Research Article

Relation Extraction Based on Fusion Dependency Parsing from Chinese EMRs

Pengjun Zhai,¹ Xin Huang ,^{1,2} Beibei Zhang,¹ and Yu Fang ¹

¹Department of Computer Science and Technology, Tongji University, Shanghai 200082, China

²School of Software, Jiangxi Normal University, Nanchang 330047, China

Correspondence should be addressed to Yu Fang; fangyu@tongji.edu.cn

Received 8 February 2020; Accepted 20 April 2020; Published 8 June 2020

Academic Editor: Chenxi Huang

Copyright © 2020 Pengjun Zhai et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The Electronic Medical Record (EMR) contains a great deal of medical knowledge related to patients, which has been widely used in the construction of medical knowledge graphs. Previous studies mainly focus on the features based on surface semantics of EMRs for relation extraction, such as contextual feature, but the features of sentence structure in Chinese EMRs have been neglected. In this paper, a fusion dependency parsing-based relation extraction method is proposed. Specifically, this paper extends basic features with medical record feature and indicator feature that are applicable to Chinese EMRs. Furthermore, dependency syntactic features are introduced to analyse the dependency structure of sentences. Finally, the F1 value of relation extraction based on extended features is 4.87% higher than that of relation extraction based on basic features. And compared with the former, the F1 value of relation extraction based on fusion dependency parsing is increased by 4.39%. The results of experiments performed on a Chinese EMR data set show that the extended features and dependency parsing all contribute to the relation extraction.

1. Introduction

Electronic Medical Record (EMR) contains a vast of medical entities that provide rich medical knowledge. It is worth noting that there are certain interdependent relations between entities rather than isolated ones, which truly reflects the medical knowledge and the judgment and application of medical knowledge by doctors. The relations between entities in EMPs represent the health of patients from different perspectives. Relation extraction plays a fundamental role in medical knowledge graph (MKG) construction and completion and supports many other tasks, such as question answering, semantic understanding of texts, and recommender systems.

Entity relation in EMRs mainly includes the relation between treatment and disease, treatment and symptom, test and disease, test and symptom, and disease and symptom. At present, the machine learning method is widely used in the field of medical texts [1–4], including the task of relation extraction of English EMRs [5], and most of the feature

selections rely on English medical dictionaries and data sets [6] as well as syntactic analysis [7]. However, the relation extraction of Chinese EMRs is still scarce, which is reflected in two aspects: one is the relation between two specific entities and the other is neglecting the unique features of Chinese EMR texts and sentences.

To cope with the above shortcomings, we proposed a fusion dependency analysis method for relation extract of Chinese EMRs. The underlying idea is to extend features according to the unique features of Chinese EMRs, such as medical records feature, indicators feature, and extended context feature. Considering that the entity relations in two sentences with similar structure and context are often the same and the structural similarity of sentences in Chinese EMR is high, the sentence structure information is fused based on the feature extension. Among the methods of machine learning, some research studies [8, 9] have verified that SVM is a better method for entity relation extract; thus, this paper directly adopts SVM to train the model and predict.

2. Related Work

The concept of relation extraction was first put forward at the Message Understanding Conference (MUC) and supported by the Defense Advanced Research Projects Agency (DARPA) at the end of the 1980s. After that, the Automatic Content Extraction Conference (ACE) promoted the development of relation extraction technologies. Recently, the development of knowledge graph (KG) once again emphasizes the importance of relation extraction.

2.1. Relation Extraction of English EMRs. The relation extraction methods of EMRs are evolved from the early methods based on rules and dictionaries to the current classification based on machine learning, where entity relation refers to the relation between entity pairs appearing in a sentence. For the relation extraction of English EMRs, an SVM model [10] was utilized to identify the relationships among disease, symptom, test, and treatment. In this research, semantic lexical features, the order of entity pairs appearing in sentences, and syntactic features have been added as classifier and present an SR classifier, which can recognize 84% of the relations in the BIDMC corpus and achieve microaveraged F-measures of 0.89. A model was described in a study [11] to identify the semantic relations among medical concepts, including problems, tests, and treatments, from the medical texts and to analyse three types of relations which are the relation between treatment and problem, test and problem, and problem and problem. To extract the above relations, a hybrid method was proposed based on machine learning, dictionary, and rules [12]. In the view of the I2B2 (Informatics for Integrating Biology and the Bedside) 2010 (<https://www.i2b2.org/NLP/Relations/>), Rink [6] used GENIA15 to pre-process the medical record texts, and then selected the context similarity as the new feature based on the lexical feature and context feature. The task of feature extraction used knowledge bases such as Wikipedia, WordNet and general inquirer [13]. This model also uses the SVM model to achieve the F-measures of 0.74. The relations between concepts in UMLS were used as a substitute feature to solve the problem that some entities in EMRs do not have rich context features [14], and the experimental results obtained an F-measures of 0.67.

