Journal of Biomedical Informatics 47 (2014) 91–104

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Supervised methods for symptom name recognition in free-text clinical records of traditional Chinese medicine: An empirical study

Yaqiang Wang [a], Zhonghua Yu [a,*], Li Chen [a], Yunhui Chen [b], Yiguang Liu [a], Xiaoguang Hu [c], Yongguang Jiang [d]

[a] Department of Computer Science, Sichuan University, Chengdu, Sichuan 610064, PR China
[b] School of Fundamental Medicine, Chengdu University of Traditional Chinese Medicine, Chengdu, Sichuan 610075, PR China
[c] No. 1 Clinical Hospital, Beihua University, Jilin, Jilin 132011, PR China
[d] Department of Preclinical Medicine, Chengdu University of Traditional Chinese Medicine, Chengdu, Sichuan 610075, PR China

ARTICLE INFO

Article history:
Received 10 November 2012
Accepted 13 September 2013
Available online 23 September 2013

Keywords:
Symptom name recognition
Free-text clinical records
Traditional Chinese medicine
Supervised sequence classification
Natural language processing

ABSTRACT

Clinical records of traditional Chinese medicine (TCM) are documented by TCM doctors during their routine diagnostic work. These records contain abundant knowledge and reflect the clinical experience of TCM doctors. In recent years, with the modernization of TCM clinical practice, these clinical records have begun to be digitized. Data mining (DM) and machine learning (ML) methods provide an opportunity for researchers to discover TCM regularities buried in the large volume of clinical records. There has been some work on this problem. Existing methods have been validated on a limited amount of manually well-structured data. However, the contents of most fields in the clinical records are unstructured. As a result, the previous methods verified on the well-structured data will not work effectively on the free-text clinical records (FCRs), and the FCRs are, consequently, required to be structured in advance. Manually structuring the large volume of TCM FCRs is time-consuming and labor-intensive, but the development of automatic methods for the structuring task is at an early stage. Therefore, in this paper, symptom name recognition (SNR) in the chief complaints, which is one of the important tasks to structure the FCRs of TCM, is carefully studied. The SNR task is reasonably treated as a sequence labeling problem, and several fundamental and practical problems in the SNR task are studied, such as how to adapt a general sequence labeling strategy for the SNR task according to the domain-specific characteristics of the chief complaints and which sequence classifier is more appropriate to solve the SNR task. To answer these questions, a series of elaborate experiments were performed, and the results are explained in detail.

1. Introduction

Traditional Chinese medicine (TCM) provides a distinctive way to perceive the human body and is becoming a medical theory complementary to Western medicine [1–4]. TCM knowledge is held largely in the minds of clinically experienced TCM doctors. Consequently, the clinical records of TCM, which are documented by TCM doctors during their routine diagnostic work, naturally constitute an abundant clinical knowledge source of TCM that can be inherited by the next generation of practitioners. However, with the increase in the accumulation of clinical records, comprehensive summarization of complicated TCM regularities is difficult. Fortunately, with the modernization of TCM clinical practice, clinical records have begun to be digitized. This provides an opportunity for researchers to discover, with the help of data mining (DM) and machine learning (ML) methods, TCM regularities buried in the large volume of digitized clinical records.

There has been some work on applying DMs and MLs to TCM knowledge discovery, such as discovering TCM knowledge from well-structured literature by using Bayesian networks [5,6] and establishing TCM expert systems for decision support by the naïve Bayes classifier based on a limited amount of manually structured data [7]. However, the contents of most fields in the clinical records, e.g. the chief complaints, are unstructured (or free-text). The result is that the methods that are verified on the well-structured data, cannot be directly applied to knowledge discovery in the free-text clinical records (FCRs) of TCM. These methods, consequently, require FCRs to be structured in advance.

Manually structuring the large volume of TCM FCRs is tedious, time-consuming, and labor intensive. Hence there is an urgent need for the development of an effective method to automatically structure the FCRs, i.e. recognizing medical named entities in FCRs [8]. Named entity recognition (NER) in general text has been

* Corresponding author.
E-mail addresses: wangyaq2204_cn@hotmail.com (Y. Wang), yuzhonghua@scu.edu.cn (Z. Yu), cl@scu.edu.cn (L. Chen), tcmhero@126.com (Y. Chen), lygpapers@yahoo.com.cn (Y. Liu), 21224498@qq.com (X. Hu), cdtcm@163.com (Y. Jiang).

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.09.008

widely studied in the natural language processing (NLP) community [9]. For example, a hybrid Chinese NER model based on multiple features was proposed in [10], and the model was evaluated on the general text dataset called "People's Daily". In [11], a lexicalized hidden Markov model (HMM) approach to NER was designed and validated on "newswire" data, which is also a general text dataset. In addition, a pragmatic approach to Chinese word segmentation was proposed in [12]. This approach was implemented in an adaptive Chinese word segmenter (MSRSeg), which can simultaneously segment general Chinese text and perform NER. However, FCRs of TCM are different from general text. They have domain-specific characteristics [13] and therefore the methods designed for NER in general text might need a domain-specific adaptation for NER in the FCRs.

Medical information extraction in English FCRs of Western medicine has become a topic of great interest in recent years, and has been extensively studied due to the efforts of the Informatics for Integrating Biology and the Bedside (i2b2) project, which has released clinical record datasets that can be used as gold standards by the medical NLP research community. In 2009, i2b2 organized a medical information extraction challenge on extracting medications, dosages, modes, durations, etc. from the English discharge summaries [14]. In the following year, a medical problem, test, and treatment concept extraction challenge was organized by i2b2 [15]. Subsequently, based on the public clinical record datasets, a series of excellent work on NER in English discharge summaries of Western medicine has been published. These NER tasks are usually treated as a sequence labeling problem, and then the open-domain sequence classifiers, e.g. Conditional Random Field model (CRF), are adapted to the medical domain [16–26] with the help of domain knowledge and domain-specific sources. For example, the domain vocabulary used in [16], the domain-specific rules in [17], and the knowledge-rich sources utilized in [18–20] have been used for these purposes. Because of the differences between English and Chinese [27] and owing to the distinctive characteristics of TCM FCRs [13], methods, that could be borrowed from English NER in the discharge summaries of Western medicine need adaptation for NER in FCRs of TCM.

The development of NER in FCRs of TCM has fallen behind the progress of English NER in FCRs of Western medicine. NER in the TCM community was first attempted in 2012 [13]. Symptom name recognition (SNR) in the chief complaints, which is one of the most important tasks of NER in FCRs of TCM, was accomplished by the bigram-based dictionary-matching method. However, various literal forms of symptoms were generated during the routine diagnostic work of TCM doctors [28]. Consequently, the application of the dictionary-matching method in clinical practice requires maintenance of a symptom name dictionary. However, the dictionary-matching method is laborious, making it less appropriate for use in practice. SNR in the chief complaints was also studied in [29]. Based on method for English NER in discharge summaries of Western medicine, SNR in chief complaints was treated directly as a sequence labeling problem and solved by CRF with two types of useful features. This preliminary work still leaves many questions waiting to be answered, such as:

(1) Is it suitable to treat SNR in the chief complaints as a sequence labeling problem? In other words, are there any domain-specific characteristics that would facilitate SNR completion by the sequence classifiers?
(2) The chief complaints in TCM FCRs have domain-specific characteristics. Some domain-specific adaptation to the general sequence labeling strategy for the SNR task might be needed. How can an appropriate adaptation be made?
(3) Several sequence classifiers, e.g. HMM, maximum entropy Markov model (MEMM), and CRF, can be used to solve the sequence labeling problem. Each sequence classifier has its own specializations. Which is most suitable to SNR and which can achieve the best performance?

To answer these questions, we focus our attention in this paper on studying SNR in chief complaints. First, the SNR task is treated as a sequence labeling problem reasonably. Second, a new sequence labeling strategy is designed for SNR in the chief complaints based domain-characteristics. These approaches are introduced in Section 2. In Section 3, three typically supervised sequence classifiers (HMM, MEMM, and CRF) are applied to the SNR task with an empirical analysis. Elaborate experiments are performed and described in Section 4 aiming to answer the previously raised questions. Finally, Sections 5 and 6, respectively, provide further discussion and conclusions.

2. Sequence labeling for the SNR task

2.1. Why sequence labeling?

Symptom names in TCM are usually composed of three aspects of descriptions: body location, sensation, and intensity. For example, the symptom name "头痛剧烈" (a severe headache) consists of three aspects of descriptions including a body location "头" (head), a sensation "痛" (ache), and a intensity "剧烈" (severe).

These three aspects preferably appear in sequence and should not be split by other descriptions (e.g. possessives or temporal descriptions). For instance, the sensation "晕" (dizziness) should come after the body location "头", but the temporal description "昨天" (yesterday), which is used to indicate the occurring time of the symptom, should be written before the symptom name "头晕" (dizziness) rather than in between the body location and the sensation.

Furthermore, technically, words used in the symptom names of TCM should have different distributions from other words that are used outside of symptom names (see Fig. 1 as an example). Therefore, it is appropriate to treat SNR in the chief complaints as a sequence labeling problem; the characteristics described above should facilitate the completion of SNR by the sequence classifiers.

2.2. Domain-specific adaptation to the general sequence labeling strategy

The commonly-used sequence labeling strategy for NER in general text or in English discharge summaries of Western medicine is to label each word (defined as a "labeling unit") in each sentence (defined as a "labeled sequence") with a predefined tag that is used to indicate the role of the labeling unit, e.g. that it is a beginning, an

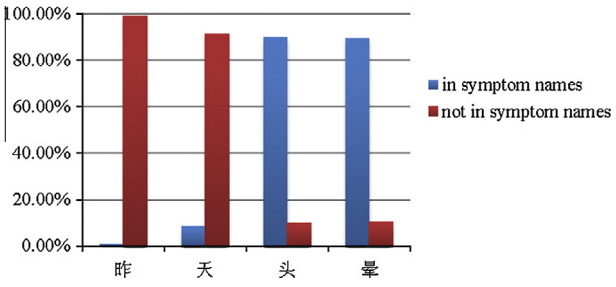

Fig. 1. An example of the differences between the distributions of the words used in symptom names and outside of symptom names.

intermediate, or an outside part of the named entity, usually symbolized by "B", "I", and "O" respectively [30]. However, chief complaints in TCM FCRs have domain characteristics [13]. The general sequence labeling strategy might need an adaptation for SNR in chief complaints. Therefore, in this section, we describe the adaptation of general sequence labeling strategy, based on several empirical reasons.

