Research on the Extraction of Medical Terms from Electronic Medical Records

Yao Yin[1], Fangfang Li[2, 3, *], Xingliang Mao[4], Hao Wang[5]

[1] School of Information Science and Engineering, Central South University, Hunan, China

[2]School of Information Science and Engineering, Central South University, Hunan, China.

[3]Mobile Health" Ministry of Education - China Mobile Joint Laboratory, Hunan, China.

[4]College of Information System and Management, National University of Defense Technology, Hunan, China

[5]Publicity Department of the Party Committee, Central South University, Hunan, China.

*corresponding author e-mail: lifangfang@csu.edu.cn

Keywords: Medical terms, CRF model, HMM model, electronic medical record

Abstract: Electronic medical records contain a large number of patient-related medical information. It is very important for doctors to diagnose diseases by mining useful information in electronic medical records. The extraction of medical terms is an important step in information mining from electronic medical records. Eyes are very important organs of human beings, therefore, this paper taking electronic medical records in the ophthalmology department in a hospital as the object, the main works of this paper are as follows. Firstly, the Conditional Random Field (CRF) model and its main steps of terms extraction are introduced. Secondly, the data set used in this paper and its annotation under the guidance of professional doctors are introduced. Finally, the extraction of medical terms by the CRF model is described in detail, and its result was compared with Hidden Markov Model (HMM) model. The experiment results show that CRF model based medical terms extraction in electronic medical records has made good performance.

1. Introduction

With the rapid development of information technology like computers, the application of this technology in hospitals can achieve digitalization of medical information about patients, and make doctors know well about patients' condition, realizing accurate diagnosis when making full use of existing resources. Electronic medical record (EMR) refers to the digital information like character, signal, diagram, figure, data and image by the means of medical information system in the medical activities, a medical record which can be stored, managed, transmitted and reappeared [1]. The electronic medical record is a full and professional description for patients' physical condition, bringing precious medical sources for us [2]. In 2010, the Ministry of Health stipulated t*he Fundamental Norms of Electronic Medical Record (trail)* to ensure the proper use and sound management of electronic medical records, prevent data missing, information stealing or other problems due to poor management [3].

Named entity (NE) means abstract or specific entity in the real world, such as human, organization, place, company and so on. When comes to electronic medical record, it refers to patients' symptom, therapy method and so forth. And name entity recognition refers to the identification of specific entity in the text, which is a valuable technology in processing Chinese information [4]. In recent years, the application of natural language processing technology in the clinical decision and support has become a new hotspot [5]. The main content extracted from electronic medical record is to recognize all entities expressing medical knowledge in these texts and establish relations among all entities [6]. Medical term is one of the named entities in this paper.

The earliest named entity recognition in electronic medical record often combines dictionary and rules [7, 8]. Min Jiang et al [9] introduced integrated medical language system (UMLS) and three natural language processing systems (Med LEE, DST, Knowledge Map) as characteristics, evaluated recognition of different machine learning algorithms with various features and proposed an entity extraction system integrating rules and machine learning algorithm. Berry de Bruijn [10] et al introduced UMLS, Ctakes, and Medline as characteristics and trained in the semi-Markov

  DOI: 10.25236/icemct.2018.091

model and gained the accuracy of 85.23%. Yaqiang Wang and others [11] constructed tagged symptom corpus including 11,613 chief complaints. Hui Wang [12] et al made trail for tumor medical records, and completed manual annotation for 12 data of liver cancer in 115 medical records with the help of two doctors. Lei Jianbo [13] et al selected 800 medical records including progress note and discharge summary, and established named entity tagged corpus among which word segmentation and part-of-speech tagging utilize tools developed by Stanford University. Xu Yan [14] et al set up annotated corpus including 336 discharge summaries targeted at four entities: medical problems, examination, treatment and medicine and put forward a united model of word segmentation and named entity recognition based on dual decomposition.

It can be seen from the above researches that no research has been done on electronic medical record in ophthalmology department about named entity recognition. For this reason, this paper studied named entity recognition based on CRF model aiming at Chinese electronic medical record in ophthalmology department. First of all, it annotated a large number of training corpus manually, then it trained CRF model by training corpus. Ultimately, it recognized entities for training corpus by use of trained models. Meanwhile, the paper conducted contrast experiment based on HMM algorithm.

2. Entity recognition based on CRF model

2.1. Introduction about CRF

Conditional Random Field (CRF) [15] is a machine learning method proposed in recent years, which counts joint probability distribution of overall annotated sequence when observed sequence is given. Conditional random field model is developed from Hidden Markov Model (HMM), and Maximum Entropy Markov Model (MEMM).

