Chronic Disease Related Entity Extraction in Online Chinese Question and Answer Services

Yan Zhang[1(✉)], Yong Zhang[1], Yanshen Yin[1], Jennifer Xu[2], Chunxiao Xing[1], and Hsinchun Chen[1,3]

[1] Research Institute of Information Technology, Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing, China
zhang-yanl4@mails.tsinghua.edu.cn, hchen@eller.arizona.edu
[2] Computer Information Systems, Bentley University, Waltham, USA
jxu@bentley.edu
[3] MIS Department, University of Arizona, Tucson, USA

Abstract. Chinese chronic disease entity extraction aims to extract health related entities from online questions and answers (QA). Our research tackles challenges in Chinese chronic disease entity extraction from three aspects: Chinese health lexicons construction, feature development, and equivalence conjunctions tagging. We construct large scale Chinese health lexicons based on expert knowledge and the Web resources; develop a feature extraction approach that draws out character, part-of-speech, and lexical features from QA data; and improve the performance of answer entity extraction by leveraging equivalence conjunctions (punctuation marks and conjunctional words) in Chinese to capture dependencies between tags of entities. Experiments on question and answer entity extraction demonstrate that the Precision, Recall and F-1 score are improved using our proposed features, and the Precision and F-1 score can be further improved by considering equivalence conjunctions.

Keywords: Entity extraction · QA · Health lexicon

1 Introduction

Aging population has become a serious problem in many countries around the world. In China, the number of seniors aged 60 and above had risen to 212 million (15.5 %) by the end of 2014 [9]. These seniors have a pressing need for high quality medical services for treating their chronic diseases, which greatly worsen their quality of life. In China, about 86.6 % of total deaths in 2012 were caused by chronic diseases, such as diabetes, hypertension, etc. [9]. Unfortunately, high quality healthcare resources have long been in short supply in China. In large cities such as Beijing, an average physician may have to see around a hundred patients every day, and literally has no time to answer patients' questions in detail or to provide personalized medical advice.

X. Zheng et al. (Eds.): ICSH 2015, LNCS 9545, pp. 55–67, 2016.
DOI: 10.1007/978-3-319-29175-8_6

This situation has been mitigated to some extent by online medical and healthcare services such as health portals, blogs, Questions and Answers (QAs), and discussion forums. For instance, QA services, whether they are community-based or expert-based, encompass a large amount of user generated content (UGC) and have become an alternative channel through which patients seek health-related information. How to leverage UGC and provide quality QA services is a nontrivial problem.

In this research, we propose a named entity extraction approach to help improve health-related QA services in China using UGC data. Entity extraction, or entity recognition (ER), aims to recognize and identify entities out of unstructured texts. A named entity is a contiguous sequence of textual tokens for representing the name of an object in a certain class (e.g., person or organization). The entity can be general (e.g., organization names) or domain specific (e.g., medicine names). The extracted entities are used to measure the similarity between questions and answers. Generally speaking, if an answer and a question have more common entities, it is more likely that the answer is for addressing the question. Therefore, entity extraction can help identify and construct answers to patient questions with higher relevance and quality.

Entity extraction is essentially a multi-class classification task, which assigns an entity label to each word in a sentence. One of the biggest challenges for Chinese entity extraction is the lack of standard domain lexicons. In techniques and applications for extracting entities out of English texts, there often are a number of domain lexicons available for text processing. For example, the Unified Medical Language System (UMLS [10]) has been widely used in text mining applications in the medical domain. Unfortunately, there has not been a standard Chinese lexicon available for the medical and health domain. Another challenge is the processing of Chinese UGC data. UGCs are usually not formal writing and tend to have a lot of "noises," such as incorrect syntax, misspellings, lay-person terms, or missing punctuations. Furthermore, some unique Chinese punctuation marks, such as the enumeration comma "、", which is used to separate items in a series, need special treatment. For example, in the sentence "治疗糖尿病的方法有药物疗法、运动疗法、饮食疗法。(Treatments for diabetes include medication therapy, physical therapy, and diet therapy.)", the enumeration comma can be leveraged to help identify entities of the same types.

