# A Bootstrapping Approach to Symptom Entity Extraction on Chinese Electronic Medical Records

Tianyi Qin[1], Yi Guan[2]

[1,] Web Intelligence Lab, Research Center of Language Technology,
School of Computer Science and Technology,
Harbin Institute of technology, 150001 Harbin,China, ralbuckle21@gmail.com

[2,] Web Intelligence Lab, Research Center of Language Technology,
School of Computer Science and Technology,
Harbin Institute of technology, 150001 Harbin,China, guanyi@hit.edu.cn

**Abstract.**Symptom entities are widely distributed in Chinese electronic medical records. Previous approaches on symptom entity extraction usually extract continuous strings as symptom entities and require massive human efforts on corpus annotation. We describe the symptom entity as two-tuples of <subject, lesion> and design a soft pattern matching method to locate them in sentences in the EMR. Our bootstrapping approach which only requires a few annotated symptom tuples and it allows iterative extraction from mass electronic medical record databases without human supervision. Furthermore, the described method annotates symptom entities in EMR by the extracted tuples. Starting with 60 annotated entities, our approach reached an F value of 81.40% in the extraction task of 3,150 entities from 992 sets of electronic medical records.

**Keywords:** electronic medical record, bootstrapping, named entity extraction, soft matching

## 1 Introduction

An electronic medical record(EMR) is the medical information of patients accessed and modified in a digital format written by medical staff in the process of medical activities.[1] There are four kinds of named entities in EMR[2]: disease, symptom, test and treatment. Among them, symptom entities have the most abundant and flexible knowledge about the physical condition of the patient, and are the starting point and primitive evidence of a clinical decision. Effectiveness of EMR varies by style of writing of medical staffs.[3-5]

With the help of medical consultants, we formulated a medical entity annotating specifications for Chinese EMR. In the specification, symptom refers to the discomfort caused by a disease or abnormal performance and explicit abnormal test results [6]. Its corresponding UMLS semantic types include signs or symptoms, mental or behavioral dysfunction and abnormal test results. A typical symptom is composed of some subjects, a lesion and some modifiers. Symptoms always occur in a certain body part, behaviors or states of patients which are called the subject in this paper. Pathological changes or abnormal states of subjects are also called lesions.[7] Some lesions happen on more than one subject, while others cannot be separated from subjects. Besides the subject and lesion, and other information in a symptom entity such as severity, frequency and so on, are called modifiers.[8]

Symptom entity extractions in Chinese EMR have difficulties due to the lack in Chinese medical knowledge base. However, symptom entities are strongly patternized in Chinese EMRs. Patternization of symptom entities is reflected in the following aspects: Most symptom entities consist of one subject,

one lesion and modifiers of them. Subject, lesion and modifiers are fixed types of words. For example, subjects are body parts and activities are words like “左下肢”(“left lower limb”),“右耳”(“right ear”) and “睡眠”(“sleep”), lesions are words describing abnormalities like “疼痛”(“pain”), “出血”(“bleeding”), and modifiers are positions, properties, degrees and negative words of subject and lesion. Our method utilizes patternized information of symptom entities in the process of extraction.

In this paper, we extract symptom entities in progress notes and discharge records of Chinese EMRs. On the basis of medical named entity annotating specifications, our method focuses on locating lesions and their corresponding subjects. Instead of extracting continuous strings as symptom entities, our method uses bootstrapping to iteratively extract two-tuples <subject, lesion> from a few annotated seed tuples in large unannotated Chinese EMRs. To realize the confidence measure of candidate tuples in the process of bootstrapping, we designed a soft pattern with the modifiers information of <subject, lesion>. Symptom entities can be formed by merging the extracted two-tuples and their modifiers in corpus according to the specification.

The remainder of this paper is organized as follows. Section 2 briefly introduces related works of our method. Details about the implementation of bootstrapping, including soft pattern initializing and updating, candidate entity extracting and confidence measure are described in section 3. The results of the experiments are shown in section 4 including the performance of our system with different parameters and also contrasting it against experiments of other researchers. Section 5 gives the conclusion of this paper and discusses future work.

