Journal of Artificial Intelligence Practice (2021) Vol. 4: 39-53
Clausius Scientific Press, Canada

DOI: 10.23977/jaip.2020.040105
ISSN 2371-8412

Research on Entity Recognition and Knowledge Graph Construction Based on Tcm Medical Records

Yanling YANG, Yan LI*, Xinyu Zhong, Lina Xu

Information Engineering Institution of Gansu University of Chinese Medicine, Lanzhou, Gansu, 730000, China

**Corresponding Author*

Keywords: Knowledge graph, Tcm medical records, Name entity recognition, Relation extraction

Abstract: Traditional Chinese medicine (TCM) medical records contain valuable medical information, and are important resources for personalized knowledge analysis, auxiliary diagnosis and treatment, clinical decision support, and drug to use pattern mining of famous TCM doctors. As an effective and novel knowledge management technology, knowledge graph can provide a new way for the inheritance and development of TCM. Constructing medical knowledge graph can potentially help to discover knowledge from clinical data, assist clinical decision-making and personalized treatment recommendation. However, the construction of TCM knowledge graph is still mainly based on structured data, and unstructured texts such as medical records, literature and electronic medical records urgently need to be extracted for mining and analysis. Aiming at the difficulties of word segmentation, entity variety and ambiguity in TCM medical records, this paper proposes a named entity recognition method of deep learning hybrid model based on two-way long-term memory (BILSTM) network and conditional random field (CRF); then by analyzing the process of TCM diagnosis and treatment, the core concepts of TCM are extracted and the ontology layer is constructed; finally, the knowledge graph is constructed by Neo4j, which can provide retrieval, visualization and other functions to help the learning and sharing of TCM knowledge.

1. Introduction

Modern famous scholar Zhang Taiyan said: “the achievements of traditional Chinese medicine, medical records most.” Based on the different understanding of TCM knowledge and different accumulation of clinical experience, each doctor has different views on the diagnosis and treatment of the same disease. Medical Records note their own diagnosis and treatment thinking mode for themselves and future generations to learn and summarize. It is a rare treasure of modern clinical data. It has a strong power for the development of clinical thinking mode and modern medicine. It is worth learning and reference. We can obtain the information of different diseases at different stages of development, diagnosis, treatment, and cure from medical records, especially the diagnosis of disease symptoms and syndromes, the determination of corresponding treatment measures, and the application and compatibility of drugs, to help us improve the ability of diagnosis and treatment.

In the aspect of data utilization, at present, the data used for research are still mainly structured data, while year by year, a multitude of medical records containing rich and valuable knowledge accumulated are left idle. In October 2019, the Central Committee of the Communist Party of China and the State Council issued "<Opinions on Promoting the Inheritance and Innovation of Traditional Chinese Medicine>", pointing out that the research and utilization of TCM medical records and classics should be strengthened, the essence of TCM treasure house should be excavated and inherited, and the modernization of TCM inheritance should be realized by combining digital and imaging technologies. TCM medical record is one major carrier among the TCM inheritance and innovation, and register the entire process of TCM diagnosis and treatment. It has rich medical knowledge and research space of artificial intelligence services, such as natural language processing, intelligent question answering, etc.

The establishment with knowledge graph has become another hot topic in current research fields. In 2012, Google proposed Semantic Web Technology to improve the performance of its Internet search engine, and the knowledge graph is essentially based on this technology [1]. Knowledge graph can show the relationship between things in the real world in the form of entity pairs. At present, disease prediction research combined with the knowledge graph is just beginning in the field of medicine in our country, we can study the construction of knowledge graph of medical knowledge systems to help to promote the development of intelligent medicine. We can aggregate the isolated reality data by knowledge graphs and form a structured graph model to describe the relation among different entities. Aiming at the problem of Knowledge Island in the field of traditional Chinese medicine, knowledge mapping technology provides a solution by integrating information resources to acquire knowledge, and realizes knowledge service better [2].

At present, the research on disease prediction based on knowledge graph in domestic medical field has just begun, so constructing the TCM medical intellectual system combined with knowledge graph owns a certain auxiliary significance for the development of smart medical. Through the research of this paper, we intend to put forward the model and method of constructing knowledge graph for TCM medical records of hypertension from the perspective of TCM inheritance path. According to these, the hypertension knowledge atlas's expression and application come true successfully, and explore a new path for the inheritance of hypertension diagnosis and treatment experience. In conclusion, it is hoped that the research work and achievements in this paper will have important theoretical significance and broad application prospects, and can provide help for the development and innovation in the field of TCM Intellectualization.

Through our discussion of the TCM elementary theory and the process of diagnosis and treatment, the fuzzy relations between TCM entities, such as disease, symptom, syndrome, treatment, prescription, and medicine are clarified in this paper. We first construct an ontology layer to represent the TCM diagnosis and treatment process based on theoretical knowledge and then extract TCM entities from texts and other unstructured data using a named entity recognition (NER) model. Finally, we combine the original ontology layer to get the knowledge corresponding to the map and form the triplet of entity relations. The combination of triplet and knowledge graph can provide reference for TCM doctors in the clinic and assist diagnosis and treatment. In all, the goal of our research is to construct knowledge graph, it plays a guiding role in helping research to build clinical decision support system and personalized medical recommendation service.

2. Related Work

Knowledge graph is a kind of semantic network describing the objective entities, concepts and their relationships in the real world. It makes full use of visualization technology, which can not only describe knowledge resources and carriers, but also analyze and depict knowledge and the

relationship between knowledge [3]. Moreover, knowledge graph is based on big data storage technology, which can draw and display massive knowledge acquired by data mining, machine learning, and information analysis. Knowledge mapping has many advantages, including knowledge semantics, data association, and easy extension. Therefore, the representation, sharing, and application of TCM knowledge through knowledge atlas have become one of the research focuses in the field of TCM [4].

