Journal of Artificial Intelligence for Medical Sciences
Vol. 3 (1–2), December 2022, pp. 8–15
Journal home: http://oapublishing-jaims.com/jaims
DOI: https://doi.org/10.55578/joaims.221108.001; eISSN: 2666-1470

RESEARCH ARTICLE

A Method of Text Information Normalization of Electronic Medical Records of Traditional Chinese Medicine*

Can Li, Dan Xie†

School of Information Engineering, Hubei University of Chinese Medicine, Wuhan, China

ARTICLE DATA

Article History

Received 17 July 2022
Revised 29 September 2022
Accepted 07 November 2022

Keywords
EMR of TCM
Normalization
Event extraction
Named entity recognition

ABSTRACT

Electronic medical records (EMR) of Traditional Chinese Medicine (TCM) contain rich contents such as chief complaints, subcutaneous symptoms, history of present illness, and past medical history, which are important reference bases for TCM diagnosis. However, there are a lot of terminology and expression irregularities since this information is frequently conveyed in natural language. In this paper, we propose a method to normalize the textual information of EMR of TCM and select the text of medical history with a strong narrative such as the history of present illness and past medical history, as well as the text of symptoms such as chief complaints and subcutaneous symptoms as the main research object. The text is then processed separately according to the type of text. For symptom texts such as chief complaints and subcutaneous symptoms, named entity recognition technology is directly applied to extract symptom entities directly; for medical history texts such as the history of present illness and past medical history, event extraction is performed first to divide the treatment events, and then named entity recognition technology is applied to extract various entities, and finally, the various entities are stored in a database. Using this method, experiments are conducted on the EMR of the orthopedic injury department of a hospital, in which the recognition rate of the symptom entity in the symptom text reaches 92.28%, and the recognition rate of entities such as symptoms and diseases in the medical history text reaches 89.86%. The validity of this method is verified. This method normalizes the natural language writing part of the EMR and stores it in a structured way, which is convenient for the subsequent data analysis and mining, and lays a solid foundation for the intellectualization of TCM.

1. INTRODUCTION

EMR is a medical record of text, symbols, charts, graphs, data, images, and other digital information generated by medical personnel in the course of medical activities using a medical institution's information system and capable of storage, management, transmission, and reproduction [1]. EMR of TCM is a way for TCM doctors to record the diagnosis and treatment process when they treat diseases for patients, mainly including the specific content, treatment, and prescription of TCM "observation, smell, inquiry and cutting" information [2]. It includes information such as symptom text and medical history text, which contains rich knowledge of Chinese medicine and has high value for clinical research. There are many phenomena such as polysemy, homonym, synonym and so on in symptom description [3]. In addition, there are also irregularities in the expression of disease names, examination tests, drugs, time, institutions, departments, and other contents. These irregularities bring many obstacles to the further application of EMR. By using the information extraction technology based on knowledge attribute, the required knowledge attribute information can be obtained from the text information of EMR of TCM, thus as to normalize and structure the textual information of EMR of TCM, which is conducive to the intelligent diagnosis of diseases and lays the foundation for the knowledge engineering of TCM clinical big data [4]. At present, the normalization of EMR texts information at home and abroad mainly focuses on the normalization of symptom information, but in addition to symptom information, other text information such as disease, examination and test, and treatment information also plays an important reference role in the diagnosis of clinical diseases. In this paper, we propose a method to normalize the text information of EMR of TCM, building a corpus with knowledge attributes based on relevant standards, and using a mainstream named entity recognition model to achieve automatic recognition and normalized storage of textual information in EMR of TCM.

*This article submitted to "Artificial Intelligence in Medical Data Governance".

†*Corresponding author. Email: dinaxie@hbtcm.edu.cn*

2. RELATED RESEARCH

With the popularization of informatization in the medical field, EMR, as one of the signs of medical informatization, is the product of medical information system [5]. It replaces paper electronic medical records and electronic medical records in the whole process of hospital diagnosis and treatment and plays an important role in the medical field. However, EMR is mainly expressed in natural language, especially the history of present illness and past medical history, which are more verbalized and narrative. It is difficult for computers to recognize them, which makes data analysis and mining of EMR difficult.

Post-structuring EMR [6] is the process of inputting medical texts in natural language form, conducting structural analysis in strict accordance with professional medical terms and relevant norms of medical records format, and storing them in a database according to a certain structure. Information extraction is a key step in the structuring of EMR. It extracts specific information, such as diseases and symptoms from EMR texts, expresses them in natural language according to a predetermined structure, and forms structured data to help people organize and analyze information in EMR, thus realizing the deep application of EMR data. Therefore, in recent years, more and more scholars have begun to carry out information extraction research on EMR documents, according to a certain structure to extract the required information, in order to achieve the normalization and structuring of EMR.