2.2.1. Adaptation to the labeling unit

Generally, labeling units are words in the text [30]. However, there are no natural whitespaces between Chinese words in general text, including chief complaints. First, a Chinese word segmentation task should be completed before recognizing symptom names in the chief complaints based on sequence labeling [12]. Unfortunately, the Chinese word segmentation methods designed for general text will not work effectively for segmenting the chief complaints into words due to the distinctive characteristics of the chief complaints [13]. Thankfully, Chinese is different from Western languages in that its characters can bear specific meanings. Therefore, in this paper, we change the labeling units from Chinese words to Chinese characters in the chief complaints.

2.2.2. Adaptation to the labeled sequence

The chief complaints in FCRs of TCM are documented by TCM doctors during their routine diagnostic work through Physician Visit procedures. In order to improve work efficiency, TCM doctors directly note the physical conditions described by patients without organizing the content and rewriting the text, which loses contextual coherence. For example, a more coherent description of the physical conditions "昨日肠鸣, 失气多, 心中不适" (Yesterday, the patient had borborygmus and more flatulence, and his/her heart was uncomfortable) may described by different patients in different incoherent forms, such as "昨日肠鸣, 心中不适, 失气多" (Yesterday, the patient had borborygmus, his/her heart was uncomfortable, and the patient had more flatulence) or "肠鸣, 心中不适, 失气多 (昨天)" (The patient had borborygmus, his/her heart was uncomfortable, and the patient had more flatulence (all these symptoms appeared yesterday)). They would be recorded by TCM doctors directly in the chief complaints section of the record.

Moreover, various kinds of punctuations can appear in chief complaints, and they are used by TCM doctors without any standard criteria. For instance, TCM doctors may like to use a comma instead of all other punctuations (e.g. the period, which is used to divide natural sentences) for convenience. The result is that several natural sentences in chief complaints are combined into one. Taking the chief complaint "昨日肠鸣, 失气多, 心中不适, 早晨大便提早, 头昏." (Yesterday, the patient had borborygmus, his/her heart was uncomfortable, and had more flatulence. This morning, the patient had a bowel movement earlier than before and felt dizziness.) as an example, it should be two natural sentences "昨日肠鸣, 失气多, 心中不适." (Yesterday, the patient had borborygmus, his/her heart was uncomfortable, and had more flatulence.) and "早晨大便提早, 头昏." (This morning, the patient had a bowel movement earlier than before and felt dizziness.), but the period, which is used to represent the end of a sentence, is replaced by a comma.

For these reasons, general labeled sequences (sentences in the text) need a domain-specific adaption. We define *clauses* in the chief complaints as the labeled sequences instead of the sentences. The clauses in this paper are groups of sequential Chinese characters separated by commas in the chief complaints. For example, a chief complaint "昨日肠鸣, 心中不适, 失气多" (Yesterday, the patient had borborygmus and more flatulence, and his/her heart was uncomfortable) will be divided into three clauses "昨日肠鸣" (the patient had borborygmus), "心中不适" (his/her heart was uncomfortable), and "失气多" (the patient had more flatulence). This adaptation would bring about two advantages:

(1) Clauses in chief complaints as labeled sequences keep the internal coherence of every clause and, at the same time, reduce the possibility of incoherent sentences occurring.
(2) Defining the clauses as labeled sequences is similar to empirical feature selection. It reduces the number of noisy features extracted from the disorganized context in the chief complaints.

2.2.3. Adaptation to the predefined tag set

The predefined tag set is usually {"B", "I", "O"}, denoted as BIO. However, according to the results reported in [29], the end of symptom names in the chief complaints is more difficult to identify. Therefore, we consider defining a specific tag, e.g. "E", to individually identify the end of the symptom names in order to improve sequence labeling performance. Thus the predefined tag set is adapted to {"B", "I", "E", "O"}, denoted as BIEO.

2.3. Formulation

Referring to the introduced domain-specific adaptation to general sequence labeling strategy, SNR in chief complaints can be naturally defined as follows:

Given a clause $\mathbf{x} = x_1, x_2, \ldots, x_n$ in a chief complaint, the goal is to construct a sequence classifier to accurately label each x_i, which is the ith Chinese character in $\mathbf{x}$, with the most reliable predefined tag y_i, where y_i is one of the elements in BIEO. Accordingly, the most reliable tag sequence $\hat{\mathbf{y}} = \hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ given $\mathbf{x}$ would be:

$$\hat{\mathbf{y}} = \arg\max_y P(\mathbf{y}|\mathbf{x}) \quad (1)$$

This problem can be effectively solved by the supervised sequence classifiers.

3. Supervised sequence classifiers for the SNR task

In Section 2, SNR in the chief complaints is reasonably treated as a sequence labeling problem and, at the same time, a new sequence labeling strategy that is adapted from the general sequence labeling strategy based on several empirical reasons is proposed for the SNR task. To complete the SNR task, supervised sequence classifiers are employed. Different sequence classifiers have their own specialties. In this section, we aim to compare three typically supervised sequence classifiers (HMM, MEMM, and CRF) and introduce them to the SNR task.

3.1. HMM for the SNR task

HMM is a statistical structure with stochastic state transitions (e.g. a transition from a hidden state to another hidden state, where the hidden states refer to the corresponding predefined tags of the labeling units in the SNR task) and observation generation processes (e.g. an observed Chinese character "气" (pneuma) in the clause "失气多" could be generated from a hidden state "I"). HMM can be flexibly used to solve sequence labeling task [31]. According to the definition of HMM, Eq. (1) can be modified to:

$$\hat{\mathbf{y}} = \arg\max_y P(y_0)\prod_{i=1}^{n+1} P(y_i|y_{i-1})\prod_{i=1}^{n} P(x_i|y_i) \quad (2)$$

where y_0 and y_{n+1} refer to a default start tag "START" and a default end tag "STOP" respectively, $P(y_0)$ equals to 1, $P(y_i|y_{i-1})$ represents the state transition probability, and $P(x_i|y_i)$ is the observation generation probability. The state transition probabilities and the observation generation probabilities can be estimated through the Maximum Likelihood method based on a training dataset. And then $\hat{\mathbf{y}}$ can be efficiently predicted by using the Viterbi algorithm based

on the estimated state transition probabilities and observation generation probabilities [31].

3.2. MEMM for the SNR task

HMM models SNR in the chief complaints as a probabilistic generative process in which the Chinese characters in the chief complaints are sequentially generated by their corresponding hidden tags. However, considering the SNR task from the perspective of simulating the human labeling process, the corresponding tag of an observed Chinese character is determined by a human labeler through by observing the chief complaint, finding and analyzing the evidence that can help the labeler to judge which predefined tag is the best choice of a current Chinese character, and then labeling it with the best tag. In other words, there many useful features (e.g. the evidence mentioned previously) can be utilized to help the sequence classifiers to complete the SNR task. Unfortunately, generative models (e.g. HMM) are known to be unable to conveniently incorporate additional features, whereas discriminative models are known to have an advantage in this respect. Therefore, the typical discriminative sequence classifier MEMM [32] is investigated. It is an extension of HMM in which the state transition probability and the observation generation probability in HMM are replaced by a single discriminative probability $P(y_i|\mathbf{x}, y_{i-1})$. This discriminative probability allows MEMM to conveniently incorporate additional features into the sequence labeling procedure.

Mathematically, $P(y_i|\mathbf{x}, y_{i-1})$ is a conditional probability of current tag y_i given its previous tag y_{i-1} and the observed Chinese characters in $\mathbf{x}$. Then, Eq. (1) can be redefined as:

$$\hat{\mathbf{y}} = \underset{y}{\operatorname{argmax}} \prod_{i=1}^{n} \frac{\exp\left(\sum_{j=1}^{N} \omega_j f_j(\mathbf{x}, y_{i-1})\right)}{\sum_{y_i=t}^{BIEO} \exp\left(\sum_{j=1}^{N} \omega_j f_j(\mathbf{x}, y_{i-1})\right)} \tag{3}$$

where y_0 refers to a default start tag "START" of every labeled sequence, n is the length of $\mathbf{x}$, ω_j is the weighting parameter of the feature $f_j(\mathbf{x}, y_{i-1})$, $j \in [1, \ldots, N]$, and N represents the number of features incorporated. The features are fetched from $\mathbf{x}$ and the previous tag y_{i-1} of y_i, and y_i is the potential corresponding tag of x_i. In Eq. (3), t is one of the predefined tags chosen from BIEO.

The key to MEMM is to learn the weighting parameter ω_j for each feature. This can be efficiently solved using quasi-Newtonian methods, such as L-BFGS, gradient descent, conjugate gradient descent, and various iterative scaling algorithms [33]. In this paper, L-BFGS [34] is employed. Finally, $\hat{\mathbf{y}}$ can also be predicted by using the Viterbi algorithm efficiently based on the learned weighting parameters [31].

3.3. CRF for the SNR task

Compared to HMM, MEMM has an advantage in incorporating useful features into the sequence labeling procedure. However, MEMM suffers from the label bias problem [35]. This problem is serious if the hidden states have a small number of possible outgoing transitions. In other words, the hidden states would like to make light of their corresponding observations when they have low-entropy next-state distributions [35]. As shown in Fig. 2, SNR in the chief complaints by MEMM is, unfortunately, a victim of the label bias problem. In order to cope with the label bias problem in MEMM, we employ CRF [35] which is an undirected graphical model. To avoid the label bias problem, CRF represents the conditional distribution over $\mathbf{y}$ *globally* conditioned on $\mathbf{x}$. Therefore, Eq. (1) can be modified to:

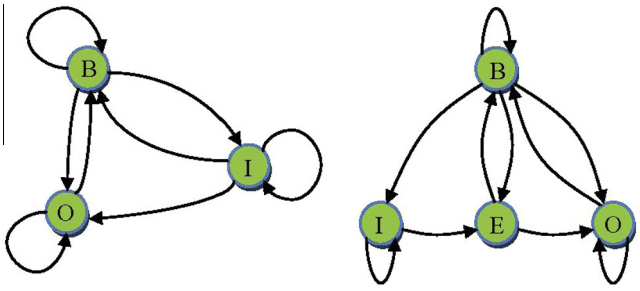

Fig. 2. State transition graphs of two predefined tag sets BIO (left) and BIEO (right) of SNR in the chief complaints. States in both of these graphs both have a small number of outgoing transitions, and thus they are likely to have label bias problem.

$$\hat{\mathbf{y}} = \arg\max_{y} \frac{\exp\left(\sum_{i=1}^{n}\sum_{k=1}^{K} \lambda_k f_k(y_{i-1}, y_i, y_{i+1}, \mathbf{x}, i)\right)}{\sum_{TS} \exp\left(\sum_{i=1}^{n}\sum_{k=1}^{K} \lambda_k f_k(y_{i-1}, y_i, y_{i+1}, \mathbf{x}, i)\right)} \tag{4}$$

where f_k is an indicator feature function, $k \in [1, K]$, and K is the number of global features. Moreover, λ_k is a weighting parameter of f_k that would be learned based on a training dataset. *TS* is the number of all possible tag sequences. In this paper, CRF is implemented by the CRF++ tool [36] for estimating the weights, and the default settings are used.