At the terminal of the twentieth century, medical informatization's development in the global has reached a certain mature stage, with large-scale corpus and research methods, and a Unified Medical Language System (UMLS) has been established. In medical research, Named Entity Recognition (NER) and Entity Relation Extraction (ERE) in Natural Language Processing (NLP) have always been hot and difficult issues.

In the information extraction stage, the major duty of entity recognition is intended to find the concept words existing on the basis of the current knowledge architecture from medical records, including diseases, symptoms, syndromes, treatment and prescriptions. The main task of ERE is intended to discover and establish the association among different entity pairs. Including the relationship between disease and syndrome, treatment and syndrome, etc. These two stages also make a good preparation for the coming building about the TCM knowledge graph.

For the research on entity recognition and relation extraction, the widely used methods can be divided into three categories: dictionary-based method, rule-based method and machine learning-based method. Collobert et al [5] first systematically applied deep learning to multiple information extraction links in 2011.They proposed a unified neural network framework and learning algorithm to jointly solve the labeling problem in NLP, which greatly enhanced the portability of the model. This model is a deep learning typical adhibition used in NLP, also as a representative case used in multi-task learning. Lample et al [6] also put forward the BILSTM-CRF model which was on the basis of the influence between words. The results of the BILSTM model were used as the input of CRF, and the state transition matrix between labels was introduced through CRF. The model was trained by integrating the meaning of words and context information. The outcome indicated that the BILSTM-CRF model has achieved more excellent effects to the separate model in English, German, and other open corpora. At present, in the medical field. Jia et al [7] discussed the data sources, research contents and application prospects of TCM knowledge graph architecture. Ruan et al [8] combined the multi-strategy learning with TCM knowledge graph to come true the semi-automated construction in it, the more is helpful with the intelligent service at clinical prescription.

This paper solves the problem of knowledge base acquisition from the root by constructing the relevant named entity recognition model with high accuracy and practicability and the relationship extraction model. The system performs the following tasks to extract knowledge graph from the text: (i) Named Entity Recognition (NER) and (ii) Relation Identification. On this basis, the knowledge graph and visualization of hypertension TCM medical records are constructed by using Neo4j graph database. The constructed knowledge graph realizes the explanation of invisible knowledge such as clinical experience and medication rules of hypertension, and the reasonable classification of syndromes is carried out, and the corresponding diagnostic criteria of syndromes and the visualization of the relationship among TCM syndromes and prescriptions are established. This paper divides the task into two phases:

(a)Knowledge extraction stage: In this stage, NRE and ERE are the main tasks. Jointed with the constructed corpus, the BILSTM-CRF model is applied to distinguish named entities, and BIO + relational model + location labeling is used to deal with entity relation extraction [9].

(b)Knowledge storage stage: The major duty in this step is intented to store knowledge by graph structure and realize the visualization stage through Neo4j. Through the form of knowledge graph,

diseases, symptoms, syndromes, treatment and prescription and their relationship are related. The model framework is shown in Figure 1.

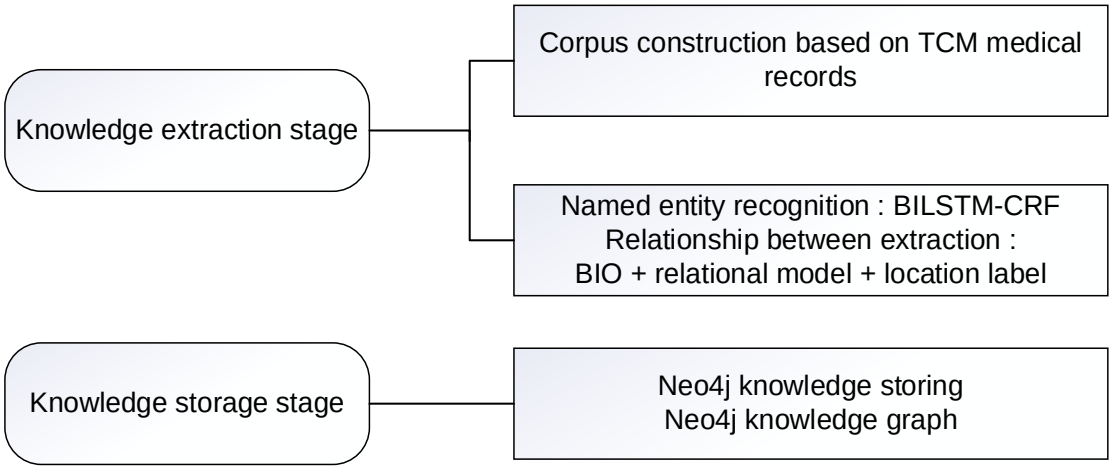

Fig.1 Overall Model Framework

3. Construction Model

3.1 Design of Knowledge Extraction Model: Bilstm-Crf

In the traditional neural network, the existing problems are: a) no connection between the same levels; b) impossible to capture the results after the words are labeled.

The RNN connect to hide layers is used to solve the above problems. However, in practical applications, due to the gradient dispersion problem, it is usually hypothesized that the present condition just has associated with the prior adjacent node state, which reduces the complexity of the model [10]. Adopt for utilizing the context messages effectively, the standard RNN one-way sequential processing method is extended to BILSTM network. BILSTM is constituted with FLSTM and BLSTM, modeling statements from forward and backward directions. Both directions are connected to an output layer, making the word sequence of the output layer contain the text messages not just from the front statement but also the below statement. Using CRF model to realize Chinese word segmentation can achieve the recognition of new words to a certain extent while having good learning performance [11]. In this paper, the neural network model combined BILSTM and CRF is used to classify entities. The structuration of this model shows in Figure 2. It combines the context-related word messages, recommends the words distributed expression into the extraction of features, and fully utilized the relationship in words to tags so that the recognition effect be maximizing.