## 2 Related Work

There has been several previous works for medical entity extracting. The most effective methods of named entity extraction on EMR are based on rule and dictionary, or supervised machine learning. In the concept extraction task of the 2010 i2b2/VA challenge, unannotated text of patient reports were given for systems to identify and extract the text corresponding to patients medical problems, treatments and tests[9]. Savova G et.al[10] built a NLP system for information extraction from EMRs by trained clinical domain dictionary on CTAKES. Feng Y, Li D, Jiang M et.al[11-13] proposed supervised machine learning methods for clinical named entity extracting. Some researchers choose semi-supervised machine learning methods rather than supervised machine learning methods. Jonnalagadda[14] et al. used a semi-supervised sequential discriminative classifier (Conditional Random Fields) to extract the mentions of medical problems, treatments and tests from clinical narratives. In addition to the traditional features such as dictionary matching, pattern matching and part-of-speech tags, their method also used as a feature words that that appear in similar contexts to the word in question. Words that have a similar vector representation measured with the commonly used cosine metric, where vector representations are derived by using methods of distributional semantics. The F-value of this method was 0.823. De Bruijn[15] et al. realized concept tagging by a discriminative Semi-Markov model. Semi-Markov models are Hidden Markov Models that tag multi-token spans of text, as opposed to single tokens. Only four tags: outside, problem, treatment, and examination are needed. Concept mapping features include context features and word-level features extracted from cTAKES and UMLS output. Their method reached an F-value of 0.8523, which was the best in the concept extraction task of the i2b2 2010 challenge.

Our approach is partly inspired by the medical information extracting method of [16]. That method proposed a definition of two-tuples <target, description> for medical information. With a series

of algorithms including pattern generalization, pattern automatic extraction and medical information extraction, their method iteratively extracts two-tuples by generalized patterns. Our work applies a similar two-tuple definition in symptom entity extracting and introduces soft pattern and fuzzy matching from Zhao J's inspiration. Their method[17] offers a solution to generating a soft pattern. It generates pattern examples through segmenting sentences by event examples and filters them by notion words, stop words and trigger words.

## 3 Method

The main idea of our method is that we can iteratively extract two-tuples of <target, description> from only a few annotated tuples by bootstrapping, and form symptom entities by extracted tuples. Bootstrapping is a simple and intuitional method which is widely used in natural language processing.[18] Two main difficulties of the bootstrapping method are: (1) Effectiveness of bootstrapping is directly affected by the precision and covering ability of the initial seed set. (2) Errors occur in the self-training process which will be magnified during iteration. For the first difficulty, we invite trained medical staffs to accomplish the annotating task. Since Chinese EMRs from various departments have common linguistic features, we are able to guarantee the quality of seed set. For the second difficulty, a soft pattern provides us a solution for the measuring confidence of a candidate's two-tuples. We calculate the similarity between word frequency vectors in the slots of the soft patterns and word frequency vectors of <subject, lesion>'s context information. In each iteration we add symptom entities with the highest confidence value and adjust the soft pattern according to the context information of the newly added entities.

Our symptom entity extraction task is similar to the concept extraction task of the i2b2 2010 challenge but focuses on symptom information. Given Chinese EMRs from several departments, we located the border of two parts: subject and lesion. Examples of two-tuples extraction are described as follow:

Source Example 1: 右侧 口角 流涎 ("Right side of the mouth drooling.")

Result 1: <symptom> 右 侧 <subject> 口 角 </subject> <lesion> 流 涎 </lesion></symptom> ("<symptom>Right side of the <subject>mouth</subject> <lesion>drooling</lesion></symptom>").

Source Example 2: 左 上肢 及 双 下肢 骨折("Left upper limb and lower limbs fracture.")

Result 2: <symptom>左 <subject>上肢</subject> 及 双 <subject>下肢</subject> <lesion>骨折 </lesion></symptom> ("<symptom>Left <subject>upper limb</subject> and <subject>lower limbs</subject> <lesion>fracture</lesion></symptom>")

Source Example 3: 心脏 各 瓣膜区 未 闻及 病理性 杂音 ("Heart valve areas found no pathological murmurs.")

Result 3: 心脏 各 <subject>瓣膜区</subject> 未 闻及 <symptom>病理性 <lesion>杂音 </lesion></symptom> ("Heart valve areas found no <symptom>pathological <lesion>murmurs</lesion></symptom>.")