2.2. Relation Extraction of Chinese Text. At present, the research studies of relation extraction in Chinese mainly focus on the open domain and the methods of relation extraction in Chinese EMRs are still in the preliminary stage. A pipeline of NLP techniques was employed [15], a.k.a., word segmentation, POS-tagging, and syntactic parsing, to extract entity relations for an open domain. This system was considered as the first attempt to handle Chinese open relation extraction. In the medical field, the dependency graph was used to automatically learn the syntactic pattern of relation extraction and extracted the relation between disease and symptom by this model [16]. Also, a rule-based method was used to extract medical information for unstructured text data in EMRs [17]. The bootstrapping

framework based on semisupervision was proposed in a study [18], combined TCM bibliographic literature database in China and MEDLINE (<https://jgc128.github.io/mednli/>), to discover the knowledge of gene functional including extracting the relation between symptom and gene, symptom and disease, and disease and gene. According to the characteristics of the relation between entities in the EMRs, a semisupervision learning method was used [19], SVM was adopted as the classifier to predict the labeled samples combined with auxiliary classification information, and then the classification after adding the samples with low confidence to the training set was repeated, which shows that the entity relation can be extracted effectively by the method of classifying and calculating entity co-occurrence.

3. Methods

In this section, we first introduce the preprocessing method of Chinese EMR data. Second, we briefly describe the basic features for relation extraction of Chinese EMRs. And we extend the features based on the basic features, according to the characteristics of Chinese EMR texts. Finally, by fusing sentence structure information, a method of relation extraction based on dependency parsing is proposed. The relation extraction process is shown in Figure 1.

3.1. Data Preprocessing. The data set used in this paper for the research of entity relation extraction comes from XML EMR texts that were preprocessed initially and the files of entity and entity relation that have been tagged from EMR by a semiautomatic annotating method, which is described in Section 4.1. Among them, discharge summaries and progress notes [20] are selected as the Chinese EMR texts. The discharge summary includes the basic information of the patient at the time of admission, the diagnosis of the doctor, the tests and treatments received in the process of hospitalization, the basic information and the doctor's advices at the time of discharge, and the final treatment results. The details of discharge summaries are shown in Figure 2. The process note mainly records the clinical manifestations of the patients during hospitalization and the medical behaviours such as test and treatment received.

The process of data preprocessing is roughly divided into three parts. First, the EMR texts should be segmented by using “。” “,” “;” and “\n” as the boundary of sentences. Then, entity pairs need to be identified from the EMR texts. Finally, the EMR texts that have completed sentence segmentation are tagged with word segmentation and part-of-speech, with the help of NLPir (<http://ictclas.nlpir.org/>) that is a word segmentation tool.

3.2. Relation Types. Relation extraction is used to find the relation between entities from the text, while the relation extraction of the EMRs entity mainly studies the relation between entities such as disease, symptom, test, and treatment recognized from the EMRs. These entity relations reflect the health information of patients and medical treatment measures for patients, as well as the professional

FIGURE 1: Relation extraction process.

Name	Section office : hepatobiliary department No.1 of general surgery (inpatients)	Inpatient ward :	Bed number :	Admission number :
Name	Gender : male	Age : 68	Radiation number :	
Admission time	28/02/2015	Discharge dates : 01/03/2015		
Condition of hospitalization	The patient came to the hospital because of "The mass of right neck has existed for 5 years and increased significantly in more than 2 months" after physical tests, it was found that there was a 0.6*0.6 cm bag on the right neck. No redness or swelling,tender, no skin rash, no ulceration			
Admitting diagnosis	Right cervical mass			
Major laboratory tests and special examination	Null			
Diagnosis and treatment process	After admission, relevant tests were carried out, and the right neck mass was removed under local anesthesia immediately. The operation was smooth and the patient returned safely. Now the patient is in a stable condition and can be discharged from the hospital and become an outpatient			
Pathological results	Come to the hospital ten days later to see the pathology reports			
Discharge diagnosis	Right neck sebaceous cyst			
Hospital discharge	The patient has no discomfort			
Discharge orders	1 Follow-up, Monday, Chief physician shi baomin ward round and records. 2 Health education, five days after operation have stitches out 3 Rehabilitation guidance: avoid touching water			
Treatment results	Cured			
Attending surgeon:			*	Residents: *

FIGURE 2: Discharge summary.

knowledge of doctors. For the first time, the assessment task of I2B2 2010 systematically classifies the entity relation of EMRs, including the relation between medical problem and medical problem, medical problem and test, and medical problem and treatment. According to the characteristics of Chinese EMR texts, this paper divides the medical problem in I2B2 2010 into two categories as disease and symptom and then redefines the relation between medical entities as the relation between treatment and symptom, treatment and

symptom, test and disease, test and symptom, and disease and symptom. The specific definitions are shown in Table 1.