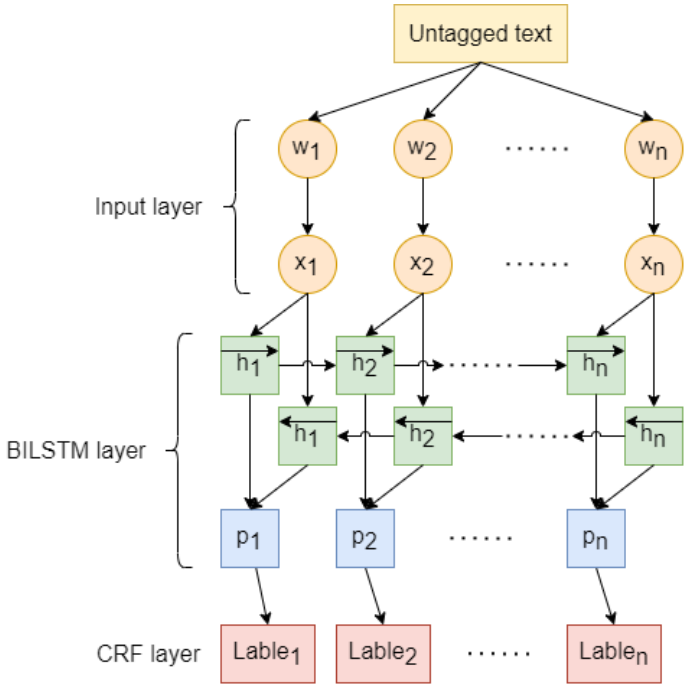

Fig.2 Bilstm + Crf Model Framework

(1) Input layer

The base layer is an input layer transforming literals as computer-processed data. Take the word sequence $\{w_1, w_2, \ldots, w_n\}$ as the initial data from the unlabeled TCM electronic texts and input it in the pretrained Word2vec model. We get the corresponding low dimensional (low than d) dense mapping vector $\{x_1, x_2, \ldots, x_n\}$, of each contains itself features and the semantic features between the characters.

(2) BILSTM layer

The medium layer is an implicit layer based on BILSTM, which consists of FLSTM and BLSTM[10]. LSTM can selectively memorize important information and forget unimportant information . Nerve cell can combine its own feature vector x_i with the hidden layer information h_{i-1} output of the previous cell to calculate the forgotten information f_i, memory information r_i, and temporary cell state $\tilde{c}_i$. Then it combine with the previous cell state c_{i-1} to calculate the current cell state c_i, and finally get its hidden layer information h_i. The relevant calculation formulas are shown as follows:

$$f_i = \sigma(W_f \times [h_{i-1}, x_i] + b_f)$$

$$r_i = \sigma(W_i \times [h_{i-1}, x_i] + b_i)$$

$$\tilde{c} = \tanh(w_c \times [h_{i-1}, x_i] + b_c)$$

$$c_i = f_i \times c_{i-1} + r_i \times \tilde{c}_i$$

$$h_i = \sigma(w_o \times [h_{i-1}, x_i] + b_o) \times \tanh(c_i)$$

Finally, the BILSTM acquires the corresponding status serials $\{\overrightarrow{h_1}, \overrightarrow{h_2}, \ldots, \overrightarrow{h_n}\}$ and $\{\overleftarrow{h_1}, \overleftarrow{h_2}, \ldots, \overleftarrow{h_n}\}$ from the hidden layer of forward and backward. The characteristic vector matrix $P = \{p_1, p_2, \ldots, p_n\}$ is spliced and mapped to the number of labels in K dimension, and p_i contains the possibility of corresponding labels.

(3) CRF layer

The third part is an output layer based on CRF, which introduced the tag transfer matrix $T_{i,j}$ with constraints to enhance the semantic correlation and to label the sequence. Combined with the output of BLSTM layer's output, the corresponding $scores(x, y)$ to n kinds label results are obtained by calculation. Finally, the global optimal sequence label results $\{Label_1, Label_2, \ldots, Label_n\}$ are obtained according to the optimal fraction. Among them. The calculating formula about $score(x, y)$ and loss l are as follows:

$$score(x, y) = \sum_{i=1}^{n+1} t_{y_{i-1}, y_i} + \sum_{i=1}^{n} p_{i, y_i}$$

$$l = \log \sum\nolimits_{y} \exp(score(x, y))$$

3.2 Entity Relation Extraction Model

The concept of relationship extraction was first put forward at the MUC-7 conference in 1998. Since it was proposed, it has been well developed in many fields, such as Finance, Health Care and Education[12]. Entity relation extraction refers to the extraction of relationship triples (entity 1, relationship, entity 2) from sentences. For example, Jack Ma founded Alibaba in Hangzhou. Among them, Jack Ma is entity 1 and Alibaba is entity 2, so the extracted triple is (Jack Ma, founder of Alibaba). The entity relation extraction method solves the problem of relationship classification between target entities in the original text. and extensive apply on text summarization, auto question answer system, knowledge graph, search engine, machine translation and so on.

In this paper, ERE's main task in TCM medical records is to study the predefined relationship between the two entities, such as the relationship between diseases, symptoms and treatment. Therefore, a new attempt is made to design a joint model, that is, entity relation model, combined with entity recognition, and using the method of BIO+ relation model + location annotation, the original annotation method is added to a set of predetermined relationships and transformed into triples, the form of <entity information, entity relationship, entity position in the relationship>.

Entity relation extraction is a follow-up task of named entity recognition. Within an identical sentence scope, the connection between two entities is relatively clear. In the process of research, most studies will only consider the relationship between two entities in a sentence, without considering the relationship between sentences and sentence entities. This paper combines entity recognition and considers the location of sentences. Then reference for the context characteristics between two entities, the two entities' semantic relationship with the highest probability is predicted. Therefore, the ERE model in this paper has four layers: first, it's a word embedding model that transformed the text content in TCM medical records into word vectors; second, it's a BILSTM+CRF model that can automatically extract the features we need. The third layer is to use unique hot coding to combine the quantized text features with the entity location, so as to form a triple format, which can be predicted more accurately. The fourth layer, a softmax function layer, realizes the transform from the third layer's relational classification to the maximum probability problem. The following figure 3 is the entity relation extraction model diagram.