Hou Weitao et al. [7] extracted medical events and carried out attribute recognition based on the BiLSTM model. Yu Jie et al. [8] proposed the method of joint extraction and realized the joint extraction of medical events. Liu Ziqing et al. [9] proposed a general framework for information extraction of EMR of TCM outpatient clinics, taking clinical manifestations and clinical events as the entry point to extract not only entities such as symptoms and disease names, but also other clinical events such as tumors, surgeries, and treatment results. Liu Kai et al. [10] used the CRF algorithm to study information extraction for content such as chief complaints and history of present illness in the EMR of TCM. Liang Wentong et al. [11] conducted comparison experiments using IDCNN-CRF and BERT-IDCNN-CRF models for medical entities in EMR of TCM and found that the BERT-IDCNN-CRF model has a better recognition effect. Chen Chen et al. [12] used the BERT-BiLSTM-CRF method to recognize named entities such as anatomical sites and diseases in EMR.

3. RESEARCH METHODS

3.1. Corpus Construction Technology

A corpus is a collection of texts. A corpus refers to a structured, representative, large-scale corpus collected specifically as a target for one or more applications. In linguistics, a corpus is a large number of categorized texts with established formats and tags [13].

A high-quality corpus plays a very important role in the automatic extraction of textual information from the EMR of TCM [14]. To construct a corpus, it is necessary to first prepare data, obtain the data to be extracted, then determine the entity categories to be extracted, establish labeling standards according to relevant standards, and finally carry out labeling to obtain the labeled entity corpus.

3.2. Event Extraction Technology

Event Extraction is one of the parts of information extraction. Event Extraction extracts events of interest to the user from text expressed in natural language form and uses them as a structured representation [15]. A complete event contains a trigger word, an event argument, an argument role, and an event type [16].

The common event extraction methods are (1) pattern matching based algorithm: the event extraction method guided by manually or automatically constructed templates of event sentence feature form representation, generally known as pattern matching algorithm. (2) Trigger word method: The trigger word method is also called the event keyword method. In the statistical processing of event sentences, there is a class of cases where more event sentences appear in the text of the sentence, and such cases are basically in the text of the sentence with a certain term or vocabulary. Therefore, it is possible to make event extraction appear better by creating a dictionary of event trigger words [17].

3.3. Named Entity Recognition Technology

Named Entity Recognition (NER) is a fundamental task in Natural Language Processing (NLP) and an important part of information extraction, which aims to recognize and isolate relevant named entities from a large amount of information and annotate them in the text [18]. It is not only an independent information extraction tool but also plays an indispensable role in NLP applications such as information retrieval, question answering system, and knowledge base construction [19].

Currently, there are commonly used named entity recognition models such as BiLSTM-CRF, IDCNN-CRF, BERT-IDCNN-CRF, and BERT-BiLSTM-CRF, which have been widely used in the medical field. In this study, these four models are selected to be trained separately to compare the performance of each model and find the optimal model.

(1) BiLSTM-CRF [20]: this model incorporates a bidirectional long and short-term memory model and a conditional random field model, combines contextual information about words, introduces distributed expressions of words into feature extraction, and maximizes the relationship between words and labels to improve the recognition effect.

(2) IDCNN-CRF [21]: this model combines iterative inflated convolutional neural network and conditional random field model, combines contextual information, adds an inflated distance in the convolutional kernel, and contains multiple inflated convolutional blocks in the neural network, which can take advantage of GPU parallelism to improve the training speed.

(3) BERT-IDCNN-CRF [22]: it consists of the BERT layer, IDCN layer, and CRF layer. The BERT layer is a vector for association extraction of text, the IDCN layer is used to extract features, and the CRF layer is used to block illegal tokens of token sequences to obtain the maximum probability of tokens.

(4) BERT-BiLSTM-CRF [23]: this model introduces an attention mechanism through the BERT model to construct word vectors before performing feature extraction, and the word vectors are constructed through a corpus, which can

improve the accuracy of entity recognition with obscure features and complex composition.

4. MODEL SELECTION

4.1. Evaluation Metrics

In the experiments in this paper, the commonly used metrics for evaluating model performance include precision, recall, and F1-measure. Where TP is the number of entities correctly identified by the model, FP is the number of other classes of entities identified by the model, and FN is the number of other classes of entities not identified by the model [24]. The specific formula is as follows.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{\text{Total True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\text{Total True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{F1} - \text{measure} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

4.2. Parameter Settings

Appropriate parameter settings will result in higher recognition rates when the model is trained [25]. The parameters to be modified in this paper include the learning rate, batch size, and epoch. The learning rate represents the speed of learning, which determines whether the loss function will converge and when local minima will be reached. When the learning rate is set too small, the rate of convergence becomes very slow. If the learning rate is too large, it will cause the gradient to explode and not even converge. The batch sample size, i.e. the number of samples used in 1 iteration, is the number of batch size samples trained together in one iteration. It determines the direction of quantitative gradient descent. If the batch size is too small, the variation between samples will be large and the model will have a hard time converging. If the batch size is too large, the gradient direction will be stable and the accuracy will be reduced. 1 epoch means training once with all the samples in the training set. epoch is not set as large as possible. When the epoch is set too large, an overfitting condition will occur. The epoch is considered appropriate when the difference between the error rates of the test and training sets is small. Therefore, this paper compares the performance metrics of the selected models by modifying the values of these parameters to find the most suitable recognition model for this study.