3.3. Basic Features. Features play an irreplaceable role in the task of relation extraction, especially for Chinese EMR texts. This paper first introduces the basic features of entity relation extraction for Chinese EMRs and references the features of open-domain text for relation extraction,

TABLE 1: Entity relation type and description.

Relation type	Relation representation	Representation description
Relation between treatment and disease	TrID	Treatment improves disease
	TrWD	Treatment worsens disease
	TrCD	Treatment causes disease
	TrAD	Treatment applied to disease
	TrNAD	Due to disease, not adopting treatment
Relation between treatment and symptom	TrIS	Treatment improves symptom
	TrWS	Treatment worsens symptom
	TrCS	Treatment causes symptom
	TrAS	Treatment applied to symptom
	TrNAS	Due to symptom, not adopting treatment
Relation between test and disease	TeRD	Test confirms disease
	TeCD	For confirming disease, adopt test
Relation between test and symptom	TeRS	Test discovers symptom
	TeAS	Due to symptoms adopt test
Relation between disease and symptom	DCS	Disease causes symptom
	SID	Symptom indicates disease

which are mainly divided into lexical feature, contextual feature, entity feature, and location feature.

- (1) Lexical: this involves the two entities themselves, which play a certain role in the relation extraction between them, because even if two specific entities appear in different places, the relation between them may be the same. For instance, the relation between “感冒 (cold)” and “发烧 (fever)” in “患者因感冒而发烧 (patients have fever due to cold)” is usually “DCS (disease causes symptoms)”, so this paper also takes the two entities themselves as a feature.
- (2) Contextual: in Chinese EMR texts, the bag-of-words and part-of-speech in a certain range before and after two entities play a key role in the extraction of the relation between the two entities. The entity relation is judged by the context information, which refers to three bag-of-words and part-of-speech before and after two entities in this paper.
- (3) Entity: the entity feature refers to the type of entity, which is an extremely important feature because the entity relation in this paper is classified by the two types of entity. Among them, the entities of test and treatment type only have relations with two types of entities that are disease and symptom, and there is a relation between disease and symptom instead of the relation between test and treatment. This feature has important guiding significance for the boundary judgment and specific type of judgment of entity relation.
- (4) Relative position: the relative position of two entities, E1 and E2, has a certain indicative function for entity relation extraction in a sentence of Chinese EMR texts. For most sentences in the Chinese EMR data set of this paper, the disease entities and symptom entities appear in front of test entities and treatment entities, while the disease entities generally appear in front of symptom entities. For example, the disease entity “胆结石 (gallstone)” is in front of the

treatment entities “全胆囊切除术 (total cholecystectomy)” in the Chinese EMR text “1974年因胆结石于瑞金医院行全胆囊切除术 (in 1974, total cholecystectomy was performed in Ruijin hospital due to gallstones),” and the relation between two entities is “TrAD (treatment applied to disease)”. There are four categories of relative positions of two entities in this paper: E1 is on the left of E2, E1 is on the right of E2, E1 is in E2, and E2 is in E1.

- (5) Distance: the distance between two entities refers to the number of words between them. In general, the more words there are between two entities, the farther apart they are, and the less likely there is a relation between them. The distance between two entities is expressed by measuring the numbers of words between two entities after word segmentation, in which words contain punctuation marks.

3.4. Extended Features. In order to achieve the task of extracting entity relation of Chinese EMRs more accurately, after analysing the texts of Chinese EMR, this paper extends the features of EMRs based on the basic features that are named extended features, which are mainly divided into medical record features, indicator features, and extended context features.

- (1) Medical record: the chapter in which the entity located has a certain effect on entity relation extraction of Chinese EMRs. For example, in the “出院情况 (discharge situation)” chapter of discharge summary in Chinese EMRs, the probability of relation related to improvement is higher than that related to worsening. In addition, the modification information of an entity is also unique information in EMRs, which is a description of the entity. To sum up, the medical record features refer to chapters and modifications of entities.
- (2) Indicator: the mapping of entity context words and the indicator word base for entity relation are

regarded as indicator features in Chinese MERs. According to the characteristics of Chinese MERs, the judgment of entity relation is related to the context words of two entities. There are some indicators that can directly classify the relation between two entities. If there are indicators such as “好转(improved),” “有所缓解(relieved),” “明显好转(obviously improved),” and “控制稳定(stable control),” the entity relation is generally “TrID” or “TrIS”. If there are indicators such as “控制不佳(poor control),” “效果一般(general effect),” “未见明显变化(no obvious change),” the entity relation is generally “TrWD” or “TrWS.” After analysis and statistics, the indicator word base of all entity relation is established, and the mapping of the two entity’s words in the indicator word base is regarded as an extended feature.