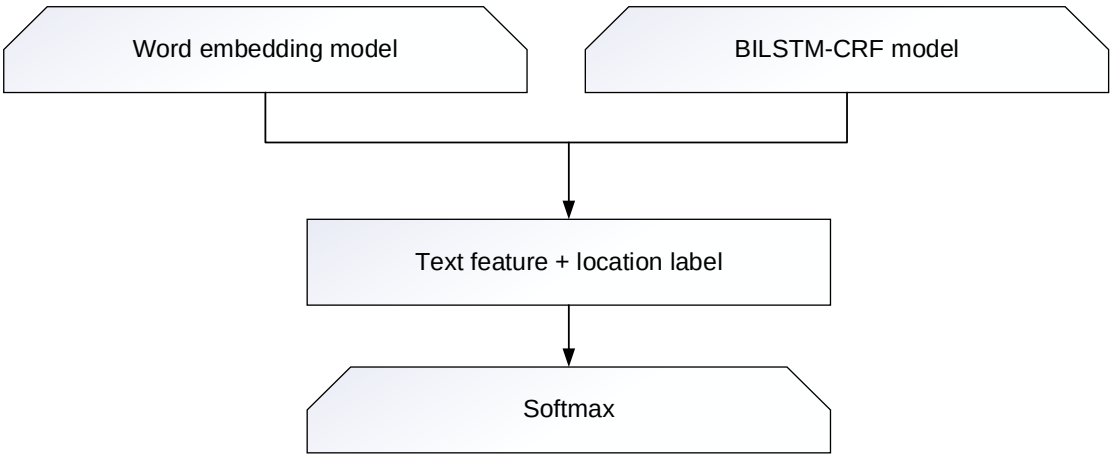

Fig.3 Entity Relation Extraction Model

3.3 Knowledge Storage Model

With the rapid development and application of knowledge graph in China, and has become an important development direction of knowledge representation and utilization on the Internet [13]. At present, there are four main steps in the process of building knowledge graph: Data acquisition, Information acquisition, Knowledge fusion and Knowledge processing. Data acquisition is the basis of the construction process, and the data source includes structured, semi-structured and unstructured data. In the construction of knowledge graph in the field of medicine, the main data sources for constructing the knowledge graph are medical professional papers, books and documents, medical records and electronic medical records. Common medical records, medical literature and other data belong to unstructured data, is the research focus of knowledge extraction. The knowledge in the existing unstructured and semi-structured data is extracted from different formats or representations and processed into the same form of data, including entity extraction, relationship extraction and attribute extraction 14]. After acquiring entity, relationship and attribute information, the next step is knowledge fusion, that is, cleaning and integration of information, including coreference resolution and entity disambiguation, to ensure the correctness and logic of knowledge. Finally, through knowledge processing, including ontology extraction, knowledge reasoning, knowledge discovery and quality evaluation, a structured and networked knowledge system is obtained to form a knowledge graph [15].

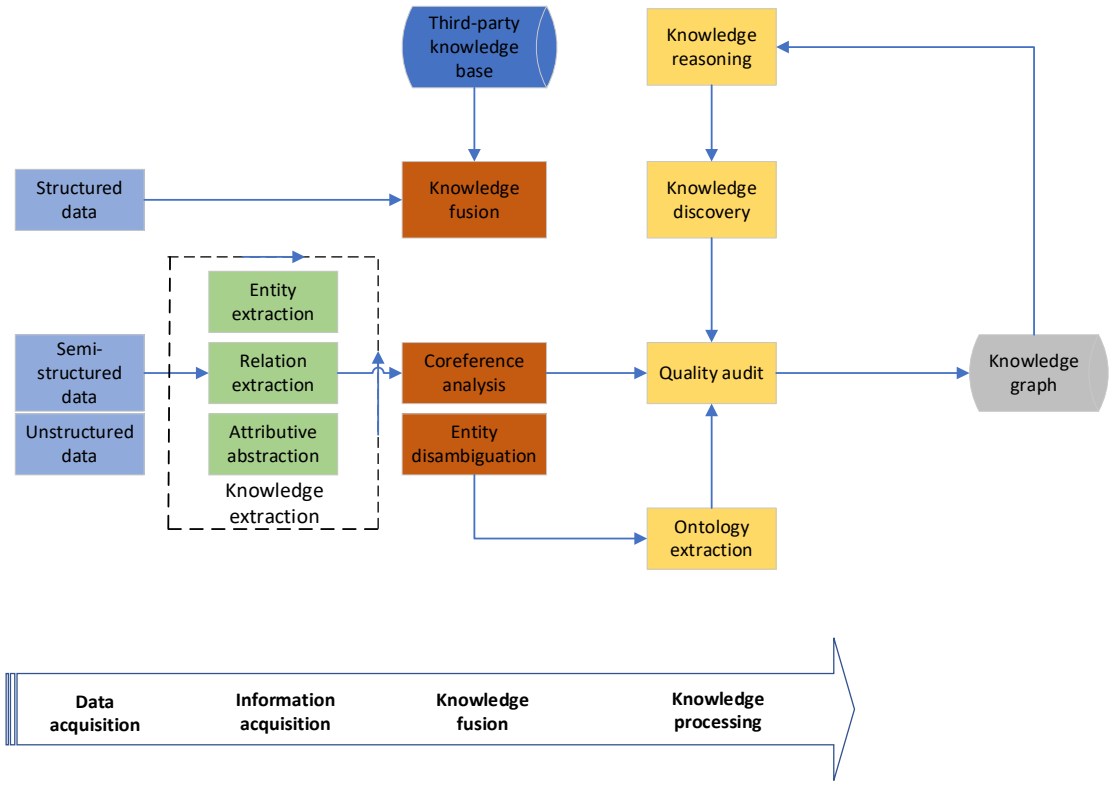

Fig.4 Construction Process of Knowledge Graph

This paper uses bottom-up construction method to construct knowledge graph. Among many knowledge base systems, performance, flexible design and good development. Neo4j is the most popular graph database, which consists of vertices (nodes), edges and attributes. Each vertex and edge can have multiple attributes [16].