4.3. Results of Model Selection

1109 diagnostic terminology for diseases and illnesses of orthopedic plus surgery are selected as training data for the selection of recognition models. The BiLSTM-CRF, IDCNN-CRF, and BERT-BiLSTM-CRF models are used to recognize the symptom entities and modify the values of the main parameters. When the BIO labeling method is adopted, when the learning rate is 0.01, the batch size is 16 and max_epoch is 100, the precision, recall, and F1 of the BILSTM-CRF model could reach 91.35%, 71.08%, and 81.20%, respectively. When the learning rate is 0.001, the batch size is 64 and max_epoch is 100, the precision rate, recall rate, and F1 of the IDCNN-CRF model could reach 89.07%, 87.94%, and 86.85%, respectively. When the learning rate is 0.01, the batch size is 64 and max_epoch is 40, the precision, recall, and, F1 of the BERT-IDCNN-CRF model could reach 89.7%, 72.1%, and 78.7%, respectively. When the learning rate is 0.001, the batch size is 32 and max_epoch is 90, the precision, recall, and F1 of the BERT-BILSTM-CRF model could reach 87.71%, 91.14%, and 88.89%, respectively. In summary, as shown in Table 1, the BERT-BiLSTM-CRF model has the best effect on the recognition of symptom entities. Therefore, BERT-BiLSTM-CRF is selected as the recognition model. When the BIOES annotation method is used, as shown in Table 2, the precision, recall, and F1 of the BERT-BiLSTM-CRF model are 80.04%, 87.20%, and 83.47%, respectively, which are lower than those when the BIO annotation method is used, so the BIO annotation method is finally selected for the experiments.

5. EXPERIMENTS

In this paper, a process for text information normalization of EMR of TCM is constructed, as shown in Figure 1. The whole process is divided into the following four stages: (1) data acquisition: obtaining data from the EMR of TCM, then determining the entity types to be recognized, constructing the corresponding entity corpus, and finally selecting the recognition model for the entity extraction stage; (2) event extraction: extracting the treatment events from the medical history text according to the trigger words; (3) entity extraction: add the constructed corpus of entities such as symptoms, diseases, institutions, departments, time, drug treatment, non-drug treatment, surgery names, and tests, etc., and use the selected entity recognition model to recognize and extract the entities from various texts of EMR of TCM; (4) data storage: store the extracted entities in the database. According to the process, an experiment is conducted on the EMR of TCM of a hospital's orthopedic injury department.

5.1. Data Acquisition

5.1.1. Data preparation

767 EMRs of the orthopedic injury department in a Chinese hospital in Henan Province in 2018 are selected as the data foundation of this study. The subcutaneous symptoms and past medical history among them are selected for the study. First, the subcutaneous symptoms and past medical history are screened out

Table 1 Table of experimental results for the four models

Model	Precision (%)	Recall (%)	F1 (%)
IDCNN-CRF	85.27	87.98	86.61
BiLTSM-CRF	**91.35**	73.08	81.20
BERT-IDCNN-CRF	89.7	72.1	78.7
BERT-BiLTSM-CRF	87.71	**91.14**	**88.89**

Table 2 Comparison table of results of labeling methods

Labeling method	Precision (%)	Recall (%)	F1 (%)
BIO	**87.71**	**91.14**	**88.89**
BIOES	80.04	87.20	83.47

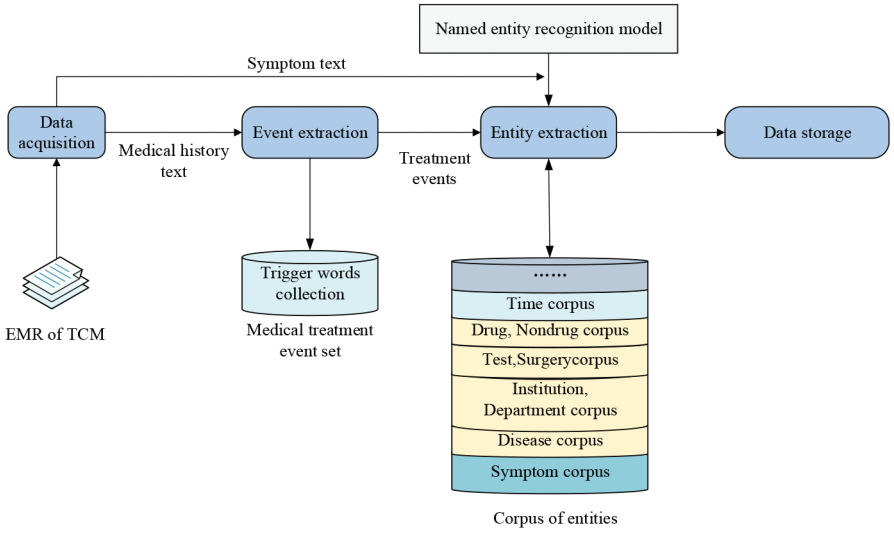

Figure 1 | Schematic diagram of the process of normalization of textual information in EMR of TCM.

from 767 EMR of orthopedic injury, and after "data cleaning", 563 data remained.