Source Example 4: <symptom> 间 歇 性 <lesion> 头 晕 </lesion></symptom> ("<symptom>Intermittent <lesion>dizziness</lesion></symptom>")

In the process of forming symptom entities, source example 2 includes two subjects corresponding to one lesion and we annotate the whole sentence as a symptom entity. Subject and lesion are separated by negative words and modifiers in source example 3, so we only annotate lesion as the symptom entity. And "dizziness" of source example 4 can't be segmented in Chinese.

Our target domain documents are progress notes and discharge records from 2003 to 2014 provided by the records room from the second affiliated hospital of Harbin Medical University (HMUSAH). Original medical records are paper records so we used OCR(Optional Character Recognition) to convert them into EMRs.

### 3.1 Definition of symptom entity and soft pattern

Subject and lesion are two main parts of symptom entities. In this paper we define lesion as a description of performance of physical abnormalities, and the subject as body parts or subjects where the lesion occurs. Subject and lesion are both single words which have direct semantic association with each other. For example, in the sentence “颈部 活动 受限”(“Neck’s activity limited”), subject of “受限”(“limited”) should be “活动”(“activity”) but not “颈部”(“Neck”). We extract two-tuples of <subject,lesion> which contain information of the symptom entity to locate them in large unannotated EMRs.

The bootstrapping used in our method starts from a few annotated <subject,lesion> seed tuples, to choosing the candidate entities by POS information, then uses fuzzy matching with soft pattern to measure confidence of candidate tuples, adds tuples of highest confidence to the seed tuple set and finally starts a new iteration. After a certain number of iterations, we form symptom entities with extracted tuples according to the annotating specification. The flow chart of bootstrapping is shown in Figure 1.

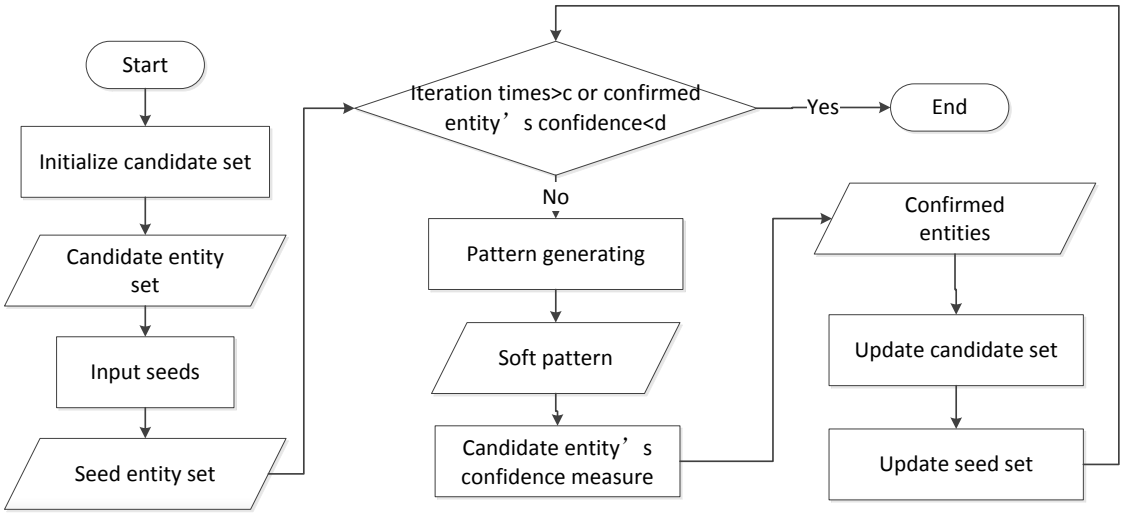


**Fig. 1**. Flow chart of bootstrapping

Although the basic flow maintains similarities with other general bootstrapping methods, our method uses fuzzy matching instead of hard matching. In order to realize the fuzzy matching, we set up a soft pattern according to the subject and lesion of the symptom information. We initialize a soft pattern by context information of the seed entities and update it in each iteration. Confidence values of candidate tuples are measured by the matching degree with the soft pattern.