Neo4j can also be regarded as a high-performance graph engine, which has all the features of a mature database. Programmers work on an object-oriented, flexible network structure instead of strict, static tables-but they can enjoy all the benefits of a fully transactional, enterprise-grade database. The main advantages of Neo4j are:

1) Neo4j graphic database uses graphics to save data, which can simply and clearly show the relevance of the data.

2) Neo4j graphics database uses the format of unstructured data, so it has very good extensibility. The data can also be modified at any time, which facilitates subsequent maintenance.

4. Data Processing

4.1 Data Source

The research data of this paper mainly comes from ancient and modern medical records cloud platform software, sorting out all TCM medical records related to hypertension on the platform. Inclusion criteria :①Patients diagnosed as hypertension or vertigo were clearly recorded in medical records.②Hypertension was the main complaint in the treatment.③Complete data, including clinical manifestations, pathogenesis analysis, treatment and medication. According to the authoritative diagnostic criteria and the guidance of famous doctors, the relevant hypertension medical records in the platform were manually retrieved and screened. The screened prescriptions were uniformly entered into the Excel database by serial number, ID, patient name, gender, age,

medical records content, TCM diseases, syndromes and medical records sources. In the course of the study, the text of each medical case included in the training set was integrated, and the content of the original medical case text was segmented by single character as an interval. All statements in the training set were classified according to diseases, symptoms, syndromes, treatment and prescriptions. Finally, a total of 435 medical case data were entered.

4.2 Artificial Sequence Labeling

Sequence labeling is one of the basic problems we often encounter in solving NLP problems. In short, given a series of sequences, each element in the sequence is labeled with a tag, which can be used for depth analysis of this series of sequences. Sequence labeling can generally be divided into two categories: one is the original labeling: each element needs to be labeled as a label. Two is joint labeling: all segments are labeled as the same label. Words are not only a syllable unit, but also a unit carrying meaning, namely morpheme. According to the information carrying characteristics of Chinese characters, a word level model is proposed, which can also avoid the problems caused by accidental poor segmentation. For the equivalent annotation of entity recognition, the composition of the annotation includes two aspects: the category of the entity and the position in the entity. In this study, BIO representation is used to represent the category and location of entities. Each element is labeled as ' B-X ', ' I-X ' or ' O ', then take the character as the smallest dimension unit. In the BIO representation, B indicates the head of the entity, I indicates the intermediate entity, X indicates the type, and O does not belong to any type.

In the labeling process, the entity of traditional Chinese medicine as the research object, according to the ' label, entity ' format, the entity is classified into categories, Table 1 introduces the relevant information of five types of entities.

Table 1 Entity-Related Information

Entity category	Class definition	Sequence labeling
Disease	Disease is composed of a group of characteristic clinical symptoms.	Vertigo.
Symptom	Symptoms are various abnormal states or discomforts shown by patients.	Light-headedness.
Syndrome	Syndrome is the diagnosis made by summarizing and analyzing various symptoms and signs of patients at a certain stage.	Blood stasis in liver and kidney deficiency.
Treatment	According to the dialectical results and the principle of “suiting time, suiting measures to local conditions, suiting people”, the treatment method is determined.	Nourishing kidney and liver disperse blood stasis and dredge collateral.
Prescription	The doctor prescribed the patient.	Nourishing liver stopping endogenous wind jusculum.

The specific steps are as follows:

Step 1: To annotate the named entity of the training and test data set, and complete the preliminary processing of the overall medical case text data set.

Step 2: In the process of preliminary treatment, the entity is divided into five categories of diseases, symptoms, syndromes, treatment, and prescription.

Step 3: BIO labeling method is used for training and testing text data. In the specific labeling process, the entity markers shown in Table 2 are used to distinguish. It should be noted that :① Every word in the training corpus should be marked.②Space between words and identifiers.③An empty line between sentences.

Table 2 Bio Tall Set

Entity labeling	Head lebel B	Midst lebel I	Tail lebel O
disease	B-disease	I-disease	O-disease
symptom	B-symptom	I-symptom	O-symptom
syndrome	B-syndrome	I-syndrome	O-syndrome
treatment	B-treatment	I-treatment	O-treatment
prescription	B-prescription	I-prescription	O-prescription

A relatively complete knowledge graph of hypertension can be obtained by using the form of "entity, relationship, entity". As shown in figure 5 below, the five most basic categories of syndrome, disease, symptom, treatment and prescription are defined, and five new semantic relationships are created under the advice of professional doctors. These five semantic relation tables are listed in Table 3 below.