The main contribution of our study is two-fold. First, we construct Chinese chronic disease lexicons to facilitate chronic disease entity extraction in online Chinese QA services. Second, we propose a CRF-based entity extraction approach for Chinese chronic disease entity extraction. In this approach, we define QA entity types based on analysis of large amounts of QA data, extend the tags to leverage unique punctuation marks (i.e., enumeration comma) and conjunctions in Chinese to capture dependencies between tags of entities, and propose a feature extraction approach to extracting character, part-of-speech, and lexical features from online QA. Experiments on health related QA entity extraction show promising performance of our approach.

The remainder of the paper is organized as follows. Section 2 introduces the related work. Section 3 presents our research design. We report on our experiments and results in Sect. 4. The last section concludes this paper.

2 Related Work

Entity extraction in our study aims to extract entities related to chronic diseases from the UGC in online QA services. The extracted entities are used to measure the similarity between the questions and answers by identifying the common entities. In this section, we review the literature on QA systems and entity extraction methods.

QA Systems. There are two types of QA systems: domain-independent and domain-specific. Here we focus on the domain-specific QA systems. In the medical domain, two QA systems, MedQA [6] and AskHERMES [2], are well-known. MedQA was developed in 2006 and is the first QA system for physicians. MedQA uses the records from MEDLINE and Internet. AskHERMES was released in 2011 and employs the UMLS in question analysis. It relies on a dynamic hierarchical categorization model to select answer sentences, and uses question-oriented keywords to assemble the final answers. It not only provides the answers to a user question, but also suggests related questions. Few studies can be found for Chinese QA systems. Peng et al. [13] developed a Chinese QA system based on an enterprise knowledge base. Zhang et al. [15] designed a document retrieval method for a Chinese QA system using professional documents.

Entity Extraction. Feature extraction and entity classification are two major components in entity extraction techniques. Feature extraction is to extract a set of relevant features for building robust learning models. In machine learning approaches, feature extraction is very critical and can significantly affect the performance of entity classification. Prior entity extraction studies have adopted various types of features, including word and contextual features [3], structural features and denotation features, which consider the coherence or appropriateness of the selected entity strings [12]. [14] groups features into different dimensions: word-level features, list lookup features, contextual features, and language-specific features.

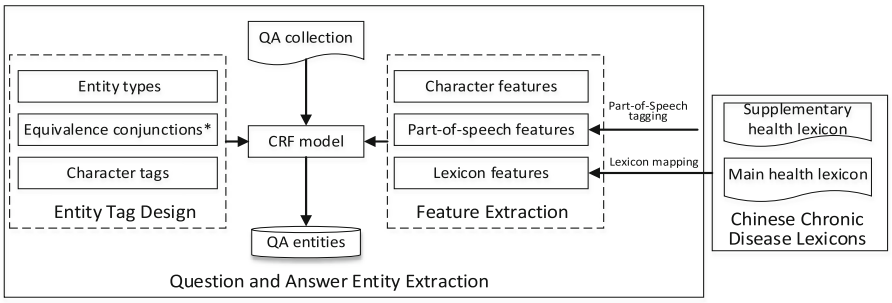
Entity classification is to classify extracted entities (terms or words) into predefined target classes, i.e., to assign a class label to each entity. Three types of entity classification methods have been proposed in the literature, namely, supervised learning, semi-supervised learning, and unsupervised learning [8]. Supervised learning includes a training stage and testing stage. During the training, records with class labels are used to construct the classification model (a.k.a. classifier). The classifier is then tested using the test data (i.e., records with their known class labels removed). Among the many supervised methods, Hidden Markov Model (HMM), Maximum Entropy, Support Vector Machine (SVM), and Conditional Random Field (CRF) are the most widely used. Unsupervised learning does not require model training. The most widely used approach is clustering. Other unsupervised techniques employ lexicon, words' pattern, and unlabeled corpus. Semi-supervised learning [11] uses two kinds of samples in the training set: one set with labeled records and the other with no labels. The semi-supervised learning combines the supervised learning with unsupervised learning to reduce the effort to label samples and achieve high accuracy at the same time. In our experiments we choose a supervised learning method: CRF [7]. CRF is one of the most effective methods for sequence labeling such as part-of-speech tagging and entity