The definition of conditional random field: suppose G= (V, E) is the undirected graph, $Y = \{Y_v \mid v \in V\}$ is a group a random variables represented by set of nodes in the graph. When X is given, if each random variable Y_v complies with Markov characteristics, then conditional probability distribution $P(Y \mid X)$ is a conditional random field [16].

$$P(Y_v \mid X, Y_w, w \neq v) = P(Y_v \mid X, Y_w, w \sim v) \qquad (1)$$

In the above equation, $w \sim v$ denotes all nodes connected with node v in the graph G= (V, E), $w \neq v$ represents all nodes except v, and Y_v, Y_w are random variables of w corresponding to node v and w.

2.2. CRF based entity recognition

When conditional random field is applied to named entity recognition, the model can be regarded as a chain structure chart G= (V, E), in which adjacent two nodes and ligature among nodes constitute the largest connected ring. e= (v_{i-1}, v_i) represents undirected edge in a largest connected ring. Before defining the state transition Eigen function and State function, the real characteristic set of observed sequence must be constructed, that is the inputted electronic medical document.

Named entity recognition of electronic medical record consists of three steps:

Step 1: extract characteristics. Select characteristics in the electronic medical record, meaning makes special annotation for needed characteristics to identify Eigen function.

Step 2: evaluate parameters. Train training data by selected characteristics and get corresponding weight of all Eigen functions.

Step 3: annotation results. Complete sequential labeling for testing data by trained conditional random field model and finish named entity recognition. Testing data is the electronic medical record text which requires named entity recognition.

The medical records of ophthalmology used in this paper includes personal information like name and gender, chief complaint, history of present illness, previous history, personal history and

family history and so on. Since the paper only takes named entity recognition relevant to symptoms into consideration, and descriptions for symptoms are mainly in history of present illness, the paper merely extracts history of present illness of patients to finish entity recognition for symptoms. And the training corpus covers 30,414 characters and testing corpus covers 5100 characters. The flow chart of the algorithm is shown as in figure1.

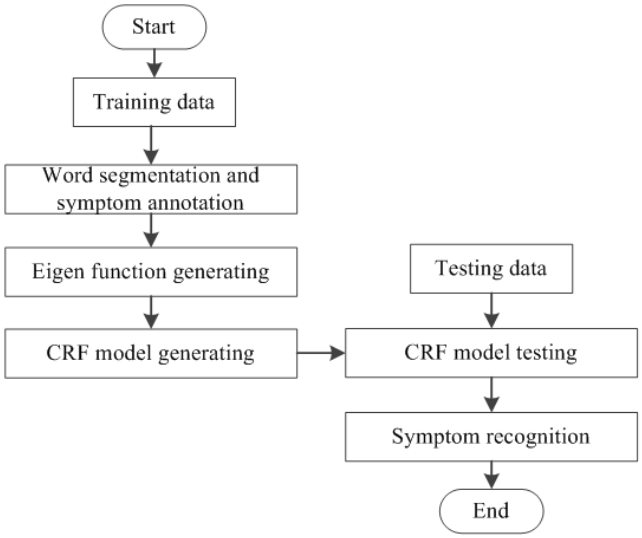

Figure 1. Algorithm flow chart

a) Preparation for training corpus

Each character is tagged in the training corpus, and each character as well as its characteristics occupies a line according to the format of training corpus required by CRF++ tool. Every sentence is separated by null string to represent a training block. There are eight annotations in total: B, M, E, S, BSY, MSY, ESY, and SSY. Among them, B, M, and E denote start tag, middle tag and ending tag of the entity other than the symptomatic entity respectively. While S represents single-byte characters other than the symptomatic entity. BSY, MSY, and ESY denote start tag, middle tag and ending tag of the symptomatic entity respectively. While SSY represents single-byte characters about the symptomatic entity. For example, left eye swelling one month. Then, this sentence is annotated as: B/E/SSY/B/E.