### 3.2 Initialization of the candidate set and soft pattern

Our initial candidate set contains all the two-tuples of <subject,lesion> which are possible to be part of

the symptom entities. For extracting the candidate two-tuples, we traverse segmented and unannotated corpuses sentence by sentence since symptom entity and its related context features are distributed mainly in its locating sentence. In each sentence, the subject always appears in front of the lesion. Every noun can be extracted as a subject, forming a tuple with all nouns/verbs/adjectives behind it as lesions. Specially, for sentences which have no nouns, we only extract candidate lesions and the subject remains empty.

For each candidate tuple, we divide its locating sentence into three parts according to the position of subject and lesion:

$(L_1,L_2,...L_a)$, subject, $(M_1,M_2,...M_b)$, lesion, $(R_1,R_2,...R_c)$

Where $L_n,M_n,R_n$ are modifiers in certain position. For example, there is a seed tuple <肺,啰音>(“<lung, rale>”) in a seed set and we find the sentence “双 肺 未 闻及 明显 啰音”(“Both lungs heard no obvious rale.”) containing it. Then we save the information of the sentence in three slots: {双},{未,闻及,明显}{null} ({“Both”},{“heard no obvious”},{null}). And these three slots become the basic form of our soft pattern:

<slotL>,subject,<slotM>,lesion,<slotR>

Each slot in a soft pattern stores a word frequency vector. We count all the sentences with seed tuples and form three slot vectors. SlotL and slotM are considered as modifiers of a subject while slotM and slotR are considered to be modifiers of the lesion. Specially, if a tuple has no subject, we count the words in front of the corresponding sentence into slotM because only slotM stores modifiers of the lesion. The feature vector for each candidate tuple is formed in the same way with the slot vectors of a soft pattern.

### 3.3 Candidate entity’s confidence measure

Most standard bootstrapping methods have to measure the confidence of both the entities and patterns, but our method updates the soft pattern by the confirmed two-tuples instead of measuring its confidence. Based on the generalized soft pattern and slot vectors containing standard information of modifiers, we are able to transform a candidate tuple evaluation problem to a word vector similarity calculating problem in which vectors have more common terms obtain a higher similarity value. The matching degree for tuples can also be represented as a cosine similarity value between feature vectors of candidate tuples and standard vectors in the slots of the soft pattern. The cosine similarity between two vectors $\bar{x}$ and $\bar{y}$ is:

$$Sim(\bar{x},\bar{y}) = cos(\bar{x},\bar{y}) = \frac{\bar{x}\cdot\bar{y}}{|\bar{x}||\bar{y}|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \quad (1)$$

The cosine similarity value between Left_Gram, Middle_Gram , Right_Gram and their corresponding standard vector in the slots of the soft pattern are calculated, respectively. In addition, we combine Left_Gram and Middle_Gram as modifiers for the subject, and Middle_Gram and Right_Gram as modifiers for lesion. The standard vectors in the soft pattern are also combined in this same way. Thus, we calculate five similarity values shown in following table:

**Table 1.** Similarity values measured in confidence calculating

| Symbol | Related feature vector |
|---|---|
| $SimL(\bar{x}, \bar{y})$ | Left_Grams |
| $SimM(\bar{x}, \bar{y})$ | Middle_Grams |
| $SimR(\bar{x}, \bar{y})$ | Right_Grams |
| $SimLM(\bar{x}, \bar{y})$ | Left_Grams, Middle_Grams |
| $SimMR(\bar{x}, \bar{y})$ | Middle_Grams, Right_Grams |

We obtain the average value of these five similarity values to get the confidence of a tuple:

$$Sim(\bar{x}, \bar{y}) = \frac{SimL(\bar{x}, \bar{y}) + SimM(\bar{x}, \bar{y}) + SimR(\bar{x}, \bar{y}) + SimLM(\bar{x}, \bar{y}) + SimMR(\bar{x}, \bar{y})}{5} \quad (2)$$

For entities with empty subjects, we ignore the similarity value using the slotL:

$$Sim(\bar{x}, \bar{y}) = \frac{SimM(\bar{x}, \bar{y}) + SimR(\bar{x}, \bar{y}) + SimMR(\bar{x}, \bar{y})}{3} \quad (3)$$

All candidate symptom entities are scored by the above formula. For each iteration, we preserve N tuples(N is a presupposed constant) with the highest similarity value is updated for the seed tuple set.