- (3) Extended context: in a sentence, there are many entities juxtaposition, which makes it impossible to find the words with indicative meaning to the relation in the entity context. For instance, in the sentence “患者7天前无明显诱因出现腹胀 (the patient had no obvious inducement to develop abdominal distention), 伴右上腹钝痛 (accompanied by dull pain in the right upper abdomen), 伴乏力, 食欲减退 (fatigue and anorexia), 伴皮肤, 巩膜黄染 (yellow staining of skin and sclera), 无腹泻, 黑便 (no diarrhea or black stool), 无寒战 (no shivering), 高热 (fever), 恶心 (nausea), and 呕吐 (vomiting),” there are many words which are useless for relation extraction in a certain range near the entity. Therefore, this paper extends the general context feature and selects verbs near the entity as the extended context feature.

3.5. Dependency Parsing. Most of the Chinese EMRs are long sentences, and the content and form of the sentences are relatively patterned, especially the structure of sentences that are mostly similar. Therefore, it is worth adding the structure information of sentences to the task of entity relation extraction from Chinese EMRs. Dependency parsing reveals the syntactic structure of a sentence by analysing the dependency among its components. In a word, it is to recognize the grammatical components such as “subject predicate object” and “attributive adverbial complement” and analyse the relationships between them. It claims that the dominator of a sentence is the core verb [21] and that all the dominators depend on the core verb in one way or another.

The language technology platform (<http://ltp.ai/>) (LTP) of the Harbin University of Technology is a complete set of Chinese language processing system developed by the social computing and information retrieval research center of the Harbin University of Technology. It provides rich, efficient, and accurate natural language processing technologies, including Chinese word segmentation, part-of-speech tagging, dependency parsing, and semantic role tagging. Using the LTP to analyse the dependency of the sentence “the patient

having symptoms of wheezing and fever was given anti-infection treatment and relieved after the treatment of antiasthmatic (患者出现喘息, 伴发热, 予抗感染, 平喘治疗后缓解).” The results are shown in Figure 3.

Dependency parsing is to analyse the structural information of a sentence, recognize the “subject predicate object” and “attributive adverbial complement,” and analyse the relationships between the components. According to the dependency parsing of example sentences in Figure 3, the core predicate of the sentence is “出现 (has),” the dependency of entity “喘息(wheezing)” and “出现(has)” is VOB, and the dependency of the entity “抗感染(anti-infection)” is VOB as well. Table 2 shows the annotation relation obtained from dependency parsing by LTP.

3.6. Dependency Syntactic Features. In this section, sentence structure and features will be integrated to get dependency syntactic features for better mining syntactic construction and semantic features, where the sentence structure is reflected in dependency parsing and sentence similarity calculation by using the algorithm of edit distance. The specific dependency syntactic features are defined as follows:

- (1) Sentence dependency relation of binary entities: this is referred to the syntactic relations between two entities in the syntactic structure of a sentence after dependency parsing. For instance, the dependency relation of entity “喘息 (wheezing)” after parsing is VOB in the above example (Figure 3) and the dependency relation of the entity “发热 (fever)” is COO. Therefore, this paper takes the dependency parsing value of each entity in the entity pairs as a feature.
- (2) Dependency relation combination of entity pair: the last feature is to take the dependency relation of entity pair as a feature input, while this feature refers to the dependency relation combination of entity pair, which is sequential. Because of this sequential, the syntactic structure of entity pairs in sentences can be shown more clearly by analysing the combinatorial feature than by analysing the independent dependency relation feature. For example, the dependency relation of entity pair <喘息(wheezing), 抗感染(anti-infection)> in the above example (Figure 3) is VOB-VOB, indicating that both entities act as an object in VOB. Different types of relationships have different dependency relation combinations, so this dependency syntactic feature can better reflect the differences of relation types between different entities.
- (3) The distance between a binary entity and core predicate: after a lot of research studies and experiments on dependency parsing, it is found that the core predicate plays an important role in the extraction of entity boundary and entity relations. In a sentence, the distance between the entity and the core predicate is obviously different from that between the entity and the common predicate, so this paper takes the former as a feature. After the core

FIGURE 3: Dependency parsing of example sentences.

TABLE 2: Tag set of dependency parsing.

Labels	Relation types	Description
ADV	Adverbial-centered relation	Adverbial
ATT	Attribute-head relation	Attribute
COO	Coordinate relation	Coordinate
HED	Head relation	Head
SBV	Subject-verb relation	Subject-verb
VOB	Verb-object relation	Verb-object
WP	Punctuation	Punctuation

predicate of a sentence is obtained by dependency parsing, the distance between the entity and the core predicate is calculated by calculating the number of words between them based on the location of the core predicate.