extraction. Entity extraction is different from general classification in that labels of neighboring items are dependent. For example, while "heart" is an organ, "heart disease" is a disease. Unlike traditional classification methods that classify each word separately, CRF considers the interdependency between labels by incorporating graphical models.

Although entity extraction has been included in medical QA systems in English, little research has been done for systems in Chinese due to the lack of standard domain lexicons and the unique characteristics of Chinese. In our research, we build Chinese lexicons for chronic diseases, design entity tags and extract features considering the QA and Chinese characteristics, and apply a CRF-based machine-learning approach to classifying the entities when processing questions and constructing answers.

3 Research Design

Figure 1 presents the system architecture for our chronic disease related entity extraction approach. It contains two components: Chinese Chronic Disease Lexicons and Question and Answer Entity Extraction. There are two main steps in our CRF-based QA entity extraction approach: Entity Tag Design and Feature Extraction.



* for answer entity extraction only.

Fig. 1. System Architecture

3.1 Chinese Chronic Disease Lexicons

The extraction of Chinese medical and health-related entities relies on a high quality Chinese chronic condition lexicon. We create two lexicons in our study: an expert-based main lexicon, and a Web-based supplementary lexicon.

Main Health Lexicon. We collaborated with domain experts to create our main lexicon based on professional dictionaries and medical textbooks. The lexicon consists of terms and phrases for diseases, symptoms, diagnosis, medicines, and their relationships. The lexicon is organized in a database, whose metadata are presented in Table 1.

Table 1. The schema of the main health lexicon

Table name (attributes)	#Records
Disease (Name, Subject, System, Department (>=1), Body Part (>=1))	512
Symptom (Name)	2,162
Diagnosis (Name, Abbreviation)	1,080
Medicine (Name, Manufacturers(>=1), URL, Alias)	3,200
Disease-Symptom (Disease, Symptom)	4,738
Disease-Diagnosis (Disease, Diagnosis)	3,049
Disease-Medicine (Disease, Medicine)	6,673

Supplementary Health Lexicon. Manually constructing a domain lexicon is labor-intensive and time-consuming as experts must identify not only concepts but also their relationships. As a result, this manual construction approach is only applicable for small-sized lexicons and can hardly scale up. To construct a more complete, large-scale health lexicon, we collected entries from existing health lexicons on the Web (pinyin.sogou.com, dict.bioon.com, zzk.xywy.com, jib.xywy.com, and yao.xywy.com), and created a supplementary lexicon. Although entities contained in these online resources are concepts without associated metadata and relationships, they greatly reduced the time and effort required by the manual approach. We categorized the concepts into nine classes and placed them in eight tables, as shown in Table 2.

Table 2. Supplementary health lexicon

Table name	Examples	#Records
Disease and symptom	Diabetes mellitus, gingival bleeding	31,452
Medicine	Melbine	38,726
Food	Banana, rice, coffee	26,893
Organ	Lymph gland, heart	6,089
Sign	Urine, sweat	149
Index	Temperature, blood-sugar, lymphokine	3,314
Diagnosis	NMR, CT, pregnancy test	3,473
Treatment	Nasal fistula excision, bone transplantation	8,493

3.2 Question and Answer Entity Extraction

Entity extraction is essentially a sequence labeling task that assigns words and phrases in a sentence sequence to their proper entity types. We use the Conditional Random Field model (CRF) [5], which is based on the Maximum Entropy model (MaxEnt) and the Hidden-Markov model (HMM). The formula is shown as follows:

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\{\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}_t)\}, \quad (1)$$

where $\mathbf{x}$ is the input/observed data sequence; $\mathbf{y}$ is the output/hidden label sequence; $\mathbf{x_t}$ is the input feature for word at position t; y_t is the output label of word at position t; K is the number of feature function; f_k is the k_{th} feature function; λ_k is the weight of the k_{th} feature function; and Z is a normalization factor of the form $\sum_{y} \exp\{\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}_t)\}$.