4. Experimental settings

4.1. Features used by MEMM and CRF

Commonly used features are words and other higher level syntactic features. However, in Section 2 we show that splitting the chief complaints into Chinese words, let alone extracting the higher level syntactic features form the chief complaints, is not a trivial task due to the distinctive characteristics of the chief complaints [13]. Therefore, referring to [29], two types of empirical features are utilized in this paper and explained below.

4.1.1. Chinese character n-gram features

Contextual words are the obvious and informative features that can be utilized by MEMM and CRF for SNR in the chief complaints. However, as previously noted, splitting the chief complaints into Chinese words is not trivial. Fortunately, Chinese characters bear specific meanings and are viable alternatives to Chinese words as the features. Furthermore, because the content of the chief complaints recorded by TCM doctors is concise [13] and the average

Diagnostic Data	Patient Name	Gender	Age	Chief Complaint	Diagnostic	Formulas	Drugs
2004/4/19	[redacted]	女	62	昨日肠鸣，矢气多，心中不适，早晨大便提早，头昏，苔薄，足转筋，脉细.	脾肾阳虚	桂附理中汤	桂枝，附子，党参，白术，干姜，广香，砂仁，黄芪，补骨脂，肉豆蔻，吴茱萸，草果，甘草，鹿角霜
2004/6/3	[redacted]	女	21	小便镜检血尿2-3+，血尿，尿频时作，小腹胀坠，腰酸，舌尖暗红，脉弦	肝脾气陷	丹栀逍遥散	丹皮，栀子，柴胡，白芍，茯苓，薄荷，早连草，白茅根，生地，黄芪，补骨脂，炒蒲黄，甘草，益母
2004/6/8	[redacted]	男	38	胃脘胀满，引两胁胀，早晨4-5点尤甚，引胸痛，容易气郁，矢气，心中烧灼感，不吐酸，不欲吃冷食，口苦，身软，睡眠差，小便正常，苔略厚腻，脉弦.	寒热错杂	半夏泻心汤	半夏，党参，黄芩，黄连，炮姜，广香，枳壳，砂仁，乌贼骨，台乌，淮山，白术，甘草.
2008/10/6	[redacted]	男	60	长期从事跑步，脚跟痛 脚后跟痛，上肢冷，耳鸣，目干，舌淡，脉缓 下肢静脉曲张/15年前手术	血虚，血瘀	桃红四物	桃仁1.5，红花1，当归2，川芎2，白芍2，生地2，黄芪3，桂枝1.5，生姜1.5，大枣1.5，炙草1

Fig. 3. An excerpt of chief complaints from the CRD.

Table 1
Detailed information of the training and test datasets.

	CRD-BIO/CRD-BIEO	
Number of unique symptom names	9129	
	Training data	Test data
Number of unique symptom names	3630	7693
Amount of symptom names	17665	53485
Number of each type of tags	17665 ('B'), 14725 ('I'), 17556 ('I/E'), 53331 ('O')	53485 ('B'), 44627 ('I'), 53171 ('I/E'), 158026 ('O')

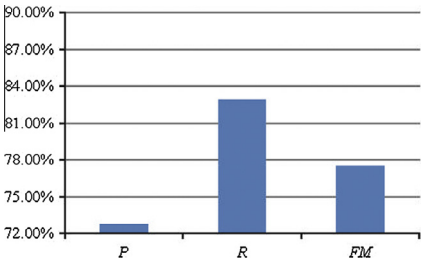

Fig. 4. P_{rec}, R_{rec}, and FM_{rec} obtained by HMM based on the general sequence labeling strategy.

length of Chinese words approximates 2 [37], Chinese character unigrams, bigrams, and trigrams cover most of the meaningful words and some other useful features in the chief complaints, e.g. the prefixes and the suffixes of symptom names, etc. Therefore, we choose Chinese character unigrams, bigrams, and trigrams as the features used by MEMM and CRF in this paper, and they are denoted by the symbols "U", "B", and "T" respectively.

4.1.2. Position features

Empirically, TCM doctors prefer initially to record background information about the patients' symptoms, e.g. the temporal information, and then note the patients' symptom information concisely. This means that symptom names are not likely to appear at the beginning of the sentences in the chief complaints. Instead, they appear more frequently at the beginning of the subsequent

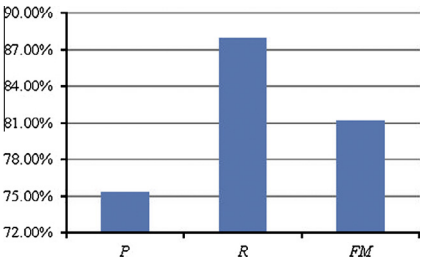

Fig. 7. P_{rec}, R_{rec}, and FM_{rec} obtained by HMM based on the domain-specific adapted sequence labeling strategy which makes an adaptation to the labeled sequence.

clauses in the sentences. For example, the chief complaint "昨日胁鸣, 失气多, 心中不适." (Yesterday, the patient had borborygmus and more flatulence, and his/her heart was uncomfortable) starts with a temporal background description (i.e. "昨日" (yesterday)), and then the following clauses in this chief complaint all start with the symptom names, i.e. "失气多" (more flatulence) and "心中不适" (his/her heart was uncomfortable). Therefore, the positions of the Chinese characters in the chief complaints are useful features for the SNR task. In this paper, the position features are represented as "[*SubSID-PosID*]", where "*SubSID*" is the index of current clause in a chief complaint and "*PosID*" indicates the position of current Chinese character in the clause *SubSID*.

4.2. Evaluation metrics

To evaluate the feasibility of the adapted sequence labeling strategy and the performance of the introduced sequence classifiers for SNR in the chief complaints, two groups of evaluation metrics are employed. The first group is symptom name recognition precision (P_{rec}), recall (R_{rec}), and F-Measure (FM_{rec}). The second group is labeling precision (P_{lab}), recall (R_{lab}), and F-Measure (FM_{lab}) of each predefined tag.

4.2.1. Definitions of P_{rec}, R_{rec}, and FM_{rec}

P_{rec}, R_{rec}, and FM_{rec} are used to comprehensively assess the feasibility of the sequence labeling strategy with a domain-specific adaptation and appraise the performance of HMM, MEMM and CRF for the SNR task. They are formulated as follows.

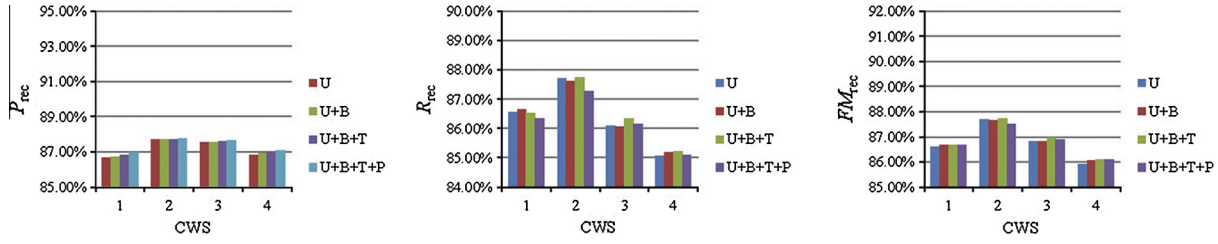

Fig. 5. P_{rec}, R_{rec}, and FM_{rec} obtained by MEMM with different features under different CWS settings based on the general sequence labeling strategy.

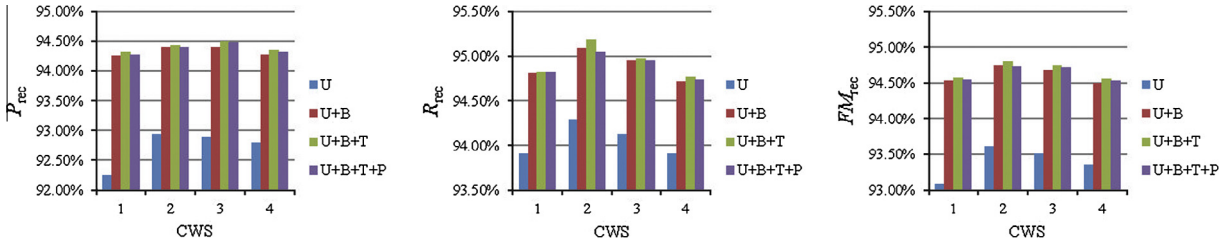

Fig. 6. P_{rec}, R_{rec}, and FM_{rec} obtained by CRF with different features under different CWS settings based on the general sequence labeling strategy.

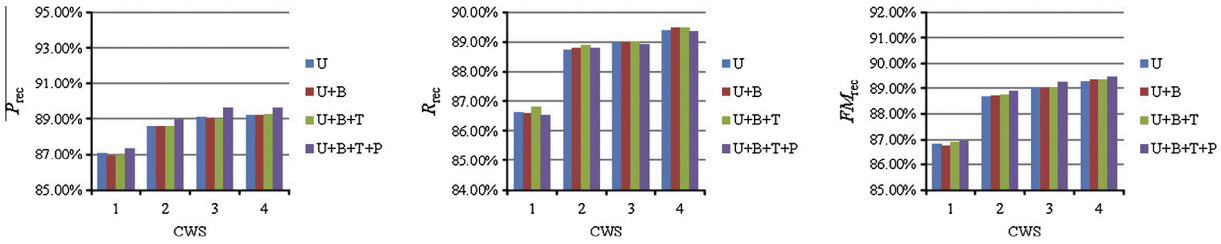

Fig. 8. P_{rec}, R_{rec}, and FM_{rec} obtained by MEMM with different features under different CWS settings based on the domain-specific adapted sequence labeling strategy which makes an adaptation to the labeled sequence.