3.7. SVM Model. The objective of the support vector machine model [22] is to find a hyperplane in an N-dimensional space (N is the number of features) that distinctly classifies the data points. To separate the two classes of data points, there are many possible hyperplanes that could be chosen. In order to find a plane that has the maximum margin, i.e., the maximum distance between data points of both classes, we turn it into a convex quadratic programming problem.

Given a training sample set $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, where $x_i \in \mathbb{R}^n$ is the i th feature vector and y_i is the label of classes, denoted as $y_i \in \{+1, -1\}$, $i = 1, 2, \dots, n$, the hyperplane is defined as follows:

$$w^T x + b = 0, \# \quad (1)$$

where $w = (w_1, w_2, \dots, w_d)$ is the normal vector of the hyperplane and defines the direction of the hyperplane and b is the intercept that determines the distance between the hyperplane and the origin. Due to the correctness of classification being judged by observing whether $w^T x + b$ and y are both positive or negative numbers, a function of margin γ' should be defined as follows:

$$\gamma' = y(w^T x + b), \# \quad (2)$$

In order to unify the measurement, constraints are added to the normal vector w :

$$\gamma = \frac{y(w^T x + b)}{\|w\|_2} = \frac{\gamma'}{\|w\|_2}, \# \quad (3)$$

The idea of the SVM is to maximize the margin, so that the distance from all points to the hyperplane is greater than or equal to a certain distance; then, all classification points are classified on both sides of the support vector, i.e.,

$$\max \gamma = \frac{y(w^T x + b)}{\|w\|_2}, \quad (4)$$

$$s.t. \ y_i(w^T x_i + b) = \gamma'^{(i)} \geq \gamma', \quad i = 1, 2, \dots, n, \#.$$

If the function of margin $\gamma' = 1$, then equation (4) is reduced to

$$\max \frac{1}{\|w\|_2}, \quad (5)$$

$$s.t. \ y_i(w^T x_i + b) \geq 1, \quad i = 1, 2, \dots, n, \#.$$

Considering that maximizing the $1/\|w\|_2$ is equal to minimizing the $1/2\|w\|_2$, the SVM model for solving the maximum partition hyperplane problem can be expressed as the following constrained optimization problem:

$$\min \frac{1}{2}\|w\|_2, \quad (6)$$

$$s.t. \ y_i(w^T x_i + b) \geq 1, \quad i = 1, 2, \dots, n, \#.$$

4. Results

In this section, we carry out three comparative experiments based on basic features, extended features, and dependency syntactic features. The experimental results show that structural information is very important for entity relation extract of Chinese EMRs, an irreplaceable role in the task of relation extraction, especially for Chinese EMR texts.

4.1. Data Set. We evaluate our approach of entity relation extraction on the medical dataset from the existing research [23]; this dataset is semiautomatic and annotated from Chinese EMRs of a grade-three general hospital in Shanghai for a whole year, and the entity set is obtained through the method of feature-enhanced entity recognition. The detailed information of the data set is shown in Table 3. We use 70% of the dataset as training data and 30% for testing. For readers interested in this data set, it is recommended to read academic study [23].

4.2. Baseline. In this paper, the task of entity relation extraction can be transformed into a multiclassification problem. The machine learning tool of LibSVM [24] is used to automatically build multiple binary classifiers according to the number of categories, which can be directly used for

TABLE 3: Entity instance distribution.

Entity types	Summary of discharge	First disease process	Total
Disease	905	1519	2424
Symptom	1407	2225	3632
Test	599	986	1585
Treatment	1045	1264	2309
Total	3956	5994	9950

multivalued classification. Therefore, this paper uses a LibSVM tool to train and test the SVM model, which has certain requirements for the data format of training and test data set, and the data format of input files is shown in Figure 4. Each row of data in Figure 4 represents a training vector, and the 'label' represents the identification of each classification label in this multiclassification, the 'index' is the number of features, and the 'value' is the value of features. In this paper, all data sets trained and tested by LibSVM are transformed into data files of this format for experiments after feature extraction and feature vector construction.

In order to compare the effects of extended features and sentence structure information on the experimental results of entity relation extraction in Chinese EMRs, three contrast experiments are set up in this paper. The first experiment is the baseline experiment, which selects the basic features including lexical feature, contextual feature, entity feature, and location feature. And the second experiment adds the extended features based on the basic features, while the last experiment adds the dependency parsing to the features to form the dependency syntactic features. The results of experiments are evaluated by 3 types of indicators [25]: Precision (P), Recall (R), and F1.