5.1.2. Corpus construction

The quality of corpus construction can determine the recognition rate of entity recognition model training, so this subsection details the process of corpus construction as a corpus for named entity recognition model training. The entity corpus constructed in this paper mainly refers to the official documents released by *Classification and Codes of Primary Symptoms in Traditional Chinese Medicine* [26], ICD10/11, and national and provincial health care committees. Since almost all clinical treatment events are covered in the past medical history, this paper constructs a corpus of entities in a total of nine categories, such as symptoms, diseases, time, institutions, departments, drugs, non-drug treatments, surgeries, and tests, based on real-world EMR of TCM, as the basis for model training, of which the symptom corpus is the most important one.

(1) The symptom entity corpus

The symptom entity corpus includes the TCM clinical basic symptom corpus and the specialized disease symptom corpus. *Traditional Chinese Medicine Terminology: Internal medicine, gynecology, and pediatrics (2010)* [27] systematically organizes the diagnostic terminology of TCM specialties, which includes information on symptomatic entities. *Classification and Codes of Primary Symptoms in Traditional Chinese Medicine* standardized the TCM symptom entities and divided the knowledge attributes such as backbone symptoms, acquisition modes, and body parts. They are an important reference basis for the construction of the symptom corpus in this study. The symptom entities and their knowledge attributes are labeled according to the standard labeling mentioned above. For example, the term "osteoarthritis" is labeled as "progressive degeneration of joint cartilage and osteophytes, with {symptom: pain-ask pain; <A: joint> pain}, swelling, local pressure pain, {symptom: limb discomfort-ask discomfort; activity limitation} as the main clinical manifestations. The main clinical manifestations of the disease." Among them, "symptom" is used to indicate the symptom entity, and before and after "-" are the necessary class attributes of the symptom "backbone symptom" and " acquisition mode", after ":" is the symptom entity description, and "<>" is the description of symptom attributes, and the annotation process is described in detail by Xie, W.L. et al. [28]. Due to the large differences in symptom attributes of specialty specialties. Therefore, to better recognize the diagnostic information in specialty-specific EMR, it is also necessary to construct a specialty-specific corpus. Thirty percent (169) of the 563 EMRs are selected for annotation to construct a corpus of specialist-specific symptoms, and the improved corpus contained the previous 1109 articles and 169 articles of specialist-specific diseases, totaling 1278 symptom diagnostic terms.

(2) The disease entity corpus

Disease names refer to ICD-10/11 to construct a corpus of disease entities. The names of common orthopedic diseases in ICD-10/11 are incorporated into the corpus of disease entities. For example, the disease name "osteoarthritis" was labeled as "{disease: osteoarthritis}". The constructed corpus of disease entities contains 510 disease entities.

(3) The time entity corpus

In medical history, there are various types of time descriptions, mostly relative time and absolute time. Relative time: for example, "2 days ago, 1 month later"; absolute time: for example, "January 10, April 22, 2021". For this reason, time entities are divided into simple time expressions, compound time phrases, and time preposition phrases, and simple time is divided into calendar time (Date), specific time (Time), time word (TimeN), periods of time (TimeD) and weekly or weekly time (TimeSet) [29]. For example, the calendar time "2012-06-14" is labeled as "{date: 2012-06-14}", and the time word "the day before yesterday" is labeled as "TimeN: the day before yesterday"; the compound time phrase "March 9, 8:30" is labeled as "Date+Time: March 9, 8:30". The time preposition phrase "1 year ago" is labeled as "{TimeD+TimeLN: 1 year ago}. The constructed corpus of temporal entities contains 26 temporal entities.

(4) The institution and department entity corpus

Since the EMR selected in the experiment are from a Chinese hospital in Henan Province, the names of medical institutions at all levels published by the Henan Provincial Bureau of Statistics and the Henan Provincial Health and Health Commission are selected here to build the institutional corpus. For example, "Yanshi City Hospital of Traditional Chinese Medicine" is labeled as "{institution: Yanshi City Hospital of Traditional Chinese

Medicine}". The corpus of departments is constructed according to the "List of Medical Institutions" issued by the state, for example, "orthopedics" is labeled as "{department: orthopedics}". The constructed corpus of institutions and departments contains 1337 institutions and 34 departments respectively.

