixia, Y., Xinghua, W., et al.: Approaches and methods of basic theory research of traditional Chinese medicine. Chin. J. Tradition. Chin. Med. **38**(02), 691–694 (2023)
2. Gao, S., Tao, H., Yanzhao, J., et al.: Sentence level joint event Extraction of TCM literature. Inf. Eng. **7**(05), 15–29 (2021)
3. Cheng, W., Lei, X., Yingyi, Q., et al.: Exploration of multilevel information extraction method for Chinese electronic medical records. China Digit. Med. **15**(06), 29–31 (2020)
4. Yongmei, L., Lingmei, B., Li, C., Zhonghua, Y., Tingting, Z., Ying, Y.: Clinical experience extraction of ancient Chinese medicine literature based on deep learning. J. Sichuan Univ. (Nat. Sci. Edn.) **59**(02), 109–116 (2022). (in Chinese)
5. Jie, M., Hongchen, L., Mo, H., et al.: Medical tacit knowledge mining based on CART algorithm: a case study of traditional Chinese medicine. Inf. Sci. **39**(06), 84–91 (2010)
6. State Administration for Market Regulation, Standardization Administration. Classification and codes of diseases and patterns of traditional Chinese medicine: GB/T15657–2021 (2021)
7. National Pharmacopoeia Commission of the People's Republic of China. Pharmacopoeia of the People's Republic of China. Beijing: China Medical Science and Technology Press (2020)
8. Ji, L.: Chinese Medicinal Formulas. China Publication of Traditional Chinese Medicine, Beijing (2016)
9. Gansheng, Z.: Chinese Materia Medica. China Traditional Chinese Medicine Press, Beijing (2016)
10. Lu, Y., Liu, Q., Dai, D., et al.: Unified structure generation for universal information extraction (2022)
11. Devlin, J., Chang, M.W., Lee, K., et al.: BERT: pretraining of deep bidirectional transformers for language understanding (2018)
12. Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF models for sequence tagging (2015)
13. Du, Z., Tang, D., Xie, D.: Automatic extraction of clinical symptoms in traditional Chinese medicine for electronic medical records. BIBM, pp. 3784–3790 (2021)
14. Wenli, X., Jun, L., Shusong, M., et al.: Study on the knowledge association of traditional Chinese medicine symptom knowledge base. J. Med. Inf. **44**(02), 35–41 (2023)

15. Xiangjun, Q., Xinrong, C., Jiahao, M.: Analysis of prescription pattern of TCM treatment of colorectal cancer based on data mining. Chin. J. Traditional Chin. Med. **46**(15), 4016–4022 (2021)
16. Cui, Y., Mi, J., Feng, Y., et al.: Effect and mechanism of Huangqi Sijunzi Decoction on breast cancer fatigue: based on 94 cases randomized controlled trial and network pharmacology. J. Southern Med. Univ. **42**(05), 649–657 (2020)