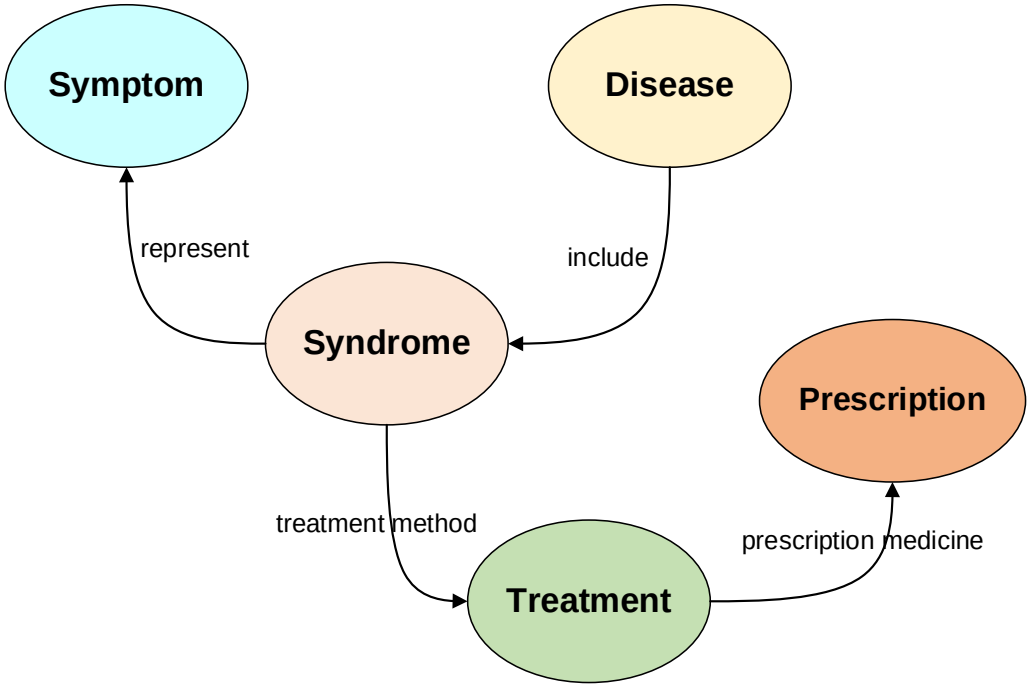

Fig.5 Ontology Layer of Tcm Diagnosis and Treatment

Table 3 Explanation of Relations in Ontology of Traditional Chinese Medicine

Semantic relation	Explanation
X has Y	Disease- syndrome
X represent Y	Syndrome- symptom
X diagnosis Y	Syndrome- treatment
X use drug Y	Treatment- prescription
Y compose X	Prescription-Drug

5. Conclusion and Analysis

5.1 Evaluation Criteria

With the above models, each model was tested and evaluated using our data set. In the evaluation of the model, we used accuracy, precision-recall, F1, and Kappa to evaluate our model. The parameter values needed to evaluate these models are shown in Table 4.

Table 4 Evaluation Index

Index	Calculation formula
Accuracy	$\frac{TP+TN}{TP+FP+TN+FN}$
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
F_1	$2\times\frac{\mathrm{Pr}\,ecision}{\mathrm{Pr}\,ecison+\mathrm{Re}\,call}$
Remarks	TP and TN respectively represent the number of positive samples and negative samples under the correct classification. FP and FN respectively represent the number of negative sample classified as positive samples and the number of positive samples classified as negative samples.

5.2 Results and Analysis

A. BILSTM-CRF model

Table 5 shows that the sample dataset is trained by the BILSTM-CRF model, and then tested by the results of the model training. The model rounds are 30 and 50. The results are compared as shown in Figure 6. The comprehensive evaluation index level of the 50 rounds model is more than 81.30 %, which is better than that of the 30 rounds model, and its recall rate has been greatly improved. For each category named entity test results as shown in table 6, disease accuracy rate is 73.87 %, symptom accuracy rate is 75.93 %, syndrome accuracy rate is 72.33 %, treatment accuracy rate is 68.13 %, and prescription accuracy rate is 89.15 %. The comparison of the detection results of various entities in Figure 7 shows that the accuracy rates of different entities are quite different, and the evaluation results of various aspects of prescription entity recognition are much higher than those of other entities. This situation may be related to the structural characteristics of entities, or may be related to the uneven distribution of various entities in the small sample annotation data set. The more entities are the higher the accuracy rate is in the model test. In this experiment, continuous iterative model training and testing were adopted to increase the number of rounds, and finally complete entity annotation data corpus was formed.

Table 5 Model General Evaluation Structure

Model rounds / rounds	precision	Recall	F-balanced	Accuracy
30	80.03%	80.50%	80.27%	88.87%
50	81.30%	83.37%	82.32%	90.13%

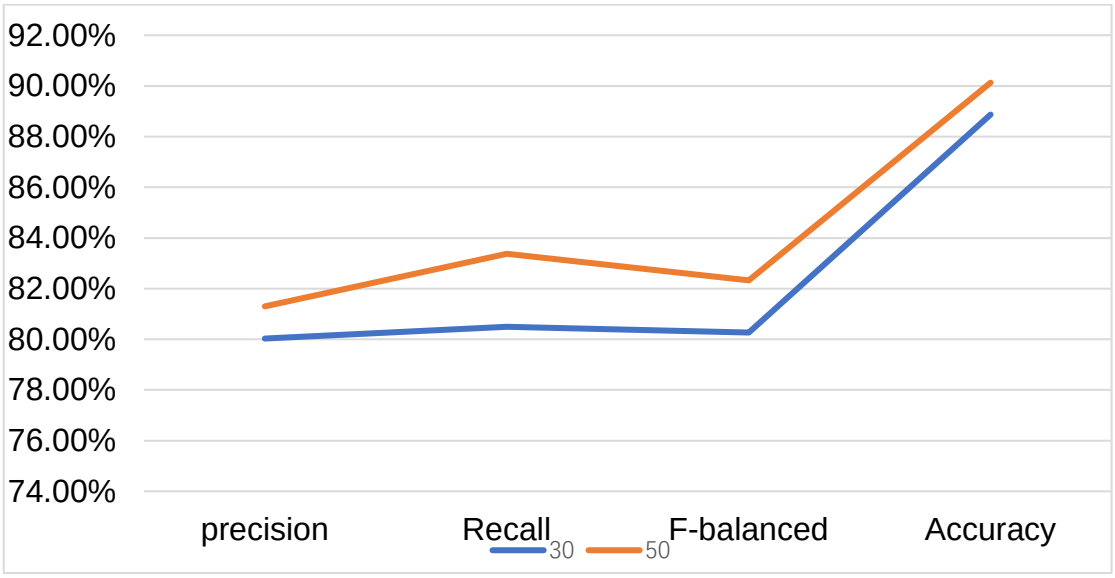

Fig.6 Model Overview Comparison Chart

Table 6 Evaluation Results Of All Kinds of Named Entities

Name entity recognition	Precision		Recall		F-balanced	
Model rounds / rounds	30	50	30	50	30	50
disease	66.53%	73.87%	70.98%	73.21%	68.68%	73.54%
symptom	71.49%	75.93%	72.33%	78.21%	71.91%	77.05%
syndrome	71.58%	72.33%	72.93%	82.32%	73.74%	77.00%
treatment	65.99%	68.13%	62.50%	65.75%	68.59%	77.92%
prescription	88.68%	89.15%	87.43%	92.98%	88.00%	91.02%