(5) The test and surgery entity corpus

For the information on tests and surgery appearing in the medical history, the orthopedic-related surgery are selected from the nationally released "Surgery Classification and Codes" standard (ICD-9-CM3), as well as the types of common orthopedic tests in the Sogou thesaurus. For example, "vertebroplasty" is labeled as "{surgery: vertebroplasty}" and "lumbar spine MRI" is labeled as "{test: lumbar spine MRI}". The corpus of tests and surgery entities constructed contains 25 common orthopedic test entities and 26 surgery entities respectively.

(6) The drug and nondrug entity corpus

Since the experiments are conducted in the EMR of orthopedics, to better recognize orthopedic drugs, the open-source platform Sogou Thesaurus is selected, which contains commonly used orthopedic drugs, and the orthopedic drugs and non-drug treatments are obtained from the annotated data of the EMR to jointly build a corpus of drug names and non-drug treatment entities. For example, "calcitriol capsules" is labeled as "{drug: calcitriol capsules}", and "acupuncture" is labeled as "{nondrug: acupuncture}". The constructed corpus of drug names and non-drug entities contains 105 entities of commonly used orthopedic drug names and 11 entities of non-drug treatments, respectively.

5.2. Event Extraction

Symptom text is not considered in this section because it only describes the symptom part. Instead, the past medical history text contains information about the patient's current and previous visits, examination tests for the disease, surgical history, and treatment. The templates for writing past medical history are different for different medical institutions, different departments, and different types of diseases. Table 3 shows the past medical history of a Chinese medicine hospital in Hubei Province for dropsy and an EMR of a hospital of Chinese medicine in Henan Province for orthopedic injury. As can be seen from Table 3, the two text descriptions are highly different, with the former being well structured and less difficult to extract information, while the latter has a large amount of colloquialism and is difficult to extract. The method in this paper is mainly oriented to the latter, which is a low-structured EMR, and the recognition rate will be higher for the former. Firstly, 30% (169) of the EMR of orthopedic injury are pre-labeled, and the contents of past medical history are analyzed to classify their treatment events.

5.2.1. Treatment events classification

In this paper, by analyzing the content of the past medical history of a Chinese medicine hospital in Henan Province, the treatment events in the EMR of TCM are divided into five sub-categories of patient's symptoms, diagnosis of disease, admission to hospital, and department, testing and examination, and treatment, as shown in Table 4. Among them, treatment events included three sub-categories of events, namely surgery, drug treatment, and non-drug treatment, as shown in Table 5. At least one of the seven sub-events of patient symptoms, disease diagnosis, admission, examination, surgery, drug treatment, and non-drug treatment can be defined as a treatment event, and the combination of several sub-events can also be a treatment event.

Both treatment events and sub-events are composed of multiple argument roles, and nine types of argument roles (corresponding to the nine types of entities) are defined in this paper, and their types are described as follows.

(1) Time: The time when the treatment event occurred. The thesis element is present in each type of event.

(2) Symptom: The manifestation of discomfort, abnormal manifestation caused by the disease.

Table 3 | Table of original information of past medical history in EMR of different medical institutions

Medical institution	Contents
A Chinese hospital in Hubei Province	History of lumbar disc herniation for more than 5 years, history of surgery, history of lumbar disc herniation surgery, history of appendectomy; denied a history of viral hepatitis; denied a history of hypertension; denied a history of diabetes mellitus; denied a history of coronary heart disease; denied a history of trauma; denied a history of blood transfusion.
A Chinese hospital in Henan Province	The patient had intermittent pain in the lumbar back for no apparent reason 3 years ago and was unable to get up. In order to seek further systemic treatment, he visited our department today and checked the lumbar spine BMD: totalBMD0.379T-6.8. He was admitted to our clinic with the diagnosis of "osteoporosis with pathological fracture".

Table 4 | Table of subtypes of treatment events and their descriptions

Sub-event type	Interpretation	Argument roles
patient's symptoms	Events describing the patient's symptoms before and after clinical treatment	time and symptom
diagnosis of disease	An event that determines that a patient has a certain disease	time and disease
admission to hospital and department	It refers to the event that the patient has some kind of discomfort and is admitted to the hospital for consultation	time, institution, and department
testing and examination	Refers to information related to tests and examinations performed on patients	time and test
treatment	Referring to what treatment the patient has taken	time, drug, nondrug, and surgery

(3) Disease: In this paper, it refers to the diagnosis of disease made to the patient.

(4) Institution: In this article, it refers to a medical institution.

(5) Department: In this paper, it refers to the functional departments of the hospital.

(6) Drug: Refers to the drug administered to the patient.

(7) Non-drug: Refers to the non-drug treatment modality performed on the patient. Labeled as nondrug.

(8) Surgery: The type of surgery performed on the patient.

(9) Examination and test: Refer to what kind of examination and test the patient has done.

5.2.2. Treatment event extraction

Since the past medical history contains a lot of information about what symptoms the patient has or has had, where the patient was seen, what tests were done, what treatment was taken, what the surgical history was, and much other information that is numerous and confusing, this subsection will extract the patient's past visits in terms of treatment events.