According to the requirements of the CRF++ tool, '/.../' occupies a single line in the training corpus. The test data is the same as the format of the training data, and each character occupies a single line. As can be seen from the above examples, in this paper, the annotation of training data set is based on the criteria of character segmentation. The characters here include a variety of types that appear in the medical records, including numbers, English letters and so on, not just Chinese characters. In the system of Chinese named entity recognition, the algorithm based on character segmentation is more concise, and the performance of character-based model is better than that of word-based and sentence-based model. Therefore, in this paper, the CRF model is obtained based on the characters.

b) Setting of feature template

Feature template is used to generate Eigen function. Different feature template can generate different Eigen functions. Compiling and adjusting feature templates can generate different function feature sets, therefore, different CRF models will also be obtained. The input format of the feature template is as %x [row, col]. While U00, U01... is the label of each feature function template, row is the relative row and col is the relative column. If there is only one marker in the training corpus, the value of col is 0. If there are many markers in the training corpus, such as character markers, word markers, the value of col is 0, 1, 2.... In this paper, the labeling features of the training corpus are only based on characters, so the col values are all 0, that is, the CRF model only considers the character features of the text context, not the word features or other. Table1 shows the feature template devised in this paper.

TABLE 1 FEATURE TEMPLATE

No.	
Line 1	# Unigram
Line 2	U00: %x[-2, 0]
Line 3	U01: %x[-1, 0]
Line 4	U02: %x[0, 0]
Line 5	U03: %x[1, 0]
Line 6	U04: %x[2, 0]
Line 7	U05: %x[-2, 0]/ %x[-1, 0]
Line 8	U06: %x[-1, 0]/ %x[0, 0]
Line 9	U07: %x[0, 0]/ %x[1, 0]
Line 9	U08: %x[1, 0]/ %x[2, 0]
Line 9	# Bigram
Line 9	B

U00, U01...U08 are grades of each Eigen function template. The first number in the parentheses represents the position relative to the current character, and the second number represents the type of context feature, which is shown in table 2.

TABLE 2 CONTEXT ANNOTATION

	column1	column2	
row-2	left	B	
row-1	eye	E	
Row 0	swelling	SSY	Current row
Row 1	one	B	
Row 2	month	E	

In the paper, context feature can be obtained by designing the above feature template, that is, the feature information about patients' history of present illness, which is shown in table 3.

TABLE 3 CONTEXT FEATURES

template	feature
U00:%x[-2,0]	left
U01:%x[-1,0]	eye
U02:%x[0,0]	swelling
U03:%x[1,0]	one
U04:%x[2,0]	month
U05:%x[-2,0]/%x[-1,0]	left/eye
U06:%x[-1,0]/%x[0,0]	eye/swelling
U07:%x[0,0]/%x[1,0]	swelling/one
U08:%x[1,0]/%x[2,0]	one/month

c) Preparation for testing data

According to the requirements of CRF++ tool, each typed character of testing data occupies a line.

3. Experiment

3.1 Data set

The experimental data are electronic medical records in ophthalmology department in a hospital, among which 375 medical records are selected randomly with 263 training sets and 112 training sets. The ratio of training set and testing set is 7:3. The paper mainly studies part of history of present illness in the electronic medical records. By extracting history of present illness in training

set and testing set, 29721 characters and 11156 characters are acquired respectively. There are two indictors to evaluate the precision of named entity recognition: precision rate and recall rate. The computational formulas of these two rates are as follow:

$$precision(P) = \frac{the\ number\ of\ \ properly\ recognized\ entity\ in\ the\ testing\ set}{total\ recognized\ entities\ in\ the\ testing\ set} \tag{2}$$

$$recall\ rate(R) = \frac{the\ number\ of\ \ properly\ recognized\ entity\ in\ the\ testing\ set}{total\ entities\ in\ the\ testing\ set} \tag{3}$$

3.2 Contrast experiment

Hidden Markov Model (HMM) is a kind of statistical model used to describe the Markov process involving hidden unknown parameter. It belongs to Markov chain, and its state cannot be observed directly but obtained through observing vector sequence. In the named entity recognition of electronic medical record, HMM can be adopted to perform the word segmentation. Therefore, in order to compare the effect of different models for extracting medical terms, this paper also carried out experiments on the HMM model.

Table 4 shows the two contrast experiments. Experiment A is named entity recognition based on HMM model. The trained HMM model and CRF model are used to recognize testing sets, and entity word representing symptoms (with SY as suffix) in the word segmentation are extracted. The precision rate and recall rate are indicated in table 5.

TABLE 4 CONTRAST EXPERIMENT

Number	Experiment name
A	HMM model
B	Model of this paper

From table 5, the precision rate of symptom entity of CRF model reaches 98.11% and recall rate 94.24%. While the precision rate of symptom entity in HMM algorithm is 93.19% and recall rate 91.21%, which indicates that the algorithm proposed in the paper makes good performance in named entity recognition of Chinese electronic medical records. If CRF model can make full use of the context information of characters in the modeling process, then the obtained CRF model will be more effective in entity recognition.