4.3. Results and Analysis. The experimental results of relation extraction based on different features for the data set are shown in Table 4. As we expected, the method of fusing dependency parsing outperforms the relation extraction method based on basic features or extended features. For the baseline, the extraction effect of entity relations of TrCD, TrNAD, and TrNAS is poor. This is because these three types of relationships appear less frequently (less than 5 times) in the tagging corpus. While the precision of TeRS is high, not only because this relation type appears more frequently in the training corpus but also because the characteristics of this relation are obvious, in which sentence pattern is basically “胸片示 (chest X-ray shows): 双肺纹理增多 (bilateral lung marking are increased), and 模糊 (blurred)”. In addition, the extraction effect of SID and TeRD is better, which is also due to the obvious surface features and more training data. However, the relation extraction precision of TrID, TrIS, TrWd, and TrWS is low because of the existence of long sentences in Chinese EMRs, and only the contextual features of before and after words are not obvious.

While after adding the extended features proposed in this paper, in addition to the three unextracted relation types of TrCD, TrNAD, and TrNAS, the precision and

<label>	<index1>:<value1>	...	<index n>:<value n>	...
...	...	...	...	...
<label2>	<index1>:<value1>	...	<index n>:<value n>	...
...	...	...	...	...

FIGURE 4: Data format of training and test data set.

recall rate of all other relation types have been improved, among which the improvement effect of the four types is stronger: TrID, TrWd, TrIS, and TrWS. This is because the medical record features in the extended features (including chapter information and entity modification information) have some influence on the location of entity relation. For example, in the chapter of “出院情况 (discharge situation),” the incidence of entity relation related to improvement is higher than that related to worsening. The indicator features in extended features are more effective for the relation types of improving and worsening because there are related demonstratives (好转 (improvement), 稳定 (stability), 一般 (general), 不佳 (poor), etc.) before and after the entities of improving and worsening. The verb features of entity to the front and back in the extended features are also of great significance to the entity relation extraction of Chinese EMRs. Due to the long sentence in the texts of Chinese EMR, the words before and after many entity pairs are meaningless for entity relation extraction, while the verbs before and after entity pairs generally have certain indicative meanings.

As shown in Table 4, the precision and recall rate of all entity relations have been significantly improved after adding the dependency syntactic features. Dependency parsing is mainly to mine deeper structure information of sentences based on the surface semantic features. Obviously, the three dependency syntactic features added in this paper still greatly improve the precision rate of TrID, TrWd, TrIS, and TrWS, as well as the improvement effect on the TeRD and TeRS. It is because the sentence patterns of “treatment discover symptoms” and “treatment confirmed diseases” are very similar and unified. Many sentences are the patterns of “a certain test: symptom description or disease description” or “test shows: symptom description or disease description,” so this characteristic can be mined by the dependency parsing.

The values of F1 for the relation extraction based on different features show the trend of the effects of entity relation extraction. In the case of limited training corpus, the performance of each entity relation is improved after fusing extended features and dependency syntactic features. Particularly, it is more effective for the several relation types (TrID, TrWd, etc.) that are relatively few in the corpus. However, our method is not very effective in the extraction of three types of relation TrCD, TrNAD, and TrNAS, because the number of these three relation types in the corpus is too small. The future research direction can be focused on how to generate relevant corpus or mine deeper features when the number of the corpus is small.

TABLE 4: Performance on relation extraction for different features.

Relation types	BFs			BFs + EFs			BFs + EFs + DSFs ¹		
	<i>P</i>	<i>R</i>	<i>F1</i>	<i>P</i>	<i>R</i>	<i>F1</i>	<i>P</i>	<i>R</i>	<i>F1</i>
TrID	52.76	61.67	56.87	59.75	68.32	63.75	64.34	72.09	67.99
TrWD	49.31	51.31	50.29	56.79	60.57	58.62	61.09	65.83	63.37
TrCD	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TrAD	68.57	64.97	66.72	71.19	72.76	71.97	76.93	77.93	77.43
TrNAD	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TrIS	59.13	68.12	63.31	64.41	71.43	67.74	68.01	76.83	72.15
TrWS	51.97	59.32	55.40	55.46	62.82	58.91	62.24	67.42	64.73
TrCS	69.13	73.97	71.47	72.73	86.72	79.11	77.03	90.43	83.19
TrAS	58.34	61.45	59.85	62.61	64.86	63.72	66.78	68.92	67.83
TrNAS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TeRD	73.21	77.61	75.35	77.89	80.75	79.29	82.89	84.79	83.83
TeCD	59.32	61.75	60.51	63.09	64.09	63.59	68.98	69.76	69.37
TeRS	81.78	89.31	85.38	84.98	93.73	89.14	86.92	92.67	89.70
TeAS	61.79	62.45	62.12	65.89	67.63	66.75	71.89	72.83	72.36
DCS	58.71	62.75	60.66	61.72	68.53	64.95	65.27	72.46	68.68
SID	75.82	78.97	77.36	80.62	81.41	81.01	84.81	85.42	85.11
Total	63.07	67.20	65.07	67.47	72.59	69.94	72.09	76.72	74.33

BFs represent basic features, extended features are represented as EFs, and DSFs are used to represent dependency syntactic features.