For example, the example of the past medical history of orthopedic injury in Henan Province in Table 3 is extracted and changed to: "{event: The patient had intermittent pain in the lumbar back for no apparent reason 3 years ago and was unable to get up.}" {event: In order to seek further systemic treatment, he visited our department today and checked the lumbar spine BMD: totalBMD0.379T-6.8}. He was admitted to our clinic with the diagnosis of "osteoporosis with pathological fracture". "A total of 2 treatment events".

In this paper, two approaches are considered: extracting treatment events by matching trigger words and recognizing treatment events by using named entity recognition models. It is found that the former depends on the selection of trigger words and the selection of the raw data to be extracted. When the data is not known in advance, the results of recognizing the diagnosis and treatment events vary depending on the trigger words and the selected past medical history data, and the results of recognizing the diagnosis and treatment events can be as high as 100% in good cases and only about 30% in poor cases. The latter has a recognition rate of 70% when used as named entity recognition, which is not much different from the average result of the former, but the result does not fluctuate greatly due to the difference of data in this way. Thus, a combination of both is considered to extract the treatment events, and the extracted values of precision, recall, and F1 values reach 79.20%, 83.96%, and 81.77%, respectively.

5.3. Entity Extraction

In Subsection 5.2.1, the argument roles of treatment events are the type of entities to be extracted: symptom, disease, time, institution, department, drug, nondrug, surgery, and test nine types of entities. The following will recognize and extract the content in the inscribed symptoms and past medical history respectively.

5.3.1. Entity extraction of symptom text

The entities in symptom text are mainly symptom entities, and symptom entities include necessary and additional attributes, such as backbone symptoms and body parts, etc. The recognition process of knowledge attributes was described in detail by Du Zengzhen et al. [30]. The BERT-BiLSTM-CRF model is used to recognize the symptom entities in orthopedic inscriptions, and a total of 1109 diagnostic terminology for diseases and illnesses of orthopedic plus surgery are used as the base corpus, and five proportions of the orthopedic specialty-specific corpus are randomly selected from 10% (56 terms), 15% (84 terms), 20% (112 terms), 25% (141 terms), and 30% (169 terms), respectively The content of symptom text in the corpus is recognized, and the precision, recall, and F1 values are compared. The training results are shown in Table 6. The highest precision and F1 values, 79.76% and 83.75%, respectively, and the recall rate of 89.36%, are obtained when 25% of the specialist corpus is added.

For the example of subcutaneous symptoms "Since this visit, he was clear, poor mental health, poor sleep, indwelling urinary catheter, and normal stool.", two symptom entities, "poor mental health" and "poor sleep", could be recognized.

Table 5 Table of treatment event subtypes and their descriptions

Sub-event type	Interpretation	Argument roles
surgery	Describe the surgery the patient had at the time of treatment	time and surgery
drug	Describe the medications used during treatment	time and drug
non-drug	Describe the non-pharmacological treatment modalities used by the patient at the time of treatment	time and nondrug

Table 6 Table of recognition results of BERT-BiLSTM-CRF model for symptom text

Training data	Precision (%)	Recall (%)	F1 (%)
Orthopedics + Surgical	72.92	87.81	79.67
Orthopedics + Surgical + 10%EMR	78.10	83.90	80.90
Orthopedics + Surgical + 15%EMR	74.93	83.54	79.00
Orthopedics + Surgical + 20%EMR	78.61	86.97	82.58
Orthopedics + Surgical + 25%EMR	79.76	88.16	83.75
Orthopedics + Surgical + 30%EMR	74.74	89.36	81.40

The trained model using the diagnostic terms of orthopedics plus surgery plus 25% orthopedic corpus is used to predict the symptom entities of the remaining EMR of TCM inscribed with the manual annotation, and the precision, recall, and F1 values are 92.68%, 91.88%, and 92.28%, respectively.

5.3.2. *Entity extraction of medical history text*

The BERT-BiLSTM-CRF model is used to recognize entities such as symptoms and diseases in past medical history, and a total of 1109 diagnostic terminology for diseases and illnesses of orthopedic plus surgery are used as the base corpus and added to a corpus of nine types of entities, and then five proportions of 10%, 15%, 20%, 25%, and 30% of past medical history content in the orthopedic specialty corpus are randomly selected for event extraction, respectively. The treatment events obtained afterward are subjected to entity recognition, and the precision, recall, and F1 value are compared. The training results are shown in Table 7. The recall and F1 values are the highest when 30% of the specialist corpus is added, with 85.94% and 90.48%, respectively, and the precision rate was 88.15%. The entity-specific recognition is shown in Table 8.

For example, for the second event in 4.2.2, the time entity is "today", the department entity is "our department" and "our clinic", the examination entity is "Lumbar spine bone density", and the disease entity is "osteoporosis with pathological fracture".