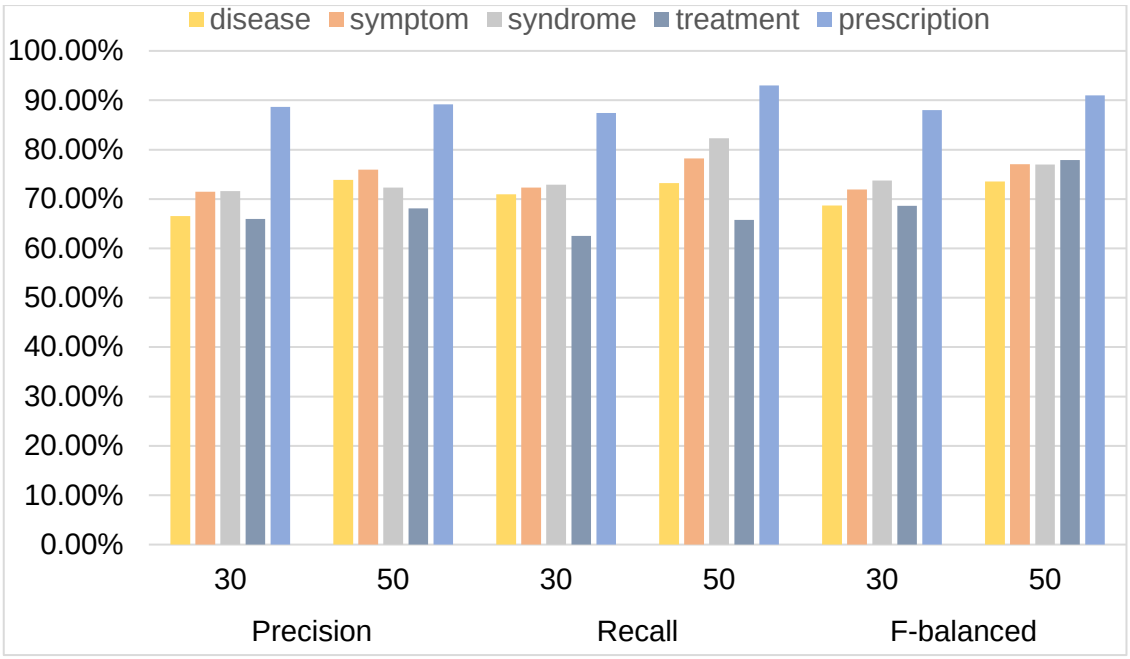

Fig.7 Comparison of Evaluation Results of All Kinds of Named Entity Recognition

In order to find the problem of great disparity in the accurate values of different categories, the entity distribution of the sample labeled data set is analyzed, we can obtain that the proportion of prescription entities in the sample data set is the largest and the proportion of treatment entities is the smallest. Therefore, it can be concluded that entities with more data have higher accuracy in model testing.

B. Knowledge graph show

The knowledge graph is shown by Neo4j. Data visualization is used to visualize the query results based on knowledge graph, including TCM knowledge query and the diagnosis and treatment path of TCM. As shown in Figure 8, this graph can show the entire diagnosis and treatment path of Disease-Syndrome-Therapy-Prescription in TCM. Users can click on the nodes in knowledge graph to view the corresponding attributes, which can help users to learn the complete knowledge of TCM diagnosis and treatment. We can clearly see the relationship between single disease and multiple symptoms and different syndromes. Figure 8 is a summary of all entities and relationships in the TCM records of hypertension. The orange nodes represent the disease entity, the red and cyan nodes the symptom entity, the green nodes the syndrome entity, the blue nodes the rule entity, and the light brown nodes the prescription entity. The diagram clearly reflects the relationship between various entities and attributes. Users can learn from the diagram more pertinently the complete diagnosis and treatment path of disease-symptom-syndrome-treatment-medicine syndrome, and provide effective data support for clinical diagnosis.

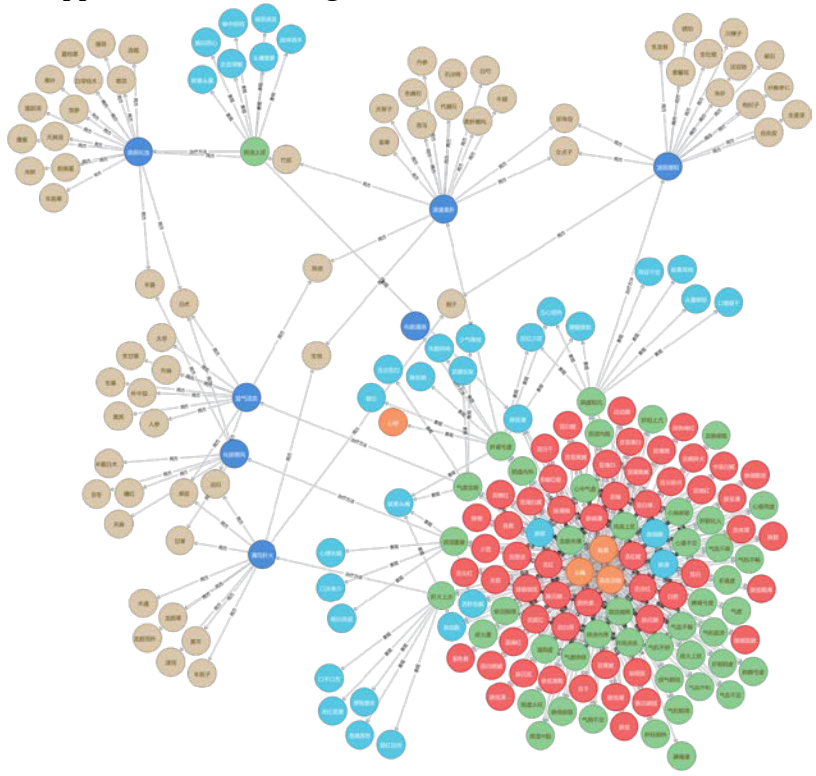

Fig.8 All Entities and Relationships

6. Conclusion

Due to the unique text characteristics of TCM medical records and the lack of large-scale corpus, and there is no widely unified annotation standard, the research is difficult. Based on the model methods widely used in the current research field, this paper combines word embedding technology with neural network BILSTM and CRF model for named entity recognition, and the research proves

that the model has good performance. However, there are still some problems in this process: for example, the amount of data in the data set is not large enough, the rules set need to be changed in different applications, and the data also need to be further optimized, so that the Agent model can play a better role.