### 3.4 Soft pattern update and iteration

According to the property of symptom entities in annotating specifications, we set up the following rules for annotating symptom entities: 1) If a symptom and its body part or subject are directly connected in a sentence, they are annotated as one symptom entity regardless of the number of body parts or subjects. 2) If the symptom and its body part or subject are inseparable, we annotate them as one symptom entity. 3) If the symptom and its body part are separated by punctuations, negative words or modifiers, only the symptom is annotated.

After one iteration, we obtain an expanded seed tuple set. The soft pattern is updated as described in chapter 3.2. In order to comply with annotating specification, we obtain the final results of the symptom entity annotation by processing the seed tuple and the soft pattern in the iteration.

As shown in the source examples, symptom entities in Chinese EMRs come in mainly four forms: 1) Single subject adjacent to lesion. 2) Multiple subjects adjacent to lesion. 3) Subject and lesion divided by negative words or punctuations. 4) Lesion has no corresponding subject. In order to comply with annotating specifications, we annotate subjects, lesion and their modifiers as symptom entities in case 1) and 2), and then annotate lesion and its modifiers in case 3) and 4).

Slot vectors of the soft pattern offer us a set of statistic information about the words in a sentence with <subject,lesion> which can be directly used to determine the boundary of the symptom entity. According to the four forms of symptom entities in Chinese EMRs, symptom entity annotating based on our bootstrapping method applies the following rules:

1) If there are negative words in slotM, ignore the subject, slotL and slotM.

2) If multiple subjects appear in one sentence, annotate them without violating rule 1).

3) If there are no subjects in the sentence, ignore slotL.

For slots not ignored, we determine the symptom entity's border by calculating word frequency in its corresponding slot. If word frequency of one word exceeds a threshold value, we annotate it as a part of the symptom entity. If the frequency of one word in slotL cannot reach the threshold, we ignore the words to its left in the sentence, and also ignore the words to the right of the first low frequency word in slotR. Specially, if the frequency of one word in slotM cannot reach the threshold, we ignore the whole sentence because the border of a symptom entity cannot be determined.

# 4 EVALUATION

## 4.1 Source data and security measures

In this section we represent the experiments for symptom entity extracting using the bootstrapping method. We randomly select 992 sets of Chinese EMRs whose privacy information was removed. Authorization of our source data were all given by the records room from the second affiliated hospital of the Harbin Medical University (HMUSAH), and their confidentiality and security are ensured by the regulations of the hospital. We intercept admission condition, discharging condition from discharge records, and disease cases from progress notes. The whole process of the experiment keeps no negative implications for the patients.

In order to guarantee the universal properties of the source data, EMRs involved in our experiment are from 20 different departments. The writing form of our EMRs are in accordance with the writing standard made by China's ministry of health.

## 4.2 Performance of symptom entity extracting system

Three measures: precision rate, recall rate and F value are indicated to be necessary according to previous experience. We evaluate these three measures of our bootstrapping method on segmented Chinese EMRs.

As mentioned in 2.1, from 992 sets of Chinese EMRs we manually marked 3,150 symptom entities. We manually annotate 60 high frequency pairs of subject and lesion as seed tuples. Starting with 60 seed tuples, we aimed to extract more seed tuples and put them into symptom entities. In each iteration, we add 60 symptom tuples with the highest confidence values to the seed tuple set, and get symptom entities ad described in section 3.5. Iteration terminates either when no entity tuple reached the confidence threshold in one iteration or after 74 iterations. The trend chart is shown in Figure 2 and parts of the detailed results are shown in Table 2.