The prediction of entities in the past medical history of the remaining EMR of TCM using the trained model with the diagnostic terminology of orthopedics plus surgery plus 25% orthopedic corpus is compared with the results of manual annotation, and the precision, recall, and F1 values are 87.14%, 92.44%, and 89.86%, respectively, and the recognition of various types of entities are shown in Table 9.

Table 7 | Table of recognition results of BERT-BiLSTM-CRF model for medical history text

Training data	Precision (%)	Recall (%)	F1 (%)
Orthopedics + Surgical + 10%EMR	81.37	86.8	84.00
Orthopedics + Surgical + 15%EMR	80.91	87.83	84.23
Orthopedics + Surgical + 20%EMR	84.18	89.79	86.89
Orthopedics + Surgical + 25%EMR	84.42	90.28	87.25
Orthopedics + Surgical + 30%EMR	**85.94**	**90.48**	**88.15**

Table 8 | Table of recognition results of BERT-BiLSTM-CRF model for various types of entities with the medical history text

Type of entity	Precision (%)	Recall (%)	F1 (%)
department	98.21	94.48	96.49
disease	89.93	97.81	93.71
drug	90.00	91.53	90.76
nondrug	75.00	85.71	80.00
surgery	78.95	88.24	83.33
symptom	78.13	83.71	80.83
test	84.38	87.10	85.71
time	94.51	92.31	91.14
institution	94.34	94.51	94.51

Table 9 | Table of prediction results of BERT-BiLSTM-CRF model for various types of entities in the medical history text

Type of entity	Precision (%)	Recall (%)	F1 (%)
department	100.00	100.00	100.00
disease	83.63	91.54	87.41
drug	88.24	94.94	91.46
nondrug	83.63	58.82	68.96
surgery	85.71	75.00	80.00
symptom	76.92	82.71	79.71
test	90.05	94.37	92.41
time	85.64	93.82	89.54
institution	94.62	99.19	96.85

5.4. Data Storage

The nine types of entities such as symptoms and diseases extracted in Section 5.3 are stored in an EXCEL table according to a certain structure so that the text information of EMR of TCM expressed in natural language form can be structured and normalized to facilitate further research on EMR of TCM.

6. CONCLUSION

In this paper, we propose a method to normalize the textual information of the EMR of TCM and construct a corpus of nine types of entities such as symptoms and diseases. Using event extraction and named entity recognition techniques, the BERT-BiLSTM-CRF model is selected by comparing several mainstream named entity recognition models. The model has a recognition rate of 92.28% for the symptom entities in the symptom text, and 89.86% for the entities such as symptoms and diseases in the medical history text. Using this method, the textual information of the EMR of TCM can be stored in a structured and normalized form in EXCEL, reducing the preliminary work costs of clinical researchers and laying a solid foundation for the implementation of TCM clinical big data knowledge engineering. In the course of the study, it is found that the quality of the corpus has a significant impact on the recognition rate of entities, so the corpus will be continuously expanded and improved in future studies, and further optimization of the named entity recognition model is also expected to improve the recognition rate of textual information in EMR.

CONFLICT OF INTEREST

The authors do not have any conflict of interest to declare.

AUTHORS' CONTRIBUTION

CL and DX design of the study and analyze the data. CL draft the manuscript and DX revise the manuscript critically for important intellectual content.

ACKNOWLEDGMENTS

This research is partially supported by the the task of Guangdong Provincial Key Laboratory of Traditional Chinese Medicine Information research for the Open fund (No. 2021B1212040007). The authors would also like to thank the Research on inheritance and innovation of TCM syndrome differentiation based on information extraction technology TCM Inheritance and innovation project supported by Hubei University of Traditional Chinese Medicine (No. 2022SZXC012)