5. Conclusions

This paper implements the extraction of entity relations in Chinese EMRs. The relation types of extraction include the relations between treatment and disease, treatment and symptom, test and disease, test and symptom, and disease and symptom. And the machine learning method is used to transform the task of relation extraction into the classification of entity pairs, which mainly uses the SVM model for training and testing. The similarity of sentences brings a lot of hints to entity relation, i.e., generally, the relation between two entities in sentences with similar sentence structures and semantics is the same. First, this paper proposes four basic features of general text, such as lexical feature and location feature. Second, due to the juxtaposition of many entities or words in Chinese EMR texts, the simple context information is redundant and noisy, so the extended feature is proposed, which is composed of chapter information and indicator feature. In addition, because the basic features and extended features are the only superficial semantic features, but ignoring the information of sentence structure, LTP tool is used to analyse the dependency parsing of Chinese EMR texts and introduce the dependency syntactic features. In this paper, an SVM model is adopted to train and test entity relation extraction. Three comparative experiments are designed for the above three types of features. The results show that the extended features and dependency syntactic features proposed in this paper improve the accuracy and recall rate of entity relation extraction of Chinese EMRs to a certain extent. However, the training set and test set used in this paper are limited in scale. In the future, it is necessary to study the deep learning method for a largescale corpus to extract entity relations more efficiently.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

The authors would like to thank the hospital for its contribution, which provides electronic medical records that are used as the data set for the experiments in this paper. This research was funded by the National Key Research and Development Program of China (No. 2019YFB2101600).

Supplementary Materials

Due to the privacy of the data set of our medical EMRs, we selected some experimental data as samples. The details of the supplementary materials file are as follows: (1) the discharge folder is the data of discharge summary, which includes the data of training, test, discharge summary relation, and discharge summary entity. The files in the folders of train and test are the discharge summaries of patients, which are used to train and test models including condition of hospitalization, admitting diagnosis, diagnosis and treatment process, discharge diagnosis, hospital discharge, and discharge orders. The files in the folders named dischargeEntity are the medical entities of discharge summaries. Every line in the files corresponds with entity information in discharge summaries, which are tagged with “ $C = \text{entity } P = \text{start: end } T = \text{entity type } A = \text{entity assertion}$,” where C represents the concepts of entities in discharge summaries, P means the start and end position of entities in medical EMR texts, and T and A stands the type of entities and the modification of entities, respectively. The files in the folders named dischargeRelation are the entity relations of discharge summaries. Every line in the files corresponds with the relation between entities in medical discharge summaries, which are tagged with $E = \{\text{entity}[\text{start-end}]\text{entity type}; \dots\} \| R = \| E = \{\text{entity}[\text{start-end}]\text{entity type}; \dots\}$, where the first E represents the first entity, including the entity concept, the start-end position, and type of entity. Similarly, the second E represents the second entity. And the middle R represents the relation type between the two entities. (2) The progress

folder is the data of progress record, which includes the data of training, test, progress record relation, and progress record entity. The files in the folders of train and test are the progress records of patients, which are used to train and test models including characteristics of case, preliminary diagnosis, and plan of diagnosis. The files in the folders named progress Entity are the medical entities of progress records. Every line in the files corresponds with entity information in progress records, which are tagged with “ $C = \text{entity } P = \text{start}; \text{end } T = \text{entity type } A = \text{entity assertion}$,” where C represents the concepts of entities in progress records, P means the start and end position of entities in medical progress records, T and A stands the type of entities and the modification of entities respectively. The files in the folders named progressRelation are the entity relations of progress records. Every line in the files corresponds with the relation between entities in progress records, which are tagged with $E = \{\text{entity} [\text{strat-end}] \text{entity type}; \dots\} \| R = \| E = \{\text{entity} [\text{strat-end}] \text{entity type}; \dots\}$, where the first E represents the first entity, including the entity concept, the start-end position, and type of entity. Similarly, the second E represents the second entity. And the middle R represents the relation type between the two entities. (*Supplementary Materials*)