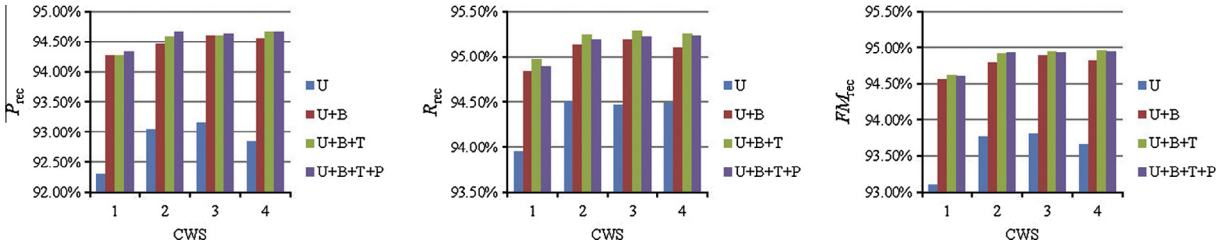

Fig. 9. P_{rec}, R_{rec}, and FM_{rec} obtained by CRF with different features under different CWS settings based on the domain-specific adapted sequence labeling strategy which makes an adaptation to the labeled sequence.

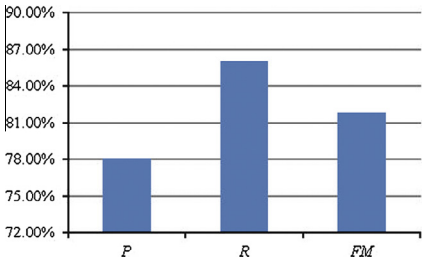

Fig. 10. P_{rec}, R_{rec}, and FM_{rec} obtained by HMM based on the domain-specific adapted sequence labeling strategy which makes an adaptation to the predefined tag set.

$$P_{rec} = \frac{|NSRC|}{|NSR|} \tag{5}$$

$$R_{rec} = \frac{|NSRC|}{|NS|} \tag{6}$$

$$FM_{rec} = \frac{2 \cdot P_{rec} \cdot R_{rec}}{P_{rec} + R_{rec}} \tag{7}$$

where $|NSRC|$ is the number of symptom names recognized correctly, $|NSR|$ is the number of symptom names recognized, and $|NS|$ is the number of symptom names in the test dataset. A symptom name is correctly recognized, if and only if its beginning, intermediate and end positions are all accurately labeled.

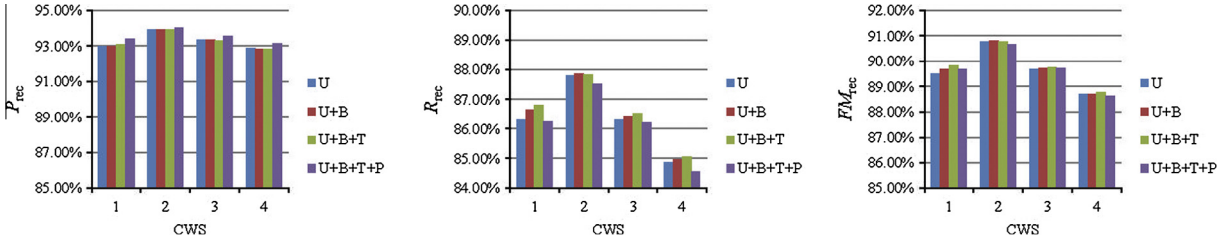

Fig. 11. P_{rec}, R_{rec}, and FM_{rec} obtained by MEMM with different features under different CWS settings based on the domain-specific adapted sequence labeling strategy which makes an adaptation to the predefined tag set.

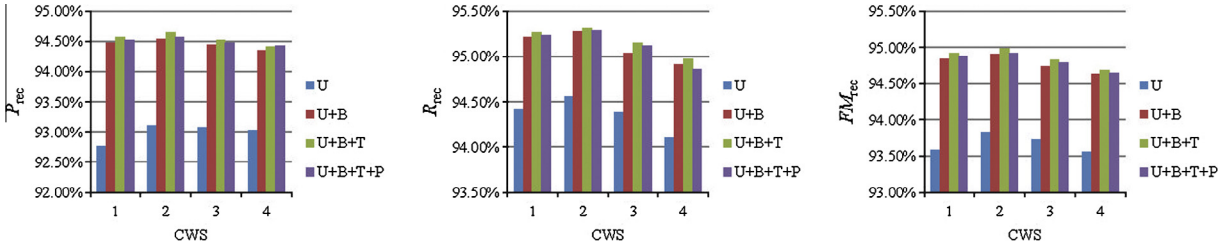

Fig. 12. P_{rec}, R_{rec}, and FM_{rec} obtained by CRF with different features under different CWS settings based on the domain-specific adapted sequence labeling strategy which makes an adaptation to the predefined tag set.

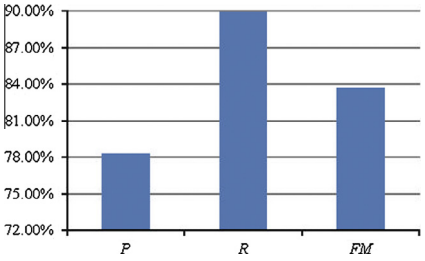

Fig. 13. P_{rec}, R_{rec}, and FM_{rec} obtained by HMM based on the domain-specific adapted sequence labeling strategy.

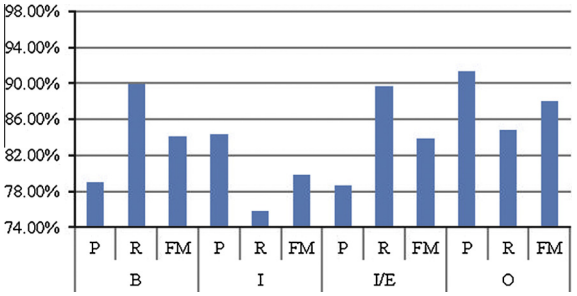

Fig. 16. P_{lab}, R_{lab}, and FM_{lab} of each predefined tag obtained by HMM based on the general sequence labeling strategy, where *P* represents P_{lab}, *R* represents R_{lab}, and FM represents FM_{lab}.

4.2.2. Definitions of P_{lab}, R_{lab}, and FM_{lab}

P_{lab}, R_{lab}, and FM_{lab} are designed for validating the feasibility of the sequence labeling strategy with a domain-specific adaptation and evaluating the performance of HMM, MEMM and CRF for the SNR task in detail. They are defined below.

$$P_{lab} = \frac{|NCLC|}{|NCL|} \tag{8}$$

$$R_{lab} = \frac{|NCL|}{|NC|} \tag{9}$$

$$FM_{lab} = \frac{2 \cdot P_{lab} \cdot R_{lab}}{P_{lab} + R_{lab}} \tag{10}$$

where $|NCLC|$ represents the number of Chinese characters in the test dataset that are correctly labeled with their corresponding tags, $|NCL|$ is the number of Chinese characters labeled, and $|NC|$ is the number of Chinese characters that should be labeled.

4.3. Experimental dataset

The gold standard experimental dataset of the chief complaints used in our experiments is from a TCM clinical record dataset, which was collected by TCM doctors during their routine diagnostic work from April 2006 to June 2008. The dataset contains 11,613 clinical records; an excerpt is shown in Fig. 3. The gold standard experimental dataset is constructed through following steps:

(1) Two TCM experts are invited to independently annotate symptom names mentioned in the chief complaints according to an annotation guideline, which was developed jointly by the two annotators and the authors in advance (the detailed information of the annotation guideline is described in Appendix 1). The Inter-Annotator Agreement [38] on the annotating results reaches 0.84. This value lies in "the almost perfect agreement interval" [39] or in "the excellent agreement interval" [40].

(2) In our experiments, only the symptom names agreed on by both annotators are treated as the gold standard symptom names. The Chinese characters in these symptom names are labeled with the predefined tags "B", "I" and "E" when validating our method or "B" and "I" when performing comparative experiments. Incompatible symptom names in the annotating results are viewed in the same way as other general descriptions, and their corresponding Chinese characters are labeled with the predefined tag "O".

(3) For convenience, a further process is performed on the gold standard experimental dataset. Every number, e.g. integers, decimals, and fractions, etc., in the dataset is uniformly replaced by an English character "*N*". Moreover, the punctuations are all replaced by the English character "*P*". These "*N*"s and "*P*"s will be all labeled with the predefined tag "O".

Finally, the gold standard experimental dataset is randomly divided into two parts. One contains 3483 chief complaints (about 30% of the original dataset), which is to be treated as the training

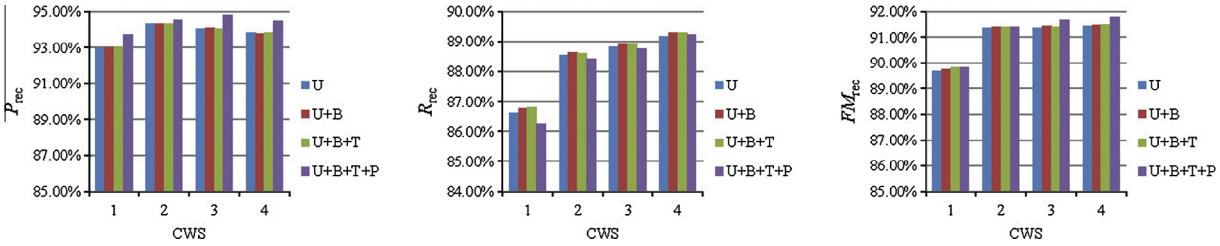

Fig. 14. P_{rec}, R_{rec}, and FM_{rec} obtained by MEMM with different features under different CWS settings based on the domain-specific adapted sequence labeling strategy.

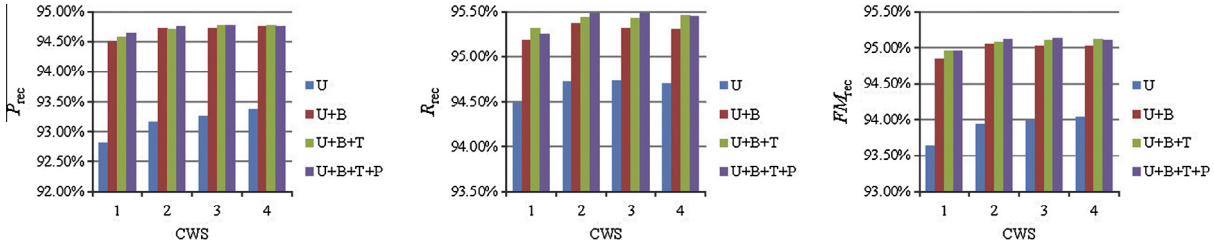

Fig. 15. P_{rec}, R_{rec}, and FM_{rec} obtained by CRF with different features under different CWS settings based on the domain-specific adapted sequence labeling strategy.