REFERENCES

[1] Basic specifications for electronic medical records (Trial), China Phar. 21 (2010), 1063–1064.
[2] Liu Yihui, Ye Hui, Yi Jun, et al., Text information extraction of traditional Chinese medicine electronic medical records based on naive, Modernization of Tradit. Chin. Med. and Materia Medica-World Sci. and Tech. 22 (2020), 3563–3568.
[3] Ren Huiling, Guo Jinjing, Sun HaiXia, et al., Thinking of the study on medical terminology standardization, J. Med. Inform. 39 (2018), 2–7.
[4] Zhang Pan, Shen Shaowu, Tian Shuanggui, et al., Knowledge engineering planning and design for clinical big data in Chinese medicine, Lishizhen Med. and Materia Medica Res. 33 (2022), 764–766.
[5] Ma Siyuan, Cheng Longlong, Huang Shuo, Cui Bingjian, Event extraction of Chinese electronic medical records based on BiGRU-CRF, In 2021 4th International Conference on Artificial Intelligence and Pattern Recognition (AIPR 2021), Association for Computing Machinery, New York, NY, USA, 2021, pp. 592–598.
[6] Cheng Nan, Hou Hao, Niu YaJun, et al., Application of post structured electronic medical record based on NLP technology, Henan Med. Res. 30 (2021), 4510–4513.
[7] Hou Weitao, Ji Donghong, Research on clinic event recognition based Bi-LSTM, Appl. Res. Comp. 35 (2018), 1974–1977.
[8] Yu Jie, Ji Bin, Liu Lei, Li Sha-sha, Ma Jun, Liu Hui-jun, Joint extraction method for Chinese medical events, Comp. Sci. 48 (2021), 287–293.
[9] Liu Ziqing, The extraction of clinical manifestations and clinical events from outpatient electrical medical records of traditional Chinese medicine, Guangzhou University of Chinese Medicine, 2021.
[10] Liu Kai, Zhou Xuezhong, Yu Jian, Zhang Run-shun, Named entity extraction of traditional Chinese medicine medical records based on conditional random field, Comp. Engg. 40 (2014), 312–316.
[11] Liang Wen Tong, Zhu Yanhui, Zhan Fei, et al., Named entity recognition of electronic medical records based on BERT, J. Hunan Univ. Tech. 34 (2020), 54–62.
[12] Chen Chen, Wu Fenlin, Named entity recognition in the electronic medical record based on BERT, Automation & Instrum. 3 (2021), 173–176.
[13] Liu Yibin, Construction and research of Chinese electronic medical record named entity recognition corpus, Guangzhou University of Chinese Medicine, 2020.
[14] Lin Feng, Research on automatic extraction and coding method of TCM clinical symptom information, Hubei University of Chinese Medicine, 2021.
[15] D. Ahn, The stages of event extraction, ARTE '06: Proceedings of the Workshop on Annotating and Reasoning about Time and Events, Association for Computational Linguistics, PA, USA, 2006, pp. 1–8.
[16] Gao Su, TaoHu, Jiang Yanzhao, et al., Sentence-level joint event extraction of traditional Chinese medical literature, Tech. Intell. Engg. 7 (2021), 15–29.
[17] Ma Chunming, Li Xiuhong, Li Zhe, et al., Survey of event extraction, J. Comp. Appl. (2022), pp. 1–20.
[18] Chen Shudong, Ouyang Xiaoye, Overview of named entity recognition technology, Radio Commun. Tech. 46 (2020), 251–260.
[19] Qingchuan Wang, E. Haihong, A BERT-based named entity recognition in Chinese electronic medical record, In Proceedings of the 2020 9th International Conference on Computing and Pattern Recognition (ICCPR 2020), Association for Computing Machinery, New York, NY, USA, 2020, pp. 13–17.
[20] Yang Yanling, Li Yan, Zhong Xinyu, et al., Named entity recognition of TCM medical records based on BiLSTM-CRF, Info. Tradit. Chin. Med. 38 (2021), 15–21.
[21] Li Ni, Guan Huan-mei, Yang Piao, Dong Wen-yong, BERT-IDCNN-CRF for named entity recognition in Chinese, J. Shandong Univ. (Nat. Sci.), 55 (2020), 102–109.
[22] Zhang Qi, Li ChengJun, Liu Jingshu, Research on name entity recognition in military field based on BERT_ IDCNN_ CRF, Aerosp. Electron. Warfare, 37 (2021), 56–60.
[23] Qu Qianqian, Kan Hongxing, Named entity recognition of Chinese medical text based on BERT-BiLSTM-CRF, Electron. Des. Engg. 29 (2021), 40–43+48.
[24] Wang Jun, Wang Xiulai, Luan Weixian, et al., Research on named entity recognition of scientific research talents field based on BERT model, Comp. Tech. Dev. 31 (2021), 21–27.
[25] Zhang Zhifei, Clinical named entity recognition from Chinese electronic medical records using a double-layer annotation model, Nanjing Univ. Posts and Telecommun. 2020.
[26] Yang Fan, Deng Wenping, Sun Jing, et al., Classification and codes of primary symptoms in traditional Chinese medicine, China Information Association for Traditional Chinese Medicine and Pharmacy, Beijing, 2019.
[27] Published by the National Committee for the Validation of Scientific and Technical Terms, Traditional Chinese Medicine Terminology: Internal medicine, gynecology, and pediatrics (2010), Science Press, Beijing, 2011.
[28] Xie WenLi, Mao ShuSong, Xie Dan, Corpus construction for TCM clinical symptom based on information coding standard, In Proceedings of the 2nd International Symposium on Artificial Intelligence for Medicine Sciences (ISAIMS 2021), Association for Computing Machinery, New York, NY, USA, 2021, pp. 493–499.
[29] Zhang Chunju, Zhang Xueying, Li Ming, et al., Interpretation of temporal information in Chinese text, Geogr. Geo-Inf. Sci. 30 (2014), 1–7.
[30] Z. Du, D. Tang, D. Xie, Automatic extraction of clinical symptoms in traditional Chinese medicine for electronic medical records, IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, Houston, TX, USA, 2021, pp. 3784–3790.