The feature function $f_k(y_t, y_{t-1}, \mathbf{x_t})$ is the key component of CRF model. It captures the co-occurrence between y_t and $\mathbf{x_t}$, which reflects the dependency between the output entity tag and the input feature of current word, and between y_t and y_{t-1}, which reflects the dependency between entity tags of two adjacent words. Note that $\mathbf{x_t}$ is a vector that can be a data sequence rather than a single value. Hence, by applying such feature function, it catches not only a large amount of observable knowledge from the input data sequence, but also the Markov chain dependency relationships between hidden entity tags that need to be inferred.

As λ_k is the parameter to be estimated, what we need to provide for the model is the alternative output entity tags for each word y_t, and the input features for each word $\mathbf{x_t}$. Hence, the CRF-based entity extraction requires two preceding steps: Entity Tag Design and Feature Extraction.

Entity Tag Design

Entity Types. Entity extraction is a multi-class classification task which assigns a label to each word in a sentence. In our study, entities are the words in questions and answers on QA websites. To identify the targeted classes to be extracted, we analyzed the QA data on three major Chinese QA services (39.net, xywy.com, and 120ask.com) and categorized the entities into the types shown in Table 3. Note that entity types frequently used in questions and answers are different, hence we need to extract them separately.

Table 3. Entity types in questions and answers

	Type	Tag	Examples
Question	Disease and symptom	d	Diabetes mellitus, renal failure
	Medicine and food	m	Plantago seed, melbine
	Diagnosis	c	Electrocardiogram, NMR
	Treatment	t	Laser therapy, cystectomy
	Organ	o	Kidney, heart
	Index	i	Body temperature, glucose
	Organ symptom	os	Pain, erythema
	Index description	is	A little high, normal
Answer	Disease and symptom	d	Diabetes mellitus, renal failure
	Medicine	m	Traditional Chinese or western medicine
	Food	f	Watermelon, millet porridge
	Diagnosis	c	Electrocardiogram, NMR
	Treatment	t	Laser therapy, cystectomy
	Organ	o	Kidney, heart
	Index	i	Body temperature, glucose

Equivalence Conjunctions. There exist some dependencies between entity tags of certain words. For example, entities around conjunction words such as "和 (and)", "或 (or)" and "及 (as well as)" are very likely to have the same tags. Similarly, the enumeration comma ("、"), a special Chinese punctuation, separates a series of terms of the same type. We call these words "equivalence conjunctions." While conjunction words in English also join two sentences, the above mentioned Chinese equivalence conjunctions usually join terms, which makes it appropriate for identifying entities with same tags. Equivalence conjunctions are frequently found in answer text where a list of symptoms, conditions, and medicines are named. The problem is that it is difficult to capture such information using a Markov model. According to the Markov property, the latter tag (y_t) is only influenced by the preceding tag (y_{t-1}), which is the tag of equivalence conjunctions instead of the tag of the preceding entity. To leverage such dependencies between entity tags, we design some tags for equivalence conjunctions in answer texts.

- Tag enumeration comma as "*dn*?", where "?" is determined by the tag of the preceding word.
- Tag equivalence conjunction words (e.g., "和", "或", "及") as "*l*?", where "?" is determined by the tag of the preceding word.

The following is an example of an answer containing enumerated Chinese medicine names, which are tagged correctly following the two rules.

- 建议在医生当面指导下服用颈复康颗粒\m 、\dnm 芬必得\m 、\dnm 舒筋活血片\m 进行治疗。
- Translation: I recommend you take Jingfukang Granules\m 、\dnm Fenbid\m 、\dnm Shujinhuoxue tablets\m under the direction of your physician.