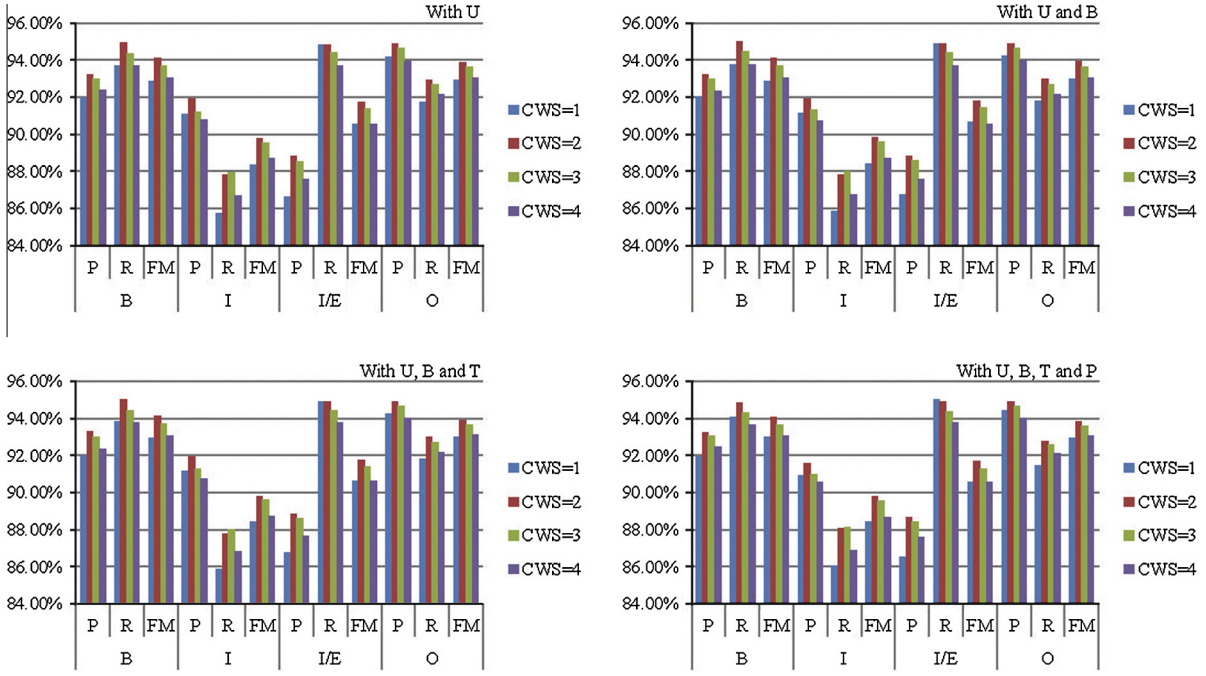

Fig. 17. P_{lab}, R_{lab}, and FM_{lab} of each predefined tag obtained by MEMM with different features under different CWS settings based on the general sequence labeling strategy, where P represents P_{lab}, R represents R_{lab}, and FM represents FM_{lab}.

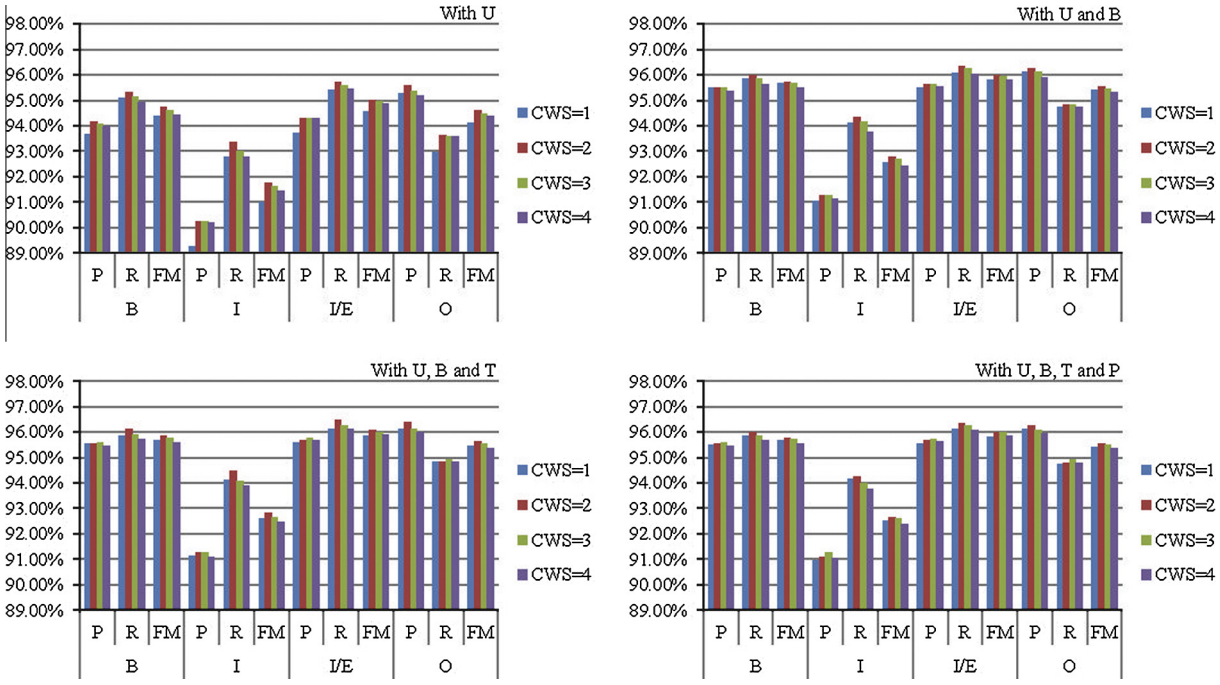

Fig. 18. P_{lab}, R_{lab}, and FM_{lab} of each predefined tag obtained by CRF with different features under different CWS settings based on the general sequence labeling strategy, where P represents P_{lab}, R represents R_{lab}, and FM represents FM_{lab}.

dataset. The remaining 8130 chief complaints form the test dataset. Detailed information of the training and test datasets is listed in Table 1.

5. Experimental results

5.1. Comprehensive evaluation

A comprehensive evaluation of the feasibility of the adapted sequence labeling strategy and the performance of the introduced classifiers for the SNR task are described below based on the symptom name recognition results.

5.1.1. Evaluating the contribution of the adaptation to the labeled sequence

The results in Figs. 7–9, compared with those in Figs. 4–6, show that after making a domain-specific adaptation to the labeled sequence based on the characteristics of the chief complaints, the SNR results obtained by different sequence classifiers are all improved. The highest improvement of FM_{rec} reaches 3.39% and

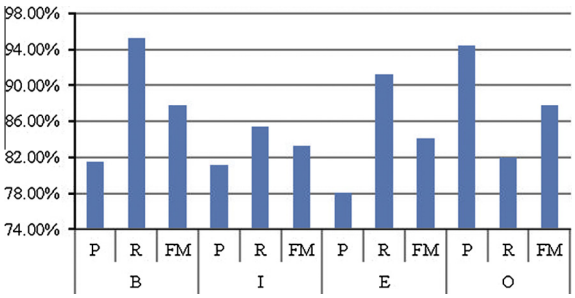

Fig. 19. P_{lab}, R_{lab}, and FM_{lab} of each predefined tag obtained by HMM based on the domain-specific adapted sequence labeling strategy which makes an adaptation to the labeled sequence, where *P* represents P_{lab}, *R* represents R_{lab}, and *FM* represents FM_{lab}.

the average improvement of FM_{rec} rises by 1.03%. Specifically, comparing the results displayed in Figs. 8 and 9 to the results in Figs. 5 and 6, after defining the clauses as the labeled sequences, P_{rec}, R_{rec}, and FM_{rec} are all improved. At the same time, with increasing CWS P_{rec}, R_{rec}, and FM_{rec} rise and then remain stable (see Figs. 8 and 9) rather than improving and then worsening as shown in Figs. 5 and 6. This demonstrates the effectiveness of the domain-specific adaption to the labeled sequence. As introduced in Section 2.2.2, this adaptation keeps the internal coherence of every clause and, at the same time, reduces the possibility of encountering incoherent sentences. This is similar to empirical feature selection, i.e. empirically reducing noisy features extracted from the disorganized context in the chief complaints.

5.1.2. Evaluating the contribution of the adaptation to the predefined tag set

Comparing the results depicted in Figs. 10–12 with the results shown in Figs. 4–6 reveals that after individually identifying the end of the symptom names, the SNR results (P_{rec}, R_{rec}, and FM_{rec}) obtained by different sequence classifiers are all improved. The highest improvement of FM_{rec} reaches 4.34% and the average improvement of FM_{rec} rises by 1.65%. These results demonstrate that the domain-specific adaptation to the predefined tag set is effective and that the strategy of individually identifying the end of the symptom names is appropriate. More detailed analysis of the contribution of the domain-specific adaptation to the predefined tag set is described in Section 5.2.2.

5.1.3. Comprehensively evaluating the adapted sequence labeling strategy

Comparing the results revealed in Figs. 13–15 to the results depicted in Figs. 4–6, vividly shows that after adapting the sequence labeling strategy based on the domain-specific characteristics of the chief complaints, the SNR results obtained by different sequence classifiers are all improved. The highest and the average improvements of FM_{rec} reach 6.18% and 3.64% respectively; the best FM_{rec} reaches 95.12% (P_{rec} 94.77% and R_{rec} 95.48%). This is achieved by CRF with the features U, B, T, and P which are fetched in CWS = 2 or 3 (see Fig. 9). This result is also obtained when the general sequence labeling strategy is made a domain-specific adaptation. Furthermore, compared to the results in Figs. 7–12, the results shown in Figs. 13–15 are better. This demonstrates that every adaptation to the general sequence labeling strategy is helpful and they all make contributions to SNR in the chief complaints. In addition, after adaptation to the general sequence labeling strategy, not only P_{rec} but also R_{rec} of every sequence classifier is improved; the results do not suffer from adverse effects when CWS is increased. All these results demonstrate that the domain-specific adaptation of the general sequence labeling strategy based on the characteristics of the chief complaints is appropriate and effective.