References

- [1] S. Gupta and A. K. Manjhar, “Relation classification from unstructured medical text using feature based machine learning approach,” in *Proceedings of the International Conference on Trends in Electronics and Informatics, ICEI*, Tirunelveli, Tamilnadu, pp. 1135–1138, May 2017.
- [2] P.-H. Chen, H. Zafar, M. Galperin-Aizenberg, and T. Cook, “Integrating natural language processing and machine learning algorithms to categorize oncologic response in radiology reports,” *Journal of Digital Imaging*, vol. 31, no. 2, pp. 178–184, 2018.
- [3] M. Lu, Y. Fang, F. Yan, and M. Li, “Incorporating domain knowledge into natural language inference on clinical texts,” *IEEE Access*, vol. 7, pp. 57623–57632, 2019.
- [4] C. X. Huang, X. Huang, Y. Fang et al., “Sample imbalance disease classification model based on association rule feature selection,” *Pattern Recognition Letters*, vol. 133, pp. 280–286, 2020.
- [5] P. Kluegl, M. Toepfer, P.-D. Beck, G. Fette, and F. Puppe, “UIMA Ruta: rapid development of rule-based information extraction applications,” *Natural Language Engineering*, vol. 22, no. 1, pp. 1–40, 2016.
- [6] B. Rink, S. Harabagiu, K. Roberts et al., “Automatic extraction of relations between medical concepts in clinical texts,” *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 594–600, 2011.
- [7] M. Jiang, Y. Huang, J.-w. Fan, B. Tang, J. Denny, and H. Xu, “Parsing clinical text: how good are the state-of-the-art parsers?” *BMC Medical Informatics and Decision Making*, vol. 15, no. S1, 2015.
- [8] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. J. Khoury, “Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes,” *BMC Medical Informatics and Decision Making*, vol. 10, no. 1, p. 16, 2010.
- [9] M. Jiang, Y. Chen, M. Liu et al., “A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries,” *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 601–606, 2011.
- [10] O. Uzuner, J. Mailoa, R. Ryan, and T. Sibanda, “Semantic relations for problem-oriented medical records,” *Artificial Intelligence in Medicine*, vol. 50, no. 2, pp. 63–73, 2010.
- [11] X. Zhu, C. Cherry, S. Kiritchenko, J. Martin, and B. de Bruijn, “Detecting concept relations in clinical text: insights from a state-of-the-art model,” *Journal of Biomedical Informatics*, vol. 46, no. 2, pp. 275–285, 2013.
- [12] T. Mikolov, K. Chen, G. Corrado et al., “Efficient estimation of word representations in vector space,” 2013, <http://arxiv.org/abs/1301.3781>.
- [13] P. J. Stone, D. C. Dunphy, and M. S. Smith, *The General Inquirer: A Computer Approach to Content Analysis*, Cambridge, London, UK, 1966.
- [14] D. Demner-Fushman, E. Apostolova, R. Islamaj Dogan et al., ““NLM’s system description for the fourth i2b2/VA challenge,”” in *Proceedings of the 2010 I2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*, i2b2, Boston, MA, USA, 2010.
- [15] Y. H. Tseng, L. H. Lee, S. Y. Lin et al., “Chinese open relation extraction for knowledge acquisition,” in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, vol. 2, pp. 12–16, Gothenburg, Sweden, April 2014.
- [16] M. Hassan, O. Makkaoui, A. Coulet et al., *Extracting Disease-Symptom Relationships by Learning Syntactic Patterns from Dependency Graphs*, 2015.
- [17] X. Y. Bao, W. J. Huang, K. Zhang et al., “A customized method for information extraction from unstructured text data in the electronic medical records,” *Journal of Peking University. Health sciences*, vol. 50, no. 2, pp. 256–263, 2018.
- [18] X. Zhou, B. Liu, Z. Wu, and Y. Feng, “Integrative mining of traditional Chinese medicine literature and MEDLINE for functional gene networks,” *Artificial Intelligence in Medicine*, vol. 41, no. 2, pp. 87–104, 2007.
- [19] R. J. Ryan, “Groundtruth budgeting: a novel approach to semi-supervised relation extraction in medical language,” Doctoral Dissertation, Massachusetts Institute of Technology, Cambridge, MA, USA, 2011.
- [20] http://www.gov.cn/zwqk/2010-03/04/content_1547432.htm.
- [21] B. Hu and X. Liao, *Modern Chinese*, Higher Education Press, Beijing, China, 2002.
- [22] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [23] B. Zhang, M. Lu, and Y. Fang, “A feature-enhanced entity recognition method for Chinese electronic medical records,” in *Proceedings of the 9th International Conference on Information Technology in Medicine and Education (ITME)*, IEEE, Hangzhou, China, pp. 9–14, October 2018.
- [24] R. E. Fan, P. H. Chen, and C. J. Lin, “Working set selection using second order information for training support vector machines,” *Journal of Machine Learning Research*, vol. 6, pp. 1889–1918, 2005.
- [25] O. Uzuner, I. Solti, and E. Cadag, “Extracting medication information from clinical text,” *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 514–518, 2010.