Character Tags. A Chinese word may consist of several characters, e.g., "糖尿病" for "diabetes". One way for entity tagging is to perform Chinese word segmentation first, and then assign a tag for each word. However, the overall performance may be influenced by the segmentation accuracy because only words included in the dictionary in the segmentation tool can be identified and assigned a tag. Character tagging, which tags every character instead of words, is a common technique for addressing this problem [16]. Hence, we extend the existing tags by adding two prefixes ***B*** and ***I*** to each tag. For example, ***d*** is extended as ***Bd*** and ***Id***. If a character is tagged as ***Bd***, it is the beginning of some disease entity. If a character is tagged as ***Id***, it is in the middle or at the end of one entity. As tags listed in Table 3 are for health related entities only, we need four extra tags for words that don't fall into this category:

- ***Bf, If*** – negation words (e.g., no, not, without, etc.).
- ***D*** – the comma sign (,).
- ***J*** – the period sign (.).
- ***O*** – characters not included in any entity.

The major difference between our approach and previous studies in entity tag design is that, rather than only tagging terms, we also design tags for Chinese equivalence conjunctions, which help identify neighboring words that share same entity

types. Furthermore, not only words but also characters found in Chinese health QA contents are tagged, which helps extract more entities from the informal and noisy UGC texts.

Feature Extraction. Feature extraction is to extract feature values from the input data. Three types of features are used in our study: character features, lexical features, and part-of-speech features. *Character feature* is the character in a sentence. Take "心脏病" ("heart disease") for example: when the character "病" ("disease") appears in a word, it usually indicates a disease entity. *Lexical feature* is the class label (tag) assigned to each character based on the health lexicons. It is derived by mapping each word in a sentence to the lexicons, and then assigning a symbol to each character in the words by extending the class label with two prefixes ***B*** and ***I***. For example, as "心脏病" is included in the disease table of the lexicon, the lexical features (i.e., tags) of these three characters are "Bd, Id, Id". Intuitively, if a word is included in a health lexicon, it probably belongs to a given entity type. *Part-of-speech feature* is the part-of-speech tags derived from NLPIR, a Chinese NLP toolkit [1]. For example, as "苹果" ("apple") is tagged as a noun by NLPIR, the part-of-speech features of these characters are then "Bn, In". Generally, nouns are more likely to be the disease entities than adjectives or verbs.

We do not use the feature of current position as input directly. Still take "heart disease" for example, if we use the lexicon symbol of "heart", it will probably be labeled as an organ. But if we take into account "disease," which is the following word of it, it should be labeled as part of a disease entity. The idea behind it is that the entity type of current word y_t is not only determined by the feature tag of current word, but also influenced by the feature around the word. Hence, by varying the observation range, we can derive various input features. We define following $\mathbf{x_t}$ based on our analysis:

Character Feature. We denote C_0 as the current character, C_{-i} as the i_{th} character before the current character, C_i as the i_{th} character after the current character. Hence, $\mathbf{x}_t$ that considers character feature can be:

- C_0. This is the most common feature template. For example, "糖Bd尿Id病Id" ("sugar Bd urine Id disease Id").
- C_{-2} and C_{-1}. For example, in "服O用O二Bm甲Im双Im胍Im片Im" ("taking metformin tablets"), "二" (the first Chinese character of metformin) is tagged as "Bm" because "服用" ("taking") usually appears before medicine Bm.
- C_1 and C_2. For example, "格Bm列Im齐Im特Im胶Im囊Im" ("Gliclazide capsule"), "特" (the last Chinese character of Gliclazide) is tagged as "Im" because "胶囊" ("capsule") usually appears after medicine Im.
- Similar unigram features also used in our study are: C_{-2}, C_{-1}, C_1, C_2.
- Similar bigram features also used in our study are: C_{-1} and C_0, C_0 and C_1.