5.1.4. Evaluating the sequence classifiers for the SNR task

Comparing the SNR results achieved by MEMM and CRF with the results obtained by HMM (see Figs. 4–15) shows that MEMM and CRF are better than HMM. This demonstrates that the discriminative models (i.e. MEMM and CRF) outperform the generative

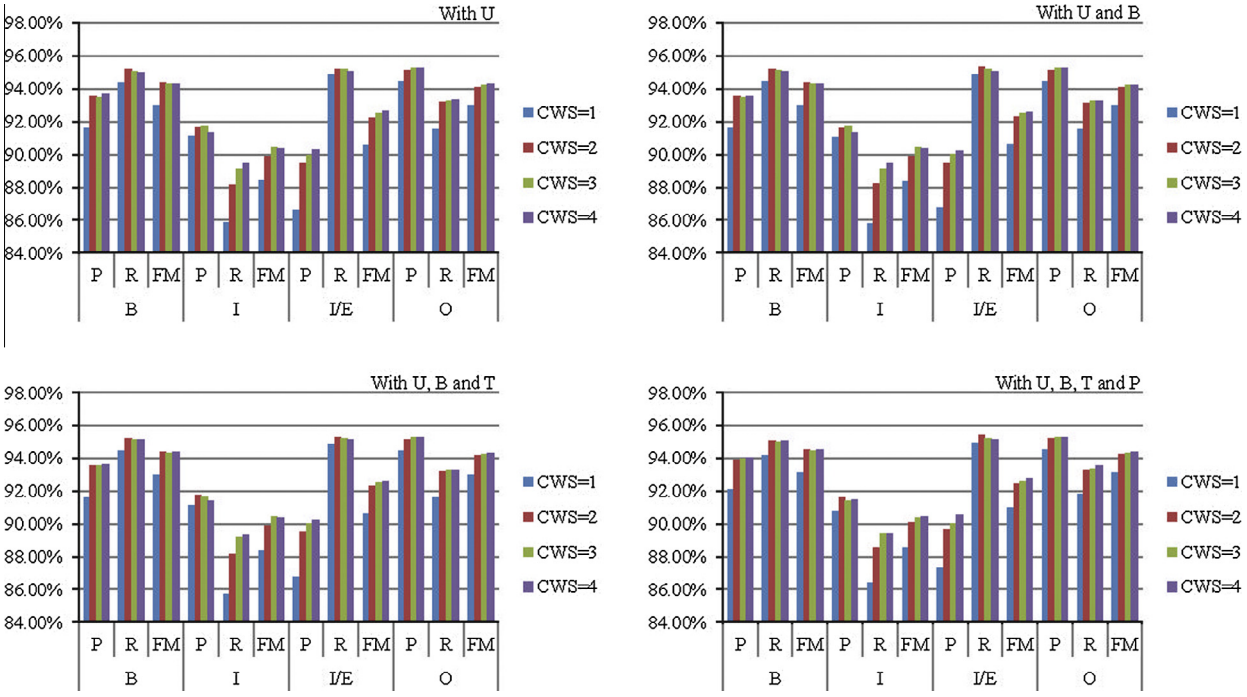

Fig. 20. P_{lab}, R_{lab}, and FM_{lab} of each predefined tag obtained by MEMM with different features under different CWS settings based on the domain-specific adapted sequence labeling strategy which makes an adaptation to the labeled sequence, where *P* represents P_{lab}, *R* represents R_{lab}, and *FM* represents FM_{lab}.

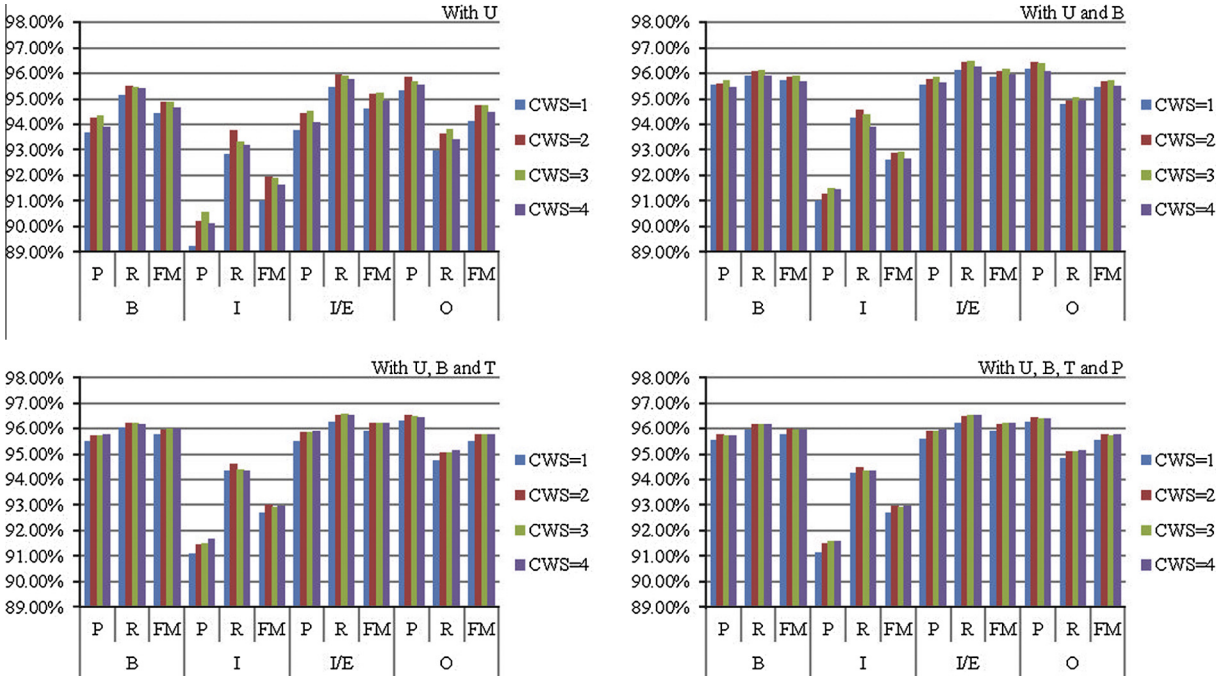

Fig. 21. P_{lab}, R_{lab}, and FM_{lab} of each predefined tag obtained by CRF with different features under different CWS settings based on the domain-specific adapted sequence labeling strategy which makes an adaptation to the labeled sequence, where *P* represents P_{lab}, *R* represents R_{lab}, and *FM* represents FM_{lab}.

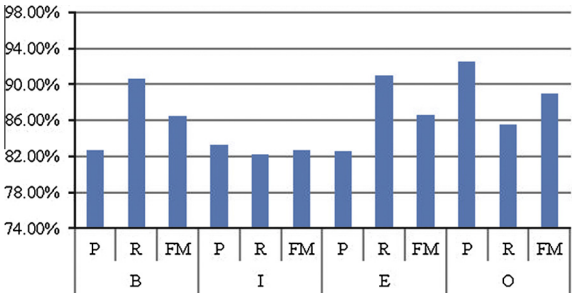

Fig. 22. P_{lab}, R_{lab}, and FM_{lab} of each predefined tag obtained by HMM based on the domain-specific adapted sequence labeling strategy which makes an adaptation to the predefined tag set, where *P* represents P_{lab}, *R* represents R_{lab}, and *FM* represents FM_{lab}.

model (i.e. HMM), because they effectively incorporate useful features into the sequence labeling procedure, thus improving the SNR performance. At the same time, a comparison of SNR results obtained by CRF to the results achieved by MEMM clearly shows that CRF outperforms MEMM, proving that CRF effectively avoids the label bias problem from which MEMM suffers.

5.2. Detailed analysis

In the previous section, the effectiveness of the domain-specific adapted sequence labeling strategy for the SNR task has been demonstrated through a comprehensive evaluation of SNR results, and the performance of different sequence classifiers for the SNR task is also described. In this section, we will give a detailed analysis of the adaption to the general sequence labeling strategy through evaluating the detailed sequence labeling results. In order to conveniently show comparisons, we will make a distinction between the tag "I" appearing at the end of the symptom names, which will be denoted as "I/E", and the tag "I" which occurs at the intermediate of the symptom names.

5.2.1. Evaluating the contribution of the adaptation to the labeled sequence

Comparing the results shown in Figs. 19–21 to the results revealed in Figs. 16–18 shows that after defining the clauses as the labeled sequences the labeling results (P_{lab}, R_{lab}, and FM_{lab}) of the tags "B", "I", "I / E", and "O" are all improved. In addition, Figs. 19–21 also show that the improvements of the labeling results are mainly due to the contribution of the substantial increase of R_{lab} of each type of the predefined tags. Moreover, Figs. 20 and 21 show that after the adaptation to the labeled sequence, P_{lab}, R_{lab}, and FM_{lab} of each type of the predefined tags do not suffer from adverse effect due to increasing CWS. These results further demonstrate the effectiveness of the domain-specific adaptation to the labeled sequence.

5.2.2. Evaluating the contribution of the adaptation to the predefined tag set

A comparison of the results revealed in Figs. 22–24 to the results depicted in Figs. 16–18 clearly show that the labeling results (not only FM_{lab} but also P_{lab} and R_{lab}) of the end of the symptom names are improved dramatically. At the same time, the labeling results of the other three predefined tags are also improved. These results have demonstrated that the adaptation to the predefined tag set described in this paper can effectively boost the labeling performance of the end of the symptom names, thereby indirectly improving the labeling performance of the other predefined tags.