TABLE 5 PRECISION AND RECALL RATE

	The number of symptoms in testing set	Recognized symptoms	Properly recognized symptoms	precision	Recall rate
HMM model	330	323	301	93.19%	91.21%
CRF model	330	317	311	98.11%	94.24%

4. Conclusion

In a word, the paper has completed the following work:

Achieved named entity recognition of electronic medical records based on CRF model.

Conducted contrast experiment on CRF model and HMM model.

According to experimental results, the named entity recognition of electronic medical records in ophthalmology department in this paper achieves high accuracy. Compared with HMM-based recognition, this method has higher precision rate and recall rate.

By analyzing the average length of all types of entities in the training data set, it can be found that the closer the window size of feature template and the average length of entities are, the better the recognition effect of the corresponding entities will be. On the contrary, the more difference between the size of feature template window and the average length of entities, the worse the recognition effect will be. It is found that, when the size of the context window is set closer to the average length of the entity, the CRF model can make better use of the feature information of the entity lexical context; therefore, it can achieve better results.

In the near future, this experiment will be optimized to complete large-scale corpus annotation and enlarge testing sets and training sets so as to improve scope of application and efficiency of the algorithm in the paper.

Acknowledgements

This study is supported by the National Natural Science Foundation of China (61573380, 61602527), China Postdoctoral Science Foundation (2016M592450), and the Hunan Provincial Natural Science Foundation of China (2016JJ4119).

References

[1] Fundamental Norms of Electronic Medical Record of Ministry of Health of People's Republic of China (trail)[EB/OL].(2010-03-04). http://www.nhfpc.gov.cn/zhuzhan/wsbmgz/201304/a99a0bae95 be4a27a8b7d883cd0bc3aa.shtml.

[2] Wasserman R C. Electronic medical records (EMRs), epidemiology, and epistemology: reflections on EMRs and future pediatric clinical research [J]. Academic pediatrics, 2011, 11(4): 280-287.

[3] Qu Chunyan, Research on named entity recognition of Chinese electronic medical record [D]. Harbin Institute of Technology, 2015.

[4] Wang Feng. Research on named entity recognition based on CRF [D]. North University of China, 2011.

[5] Demner-Fushman D, Chapman W W, McDonald C J. What can natural language processing do for clinical decision support? [J]. Journal of biomedical informatics, 2009, 42(5): 760-772.

[6] Greenes R A, Shortliffe E H. Medical informatics: an emerging academic discipline and institutional priority [J]. Jama, 1990, 263(8): 1114-1120.

[7] Friedman C, Alderson P O, Austin J H, et al. A general natural-language text processor for clinical radiology [J]. Journal of the American Medical Informatics Association, 1994, 1(2): 161.

[8] Aronson A R, Lang F M. An overview of MetaMap: historical perspective and recent advances [J]. Journal of the American Medical Informatics Association, 2010, 17(3): 229-236.

[9] Jiang M, Chen Y, Liu M, et al. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries [J]. Journal of the American Medical Informatics Association, 2011, 18(5): 601-606.

[10] De Bruijn B, Cherry C, Kiritchenko S, et al. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010 [J]. Journal of the American Medical Informatics Association, 2011, 18(5): 557-562.

[11] Wang Y, Yu Z, Chen L, et al. Supervised methods for symptom name recognition in free-text clinical records of traditional Chinese medicine: An empirical study [J]. Journal of biomedical informatics, 2014, 47: 91-104.

[12] Wang H, Zhang W, Zeng Q, et al. Extracting important information from Chinese Operation Notes with natural language processing methods [J]. Journal of biomedical informatics, 2014, 48: 130-136.
[13] Lei J, Tang B, Lu X, et al. A comprehensive study of named entity recognition in Chinese clinical text [J]. Journal of the American Medical Informatics Association, 2014, 21(5): 808-814.
[14] Xu Y, Wang Y, Liu T, et al. Joint segmentation and named entity recognition using dual decomposition in Chinese discharge summaries [J]. Journal of the American Medical Informatics Association, 2014, 21(e1): e84-e92.
[15] Charles Sutton, Andrew McCallum. An introduction to conditional random fields [C]. Foundations and Trends in Machine Learning, 2010: 18-26.
[16] Jiang Wenzhi, Gu Jiaojiao, Hu Wenxuan et al. Research on Applications of Conditional Random Fields and Its Improvement [J]. Computer and Modernization, 2011(11): 55-58.