In the future work, we will use more corpora to train the NER model, which can increase the accuracy of entity recognition. Besides, we will try our best to design new aggregation algorithms to improve the quality of knowledge graph. Finally, after completing the infrastructure of TCMKG, we will also develop more practical features, such as the answer to questions and the recommendation of knowledge.

The TCM knowledge graph constructed in this paper can be regarded as a dynamic graph, which can provide clinicians with the relationship between observation concepts in time, and can also be used as a reference for the summary of disease syndrome types.

On the other hand, the knowledge graph can generate a very complex network, which can effectively show the reasoning relationship between meta-knowledge in the network. Reasoning represents the boundary of the conceptual relationship between conceptual nodes. All boundaries contain weights and directions, where weights can be learned and adjusted to help clinicians observe and filter relationships to obtain relationship patterns. The direction represents the sequential relationship, for example, drug A can treat disease B, disease C shows the same symptoms as D, and the weight of the boundary represents the strength of the relationship.

The knowledge graph designed in this paper also plays an auxiliary role in decision-making. (1) based on the knowledge graph of ancient Chinese, the part of speech tagging collection is developed; (2) the words in the corpus are vectorized; (3) the BILSTM-CRF model is constructed to mark the part of speech; (4) entity naming recognition and entity relation extraction are carried out, and the TCM knowledge map is designed and visualized by using Neo4j graphics database.

The knowledge mapping function realized in this paper is still insufficient and needs to be further expanded. In the future, we can also try to extend medical knowledge to the corpus, so that we can identify new entity models, build new entity relationships, and greatly improve the availability of the knowledge graph of the medical system.

Acknowledgement

Fund Project: National Administration of Traditional Chinese Medicine Project (2305181101) : Health Information Platform of Traditional Chinese Medicine Diagnosis and Treatment Areas of Basic Medical and Health Institutions in Gansu Province.

References

[1] Weng H, Liu Z, Yan S, et al. A Framework for automated knowledge graph construction towards Traditional Chinese Medicine, Proceedings of the International Conference on Health Information Science, F, 2017.

[2] Yu T, Li J, Yu Q, et al. Knowledge Graph for TCM Health Preservation: Design, Construction and Applications. Artificial Med, 2017,77,48-52.

[3] Sang S. A Knowledge Graph based Bidirectional Recurrent Neural Network Method for Literature-Based Discovery; Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), F, 2018.

[4] Xie Y, Hu L, Chen X, et al. Auxiliary Diagnosis based on the Knowledge Graph of TCM Syndrome [J]. Computers, Materials and Continua, 2020, 65(1): 481-94.

[5] Collobert R, Weston J, Karlen M, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research, 2011, 12(1):2493-2537.

[6] Lample G, Ballesteros M, Suramanans S, et al.[C] Neural architectures for name identity recognition//Proc of Conference of the North American Chapter of the Association for Computational Linguistics:HumanLanguageTechnologies.2016:260-270.

[7] L.R. JIA, J. LIU, T. YU, Y. DONG, L. ZHU, B. GAO, and L.H. LIU, "Construction of traditional Chinese medicine

knowledge graph," Journal of Medical Informatics, vol. 36, pp. 51-59, August 2015.

[8] T. RUAN, C.L. SUN, H.F. WANG, Z.J. FANG, and Y.C. YIN, "Construction of Traditional Chinese medicine knowledge graph and its application," Journal of Medical Informatics, vol. 37, pp. 8-13, April 2016.

[9] Wan H, Marie-francine M, Walter L, et al. Extracting relations from traditional Chinese medicine literature via heterogeneous entity networks [J]. Journal of the American Medical Informatics Association Jamia, 2016, (2): 356.

[10] BIN JI,RUI lIU, SHA SHA LI,JIN TAO TANG, et al. A BILSTM-CRF Method to Chinese Electronic Medicine Record Named Entity Recognition[A].International Association of Applied Science and Engineering:2018:6.

[11] Dai Z,Wang X,Ni P, et al. Named recognition using BERT BILSTM CRF for Chinese Electronic Health Records[C]//2019 12th International Congress on Imagine and Signal Processing, Bio Medical Engineering and Informatics(CISP-BMEI),2019.

[12] PEI-LIN L, ZHEN-MING Y, WE-NBO T, et al. Medical Knowledge Extraction and Analysis from Electronic Medical Records Using Deep Learning [J]. Chinese medical sciences journal , 2019, 34(2):133-9.

[13] Xiu X, Qian Q, WU S. Construction of a Digestive System Tumor Knowledge Graph Based on Chinese Electronic Medical Records: Development and Usability Study [J]. JMIR Medical Informatics, 2020, 8(10): e18287.

[14] Peng Z, Song H, Zheng X, et al. Construction of hierarchical knowledge graph based on deep learning; proceedings of the 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), F, 2020 [C].

[15] ZHU, LING, GAO, et al. Knowledge graph for TCM health preservation: Design, construction, and applications [J]. Artificial Intelligence in Medicine, 2017, PP:48-52.

[16] Chai X. Diagnosis Method of Thyroid Disease Combining Knowledge Graph and Deep Learning [J]. IEEE Access, 2020, PP (99): 1-1.