Lexical Feature. We denote L as the lexicon-based tag. Hence, $\mathbf{x_t}$ that considers lexical feature can be:

- L_0. For example, "hand" is tagged as "organ" as it is included in the organ table of the lexicons.
- L_{-1} and L_0, L_0 and L_1.

Part-of-Speech Feature. We denote P as the part-of-speech tag. Hence, $\mathbf{x_t}$ that considers part-of-speech feature can be:

- P_{-1}, P_0, P_1. For example, an adjective is often followed by a noun.
- P_{-1} and P_0, P_0 and P_1.

We use CRF++ [4] as our CRF implementation.

4 Experiments

Our research aims to improve the performance of CRF-based entity extraction approach from two aspects: extracting large sets of features considering characteristics of both Chinese and health related QA, and leveraging equivalence conjunctions to identify entity of the same types. To validate the effectiveness of proposed approach, we first compare the performance of question entity extraction using different features, and then the performance of answer entity extraction including equivalence conjunctions or not. Both experiments reflect the performance and design rationality of our approach.

4.1 Data

We collected medical and health related QA data from three major Chinese QA service websites: 39.net, xywy.com, and 120ask.com. On these websites, users can ask questions and get answers and advice from real physicians and healthcare professionals. One service, 39.net, also allows other users to answer questions. 1.27 million questions and 2.26 million answers were collected in total. The date range was between July 2007 and May 2015. We then randomly sampled 1,112 questions and 1,266 answers from the dataset and manually tagged each sentence with predefined entity types. Ten-fold cross validation was conducted based on these labeled data.

4.2 Evaluation Metrics

For a N classification problem, we build a $N \times N$ confusion matrix C, where $C(i,j)$ is the number of label j predicted as i. Based on this matrix, we calculate the precision, recall, F-Measure for each class.

- Precision of i: $\mathrm{P} = \frac{C(i,i)}{\sum_j C(i,j)}$
- Recall of i: $\mathrm{R} = \frac{C(i,i)}{\sum_j C(j,i)}$
- F-Measure of i: $\mathrm{F} = \frac{\left(a^2+1\right)PR}{a^2(P+R)}$, (when a = 1, it degrades to F1-Score - $\mathrm{F}_1 = \frac{2PR}{P+R}$)

We use the average values of all indexes as performance metrics of the classifier.

4.3 Question Entity Extraction

In the question entity extraction process, we used three types of features: character features, lexical features, and part-of-speech features. As character features are the most common feature and are usually included in entity extraction, we designed four experiments by combining features on the condition that character features are used:

- Experiment (1): character features only;
- Experiment (2): character features and part-of-speech features combined;
- Experiment (3): character features and lexical features combined;
- Experiment (4): all three types of features combined.

The performance of the above combinations is evaluated as shown in Table 4. We first explore the effectiveness of using the three features. From the most commonly used character feature (experiment 1) to the combination of the three (experiment 4), performance of all types of entity extraction is improved except index description and precision of treatment. We then explore the effects of different features. The comparison between experiment (2) and (4) suggests that the inclusion of lexical features significantly improves certain performance (precision, recall, or F1-score) for entity extraction of disease, medicine, diagnosis, organ, index, and the average. The comparison between experiment (3) and (4) suggests that the inclusion of POS features significantly improves certain performance for entity extraction of diagnosis, organ, index, and organ symptom. Overall, the inclusion of character, lexicon and POS feature significantly improve the average performance of question entity extraction. From pure character feature to the combination of three features, F1-score increased from 0.79 to 0.82.