5.2.3. Detailed analysis of the domain-specific adaptation to the labeling strategy

Detailed sequence labeling results obtained by HMM, MEMM, and CRF, after the domain-specific adaptation to the sequence labeling strategy, are shown in Figs. 25–27 respectively. The labeling of all predefined tags show significant improvements when compared to the results in Figs. 16–18, with an average increase of 1.80%. In addition, the labeling results of the predefined tag "E" show dramatic improvements without suffering any adverse

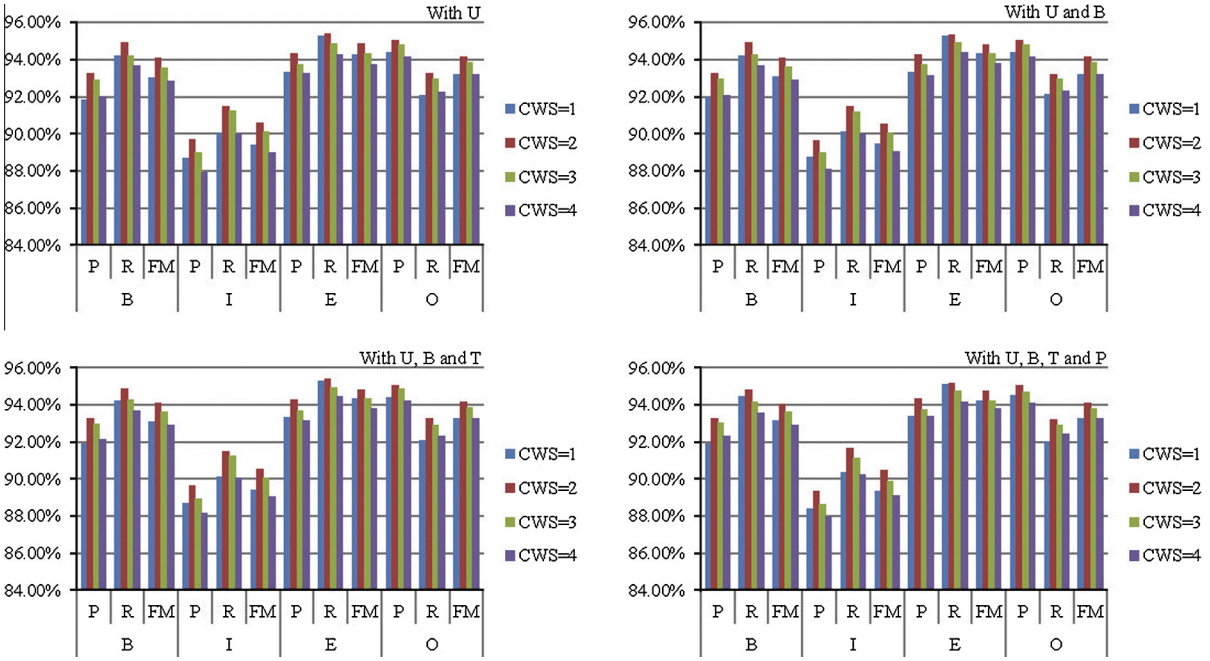

Fig. 23. P_{lab}, R_{lab}, and FM_{lab} of each predefined tag obtained by MEMM with different features under different CWS settings based on the domain-specific adapted sequence labeling strategy which makes an adaptation to the predefined tag set, where P represents P_{lab}, R represents R_{lab}, and FM represents FM_{lab}.

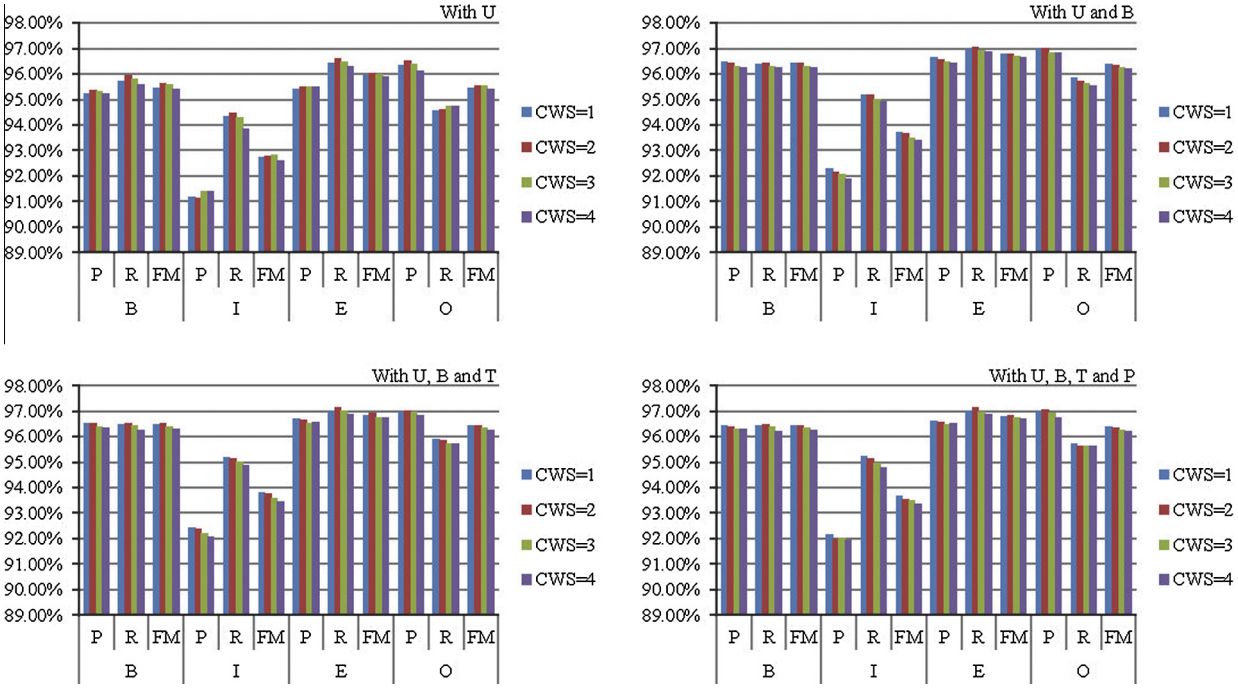

Fig. 24. P_{lab}, R_{lab}, and FM_{lab} of each predefined tag obtained by CRF with different features under different CWS settings based on the domain-specific adapted sequence labeling strategy which makes an adaptation to the predefined tag set, where P represents P_{lab}, R represents R_{lab}, and FM represents FM_{lab}.

effect from increasing CWS (see Figs. 26 and 27). All these results demonstrate that the domain-specific adaptation to the general sequence labeling strategy based on the characteristics of the chief complaints of TCM is appropriate and effective.

5.2.4. Detailed analysis of the sequence classifiers for the SNR task

The comparison of sequence labeling results obtained by MEMM and CRF to the results achieved by HMM (see Figs. 16–27) clearly shows that MEMM and CRF with useful features can achieve higher labeling performance than HMM. These results further demonstrate that the discriminative models (MEMM and CRF) are more suitable to the SNR task than the generative model (HMM), and that they can effectively incorporate useful features into the sequence labeling procedure for SNR in the chief complaints. In addition, under the same training conditions the labeling results of CRF are better than MEMM. This improvement is due to effectively avoiding the label bias problem from which MEMM suffers.

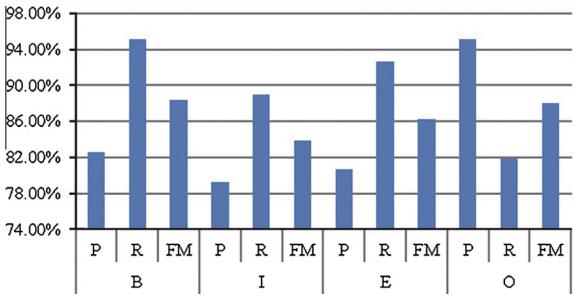

Fig. 25. P_{lab}, R_{lab}, and FM_{lab} of each predefined tag obtained by HMM based on the domain-specific adapted sequence labeling strategy, where *P* represents P_{lab}, *R* represents R_{lab}, and *FM* represents FM_{lab}.

5.2.5. More detailed analysis

Figs. 25–27 show that FM_{lab} of the predefined tag "I" is always lower than FM_{lab} of "B", "I" and "E". This reflects the fact that symptom names in chief complaints can be detected accurately, but that the exact boundaries of the symptom names are still difficult to identify. This result is also shown in NER in English discharge summaries of Western medicine [15]. Moreover, P_{lab} of the predefined tag "O" is always higher than its R_{lab}, but the results of "B", "I", and "E" are always lower. This reflects the fact that, in addition to HMM, MEMM and CRF also prefer to label Chinese characters with the "B", "I", and "E" rather than "O". Accordingly, in future work on improving sequence labeling results we should focus on the problems of improving the labeling performance of the intermediate of the symptom names and modifying the sequence classifiers to more fairly treat every type of tag in the labeling procedure.

5.3. Error analysis

In this section, the confusion matrix [41] is used to analyze errors in sequence labeling results. In an *N*-by-*N* confusion matrix, the cell (y, y') indicates the ratio of the number of gold standard Chinese characters with y to the number of incorrectly labeled characters with y'. This ratio can be used to indicate which type of labeling error occurs frequently in the sequence labeling procedure. The confusion matrices obtained by HMM, MEMM and CRF after adapting the general sequence labeling strategy based on the domain characteristics of the chief complaints for the SNR task are listed in Appendices 2–4.

Appendices 2–4 show that Chinese characters whose exact corresponding tags are "B", "I", and "E" are frequently mislabeled with the tag "O". By analyzing the labeling results, we see that this frequently occurs when a symptom name is an aggregation of several symptom names. For convenience, TCM doctors write such aggregations when there is a common prefix or suffix, provided that the aggregated symptom name does not change the meanings of the original symptom names. For example, to avoid the redundant effort, TCM doctors will aggregate the symptom names "痰多" (sputum is excessive), "痰稠" (sputum is thick) and "痰黏" (sputum is sticky) by letting them share their common prefix "痰" (sputum) to form a concise symptom name "痰多稠黏" (the sputum is excessive, thick and sticky), and the new aggregated symptom name would also express the meanings of the original symptom names. Similarly, "腕腹胀" (gastric and abdominal distension) is an aggregated symptom name of the symptom names "腕胀" (gastric distension) and "腹胀" (abdominal distension), both of which have a common suffix "胀" (distension). Unlike aggregated descriptions in general text or in English discharge summaries, chief complaints usually do not include conjunctions such as "和" (and) or "或" (or), before the last shared part of the aggregated symptom names. This results in the loss of an important feature for identifying the beginning or the end of aggregated symptom names. In future work, the domain-specific syntactic structure of the chief complaints could be explored. Domain-specific syntactic features can then be imported by the discriminative sequence classifiers into the SNR task, thereby helping to accurately recognize aggregated symptom names in chief complaints.