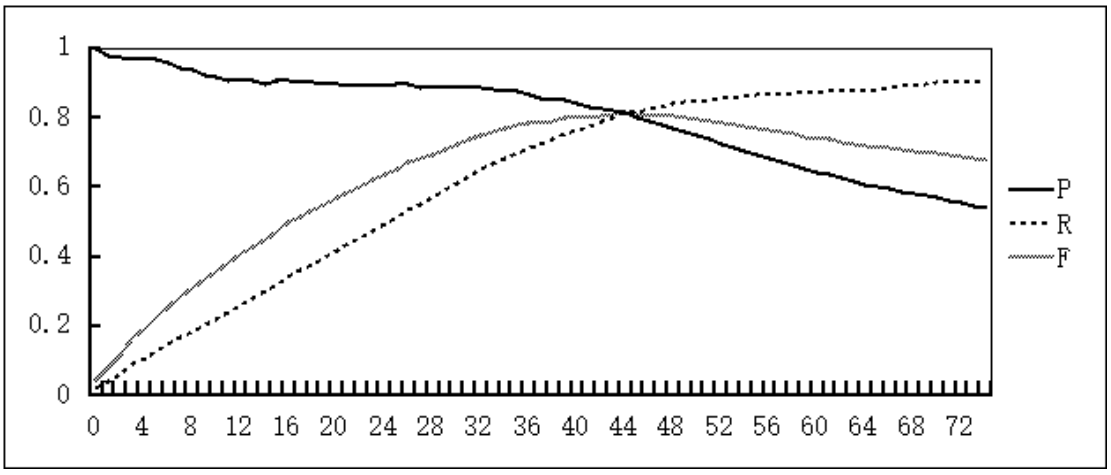


**Fig. 2.** Performance of symptom named entity extracting on unannnotated EMR

**Table 2.** A selection of experimental data from Figure 2

| Item | Iteration | Precision | Recall | F value |
|---|---|---|---|---|
| Maximum F value | 44 | 81.40% | 81.40% | 81.40% |
| Maximum R value | 75 | 54.03% | 90.35% | 67.63% |
| Other peak F values | 49 | 76.91% | 83.37% | 80.01% |
| | 58 | 71.54% | 91.50% | 80.29% |

In each iteration, 60 seed tuples are added to seed tuple set. F value of extracted seed tuples keeps monotone increasing in first 44 iterations, and keeps monotone decreasing after that. The F value of iteration 44 is the only maximum value in experiment.

Attempt of increasing number of seed tuples and decreasing number of candidate entities added to seed set in each iteration can slightly increase the F value. But trend of precision rate, recall rate and F value maintain similar with original experiment.

#### 4.3 Experiment comparison

We made contrasts with experiments which used other methods of pattern generalizing, pattern extracting and medical information extracting of Chinese EMRs.

Xu's method[16] put forward a five-tuples mode of medical information: <target modification, target, degree, property, description> and we simplified the five-tuples into two-tuples of <target,description> and their semantic description. In the process of pattern generalizing, they segmented sentences by <target,description> and arranged midterm segmentation by the order of target, modification, degree and description. The generalized pattern was recombined and stored in a pattern base. In the process of pattern extracting and medical information extracting, they also used a bootstrapping method which generalized candidate sentences as specialized patterns and measured matching degree with the pattern base by a synonym dictionary.

Compared with Xu's method, our method also segmented sentences by two-tuples <subject, lesion> and extract symptoms information iteratively. The main differences between our method and Xu's method are that we used a soft pattern to measure the confidence of the candidate tuples. Our

contrasting experiment used similar parameters as stated in 4.2. Sixty seed entities were chosen to extract 1,846 artificial marked symptom entities. The The experiment results are shown in Table 3.

**Table 3.** Results of Comparison

| Item | Our approach | Xu's method |
|---|---|---|
| Precision | 81.40% | 73.19% |
| Recall | 81.40% | 60.06% |
| F value | 81.40% | 65.98% |

In addition, we made another contrasting experiment with another bootstrapping method which estimated word frequency by an EM procedure and used the left and right branching entropy to build an appropriateness measure[17]. The experiment results are shown in Table 4.

**Table 4.** Results of Comparison

| Item | Our approach | Zhang's method |
|---|---|---|
| Precision | 81.40% | 43.65% |
| Recall | 81.40% | 17.77% |
| F value | 81.40% | 25.28% |

### 4.4 Discussions

Our experiment in 4.2 shows that our approach can extract symptom entities with highest F value of 81.40%. By updating soft pattern and confidence of candidate entities, our F value keeps increasing in front iterations. But after iteration 44, F value decreases because of candidate entities with lower confidence. Negative entities with high frequency in EMR data add false information to pattern, and leads to more negative candidate entities. Highest F value of 81.40% ensure that our approach can extract symptom entities effectively.