Table 4. The performance of different feature combinations

Entity Type	Metric	(1) Char.	(2) Char. + POS	(3) Char. + Lex.	(4) Char. + POS + Lex.
Disease	Precision	0.936	0.936	0.956	**0.955**$^{**\ \&\&}$
	Recall	0.899	0.908	0.939	**0.944**$^{**\&\&}$
	F1-score	0.917	0.922	0.947	**0.949**$^{**\&\&}$
Medicine	Precision	0.914	0.894	0.935	0.923
	Recall	0.796	0.838	0.801	**0.815**$^{\&}$
	F1-score	0.849	0.863	0.862	**0.864***
Diagnosis	Precision	0.870	0.894	0.858	0.912
	Recall	0.619	0.615	0.580	**0.635**$^{\#}$
	F1-score	0.707	0.715	0.675	**0.739**$^{\&\#}$
Treatment	Precision	0.833	0.821	0.817	0.816
	Recall	0.614	0.633	0.627	0.634
	F1-score	0.702	0.712	0.702	0.709
Organ	Precision	0.870	0.859	0.869	0.880
	Recall	0.621	0.685	0.786	**0.814**$^{**\&\&\#\#}$
	F1-score	0.721	0.759	0.823	**0.844**$^{**\&\&\#}$

(*Continued*)

Table 4. (*Continued*)

Entity Type	Metric	(1) Char.	(2) Char. + POS	(3) Char. + Lex.	(4) Char. + POS + Lex.
Index	Precision	0.936	0.932	0.952	**0.944**$^{*\ \&\&\ \#\#}$
	Recall	0.929	0.933	0.926	0.930
	F1-score	0.932	0.932	0.938	0.937
Organ symptom	Precision	0.765	0.784	0.785	0.786
	Recall	0.529	0.584	0.547	**0.601**$^{**\#\#}$
	F1-score	0.622	0.668	0.641	**0.678**$^{**\#\#}$
Index description	Precision	0.843	0.807	0.827	**0.803***
	Recall	0.698	0.688	0.680	**0.676***
	F1-score	0.759	0.738	0.742	0.730
Average	Precision	0.875	0.872	0.880	**0.883**$^{**\ \&\&}$
	Recall	0.742	0.762	0.763	**0.781**$^{**\&\&\#\#}$
	F1-score	0.794	0.806	0.809	**0.823**$^{**\&\&\#\#}$

Note: Comparison between experiment (1) and (4): * $p<0.05$; ** $p<0.01$; comparison between (2) and (4): $^{\&}$ $p<0.05$; $^{\&\&}$ $p<0.01$; comparison between (3) and (4): $^{\#}$ $p<0.05$; $^{\#\#}$ $p<0.01$.

4.4 Answer Entity Extraction

For answer entity extraction, we add a new tag - equivalence conjunction. As shown in Table 5, equivalence conjunction improves the performance of all types of the entity extractions except the recall of diagnosis, treatment, and organ, and significantly improves certain performance (Precision, Recall, or F1-score) of disease, medicine, organ, index, and the average. The reason for the decrease is that in the test set, the equivalence conjunctions are seldom used for these three entities, and, when equivalence conjunctions are considered, characters with no equivalence conjunction around them are less likely to be tagged as entities. Overall, the use of equivalence conjunction can significantly improve the precision and F1-score of answer entity extraction.

Table 5. The performance of equivalence conjunction

Entity type	Metric	No equivalence conjunction	With equivalence conjunction
Disease	Precision	0.913	**0.920****
	Recall	0.866	0.867
	F1-score	0.888	**0.892***
Medicine	Precision	0.896	0.906
	Recall	0.846	0.849
	F1-score	0.869	**0.876***
Food	Precision	0.869	0.872
	Recall	0.771	0.781
	F1-score	0.813	0.821

(*Continued*)

Table 5. (*Continued*)

Entity type	Metric	No equivalence conjunction	With equivalence conjunction
Diagnosis	Precision	0.887	0.888
	Recall	0.808	0.803
	F1-score	0.842	0.842
Treatment	Precision	0.828	0.833
	Recall	0.726	0.718
	F1-score	0.771	0.769
Organ	Precision	0.682	**0.709***
	Recall	0.587	0.584
	F1-score	0.623	0.636
Index	Precision	0.923	**0.932****
	Recall	0.891	0.895
	F1-score	0.906	**0.913***
Average	Precision	0.869	**0.877***
	Recall	0.809	0.809
	F1-score	0.835	**0.839****

* $p < 0.05$;. ** $p < 0.01$.