In contrast with other existing medical entity extracting approaches, our approach has two main advantages. First, it substantially reduces the cost of manual annotation. Second, instead of a single string, it extracts a symptom entity's main information: subject and lesion as the item for extracting. This method can locate the symptom entity more easily, offer more detailed information about a certain body part and its corresponding lesion or limited activities. Our approach maintains a stable F value trend and is able to adjust performance by modify starting conditions.

A main limitation of our approach is that few symptom entities can't be described as subject and lesion, such as "恶心"("nausea") and "晕厥"("syncope"). Our approach usually extracts random subjects for these entities with no exact subject, which causes more false information in soft pattern. Moreover, Chinese word segment cause some subjects and lesions segmented together, such as "头痛"("headache") with "头"("head") and "痛"("ache"). These candidate entities will also cause false information.

## 5 CONCLUSIONS

We have demonstrated a bootstrapping symptom entity extracting method suitable for unannotated

Chinese EMRs. We extracted the symptom entity's main information as two-tuples of <subject,attribute> and designed a soft pattern to locate them in sentences in EMR. By using a word vector oriented soft pattern, we could find a constant number of entities with the highest confidence value and update the pattern in each iteration. Experiments on Chinese electronic medical records show that our approach have reached an acceptable F value.

## References


[1] Nadeau D, Sekine S. A survey of named entity recognition and classification[J]. Linguisticae Investigationes Revue Internationale De Linguistique Française Et De Linguistique Générale, volume 30:3-26(24)(2007)

[2] Qu C,Guan Y, Yang J, Liu Y. The construction of annotated corpora of named entities for Chinese electronic medical records[J]. Chinese High Technology Letters, 2(5)(2015)

[3] Sittig D F, Singh H. Which electronic health record is better: A or B? Realities of comparing the effectiveness of electronic health records[J]. Journal of Comparative Effectiveness Research, 3(5):447-50(2014)

[4] Erica B, Field J R, Sunny W, et al. Biobanks and electronic medical records: enabling cost-effective research.[J]. Science Translational Medicine, 6(234):86-86(2014)

[5] W-Q W, Feng Q ,, Jiang L ,, et al. Characterization of statin dose response in electronic medical records.[J]. Clinical Pharmacology & Therapeutics, 95(3):331-338(2014)

[6] https://github.com/WILAB-HIT/Resources

[7] Eriksen T E, Risør M B. What is called symptom?[J]. Medicine Health Care & Philosophy, 17(1):89-102(2014)

[8] Yang J, Yu Q, Guan Y, Jiang Z. An overview on research of electronic medical record oriented named entity recognition and entity relation extraction[J]. Acta Automatica Sinica, 40(8):1537-1562(2014)

[9] Ö U, BR S, S S, et al. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text.[J]. Journal of the American Medical Informatics Association, 18(5):: 552–556(2011)

[10] Savova G K, Masanz J J, Ogren P V, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications[J]. J Am Med Inform Assoc, 17(5):: 507–513(2010)

[11] Feng Y. Intelligent Recognition of Named Entity in EMRs[J]. Chinese Journal of Biomedical Engineering, 30(2):256-262(2011)

[12] Li D, Savova G. Conditional random fields and support vector machines for disorder named entity recognition in clinical texts[J]. Proceedings of the workshop on current trends in biomedical natural language processing (BioNLP'08, 2008:94—95(2008)

[13] M J, Y C, M L, 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, 8(5):601-606(2011)

[14] Jonnalagadda S, Cohen T, Wu S, et al. Enhancing clinical concept extraction with distributional semantics[J]. Journal of Biomedical Informatics, 45(1):129–140(2012)

[15] Bruijn B D, 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, 18(5):557-562(2011)
[16] Xu G, Quan G, Wang Y. Research of electronic medical record key information extraction based on HL7[J]. JOURNAL OF HARBIN INSTITUTE OF TECHNOLOGY, 3(11):89-94(2011)
[17] Zhang L. Chinese EMR word segmentation and named entity mining based on semi supervised learning[D]. Harbin Institute of Technology(2014)
[18] Zhao J,Qin B. Design and Implementation of Event Arguments Extraction System based on BootStrapping[J]. INTELLIGENT COMPUTER AND APPLICATIONS, 2(1):16-20(2012)