5 Conclusion and Future Work

It is important to build automatic health QA systems to help people get high-quality answers to their concerns. This paper presents an entity extraction approach based on CRF, which considers both QA and Chinese characteristics by entity tag design and feature extraction. To recognize the entities, we create Chinese Chronic Disease lexicons based on expert knowledge and Web resources. Our experiments demonstrate the effectiveness of our approach. In the future, we plan to increase the size of Chinese chronic condition lexicon, add more training and test samples, and refine our entity extraction approach using different classification models and more feature types.

Acknowledgments. This work was supported by the National High-tech R&D Program of China (Grant No. SS2015AA020102), National Basic Research Program of China (Grant No. 2011CB302302), the 1000-Talent program, and the Tsinghua University Initiative Scientific Research Program. We thank the research assistance provided by Qingbo Cao at Tsinghua University.

References

1. Big Data Search and Mining Lab, BIT.: Natural Language Processing and Information Retrieval Sharing Platform. http://www.nlpir.org/
2. Cao, Y., Liu, F., Simpson, P., Antieau, L., Bennett, A., Cimino, J.J., Ely, J., Yu, H.: Askhermes: an online question answering system for complex clinical questions. J. Biomed. Inform. **44**(2), 277–288 (2011)

3. Keretna, S., Lim, C.P., Creighton, D.C., Shaban, K.B.: Enhancing medical named entity recognition with an extended segment representation technique. Comput. Methods Programs Biomed. **119**(2), 88–100 (2015)
4. Kudo, T.: CRF++: Yet Another CRF toolkit. http://taku910.github.io/crfpp/
5. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001)
6. Lee, M., Cimino, J., Zhu, H.R., Sable, C., Shanker, V., Ely, J., Yu, H.: Beyond information retrieval—medical question answering. In: AMIA Annual Symposium Proceedings, vol. 2006, p. 469. American Medical Informatics Association (2006)
7. McCallum, A., Li, W.: Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003, vol. 4, pp. 188–191. Association for Computational Linguistics (2003)
8. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes **30**(1), 3–26 (2007)
9. National health and family planning commission of the people's republic of China (2015). http://www.nhfpc.gov.cn/
10. Nlm.nih.gov: Unified Medical Language System (UMLS). http://www.nlm.nih.gov/research/umls/
11. Pasca, M., Lin, D., Bigham, J., Lifchits, A., Jain, A.: Organizing and searching the world wide web of facts-step one: the one-million fact extraction challenge. In: AAAI, vol. 6, pp. 1400–1405 (2006)
12. Pasupat, P., Liang, P.: Zero-shot entity extraction from web pages. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014. Long Papers, Baltimore, MD, USA, 22–27 June 2014, vol. 1, pp. 391–401 (2014)
13. Peng, X.Y., Chen, Y., Huang, Z.W.: A Chinese question answering system using web service on restricted domain. In: 2010 International Conference on Artificial Intelligence and Computational Intelligence (AICI), vol. 1, pp. 350–353. IEEE (2010)
14. Shaalan, K.: A survey of arabic named entity recognition and classification. Comput. Linguis. **40**(2), 469–510 (2014). http://dx.doi.org/10.1162/COLI_a_00178
15. Zhang, H., Xu, S., Li, W., Zhu, L.: XML-based document retrieval in Chinese diseases question answering system. In: (Jong Hyuk) Park, J.J., Adeli, H., Park, N., Woungang, I. (eds.) Mobile, Ubiquitous, and Intelligent Computing. LNEE, vol. 274, pp. 211–217. Springer, Heidelberg (2014)
16. Zhao, H., Kit, C.: Unsupervised segmentation helps supervised learning of character tagging for word segmentation and named entity recognition. In: IJCNLP, pp. 106–111. Citeseer (2008)