Appendices 2–4 also reveal that Chinese characters, whose exact corresponding tags are "O", are usually mislabeled with tags "I" and "B". These mistakes usually occur when symptom names include dispensable modifiers. TCM doctors have different

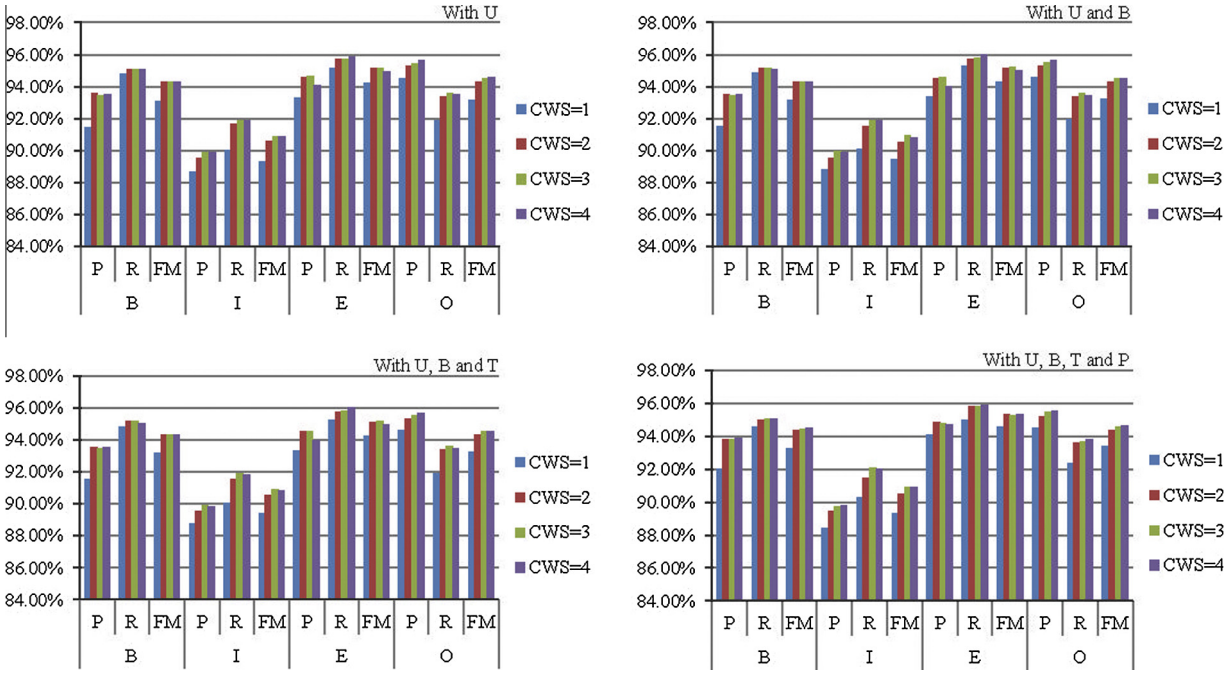

Fig. 26. P_{lab}, R_{lab}, and FM_{lab} of each predefined tag obtained by MEMM with different features under different CWS settings based on the domain-specific adapted sequence labeling strategy, where *P* represents P_{lab}, *R* represents R_{lab}, and *FM* represents FM_{lab}.

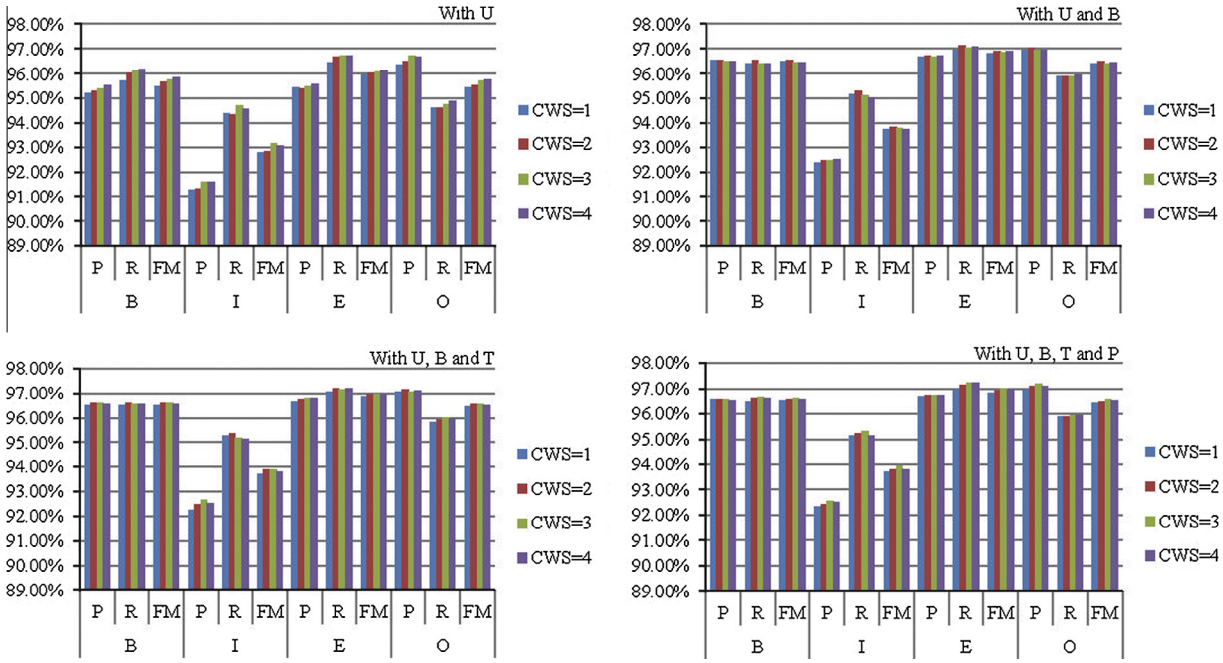

Fig. 27. P_{lab}, R_{lab}, and FM_{lab} of each predefined tag obtained by CRF with different features under different CWS settings based on the domain-specific adapted sequence labeling strategy, where P represents P_{lab}, R represents R_{lab}, and FM represents FM_{lab}.

opinions on whether to include or exclude dispensable modifiers in symptom names. For example, the Chinese character "有" (having), which is used to convey the meaning that "a patient has some symptoms", in the symptom name "有肠鸣" (has borborygmus) could be appended by several TCM doctors. This problem can be boiled down to the symptom name normalization problem and solved by normalization methods [28]. In future, we can attempt to model SNR and symptom name normalization jointly in order to further improve sequence labeling performance.

6. Discussion

With the rapid development of natural language processing, data mining, and machine learning techniques, many new state-of-the-art methods have been developed for the NER task. However, due to the distinctive characteristics of the FCRs of TCM, these general methods need domain-specific adaptation for NER in FCRs of TCM. In this paper, we study symptom name recognition in chief complaints. This is a fundamental task for other clinical research of TCM [8], such as automatic clinical diagnosis, discovering clinical knowledge from large sets of FCRs of TCM, and constructing TCM clinical expert systems. We encourage TCM researchers to pay more attention to this fundamental task.

Research on NER in FCRs of TCM is at an early stage; many practical problems remain to be solved. As shown in Section 4, we found that there are still some problems which need to be solved, for example, improving the recognition of aggregated symptom names in the chief complaints or designing an appropriately sequence classifier that could jointly model SNR and symptom name normalization. Exploration of the syntactic structures of FCRs of TCM is also a research area that may yield features useful for sequence classifiers. Our results also demonstrated that future domain-specific methods for NER in FCRs of TCM should be designed through the incorporation of more domain knowledge.

In comparing the recognition of concepts and medical information in English discharge summaries of Western medicine reported in [14,15] to the results shown in this paper, we find that the highest FM_{rec} (95.12%) of SNR in chief complaints reported in this paper is much higher than the best FM_{rec} (85.70% described in [14] and 85.20% reported in [15]) of NER in English discharge summaries of Western medicine. By comparing these two tasks and analyzing the detailed results obtained by one team (first place winners of the 2009 i2b2 medical information extraction task [16]), we find two reasons for these differences. First, unlike the naming of symptoms in English, TCM symptom names have domain-specific characteristics that usually comprise three aspects of descriptions (body location, sensation and intensity), and words used in TCM symptom names tend to have different distributions from words used in other contexts. These characteristics facilitate the SNR task as performed by sequence classifiers. Second, unlike work on NER in English discharge summaries of Western medicine, [14,15] we only recognize symptom names which eases the burden of the sequence classifiers. Several other entities remain be considered, such as treatments, temporal information, and drugs. The more entities taken into account, the greater the challenges. For example, Patrick and Min showed that FM_{rec} of the medication entities in English discharge summaries can reach 91.40%, however FM_{rec} of the reason entities only reaches 55.52% under the same experimental settings [16]. Recognition of multiple types of named entities will be needed to further exploit the FCRs of TCM.

7. Conclusions

SNR in chief complaints is an essential and fundamental task for exploiting the content of FCRs of TCM, discovering valuable clinical knowledge of TCM, and constructing clinical expert systems for TCM. It provides an opportunity to effectively and efficiently make use of an abundant clinical knowledge source – FCRs of TCM – and, consequently, further assist TCM modernization. In this paper, SNR in chief complaints is reasonably treated as a sequence labeling problem. According to the domain-specific characteristics of FCRs of TCM, the general sequence labeling strategy is appropriately adapted for the SNR task for several empirical reasons. Moreover, three typically supervised classifiers are investigated, and their specializations for the SNR task are interpreted carefully. A series of elaborate experiments were performed. The results demonstrate

that the domain-specific adaptation of sequence labeling strategies is appropriate and effective, and that CRF outperforms HMM and MEMM for the SNR task.

Acknowledgments

We would like to thank Ph.D. Ju Wang, M.S. Juan Zhao, M.S. Xuehong Zhang and M.S. Shengrong Zou for their helpful suggestions on this work and for their valuable work on manually structuring and annotating clinical records for us. The authors are also pleased to acknowledge Dr. James J. Cimino, Mr. Daniel O'Sullivan, Ms. Sky Y. Chen and Ms. Cathy F. Yu for the help in revising the paper. From all above, we are particularly thankful to Dr. James J. Cimino for taking the time out of his busy schedule to give us the constructive comments and helpful suggestions.

Funding: This work was supported by NSFC under Grants No. 61173182 and 61179071, as well as by support from Sichuan Province under Grant Nos. 2012HH0004, 2012HH0031 and 2008SZ0049 and Zhejiang Province under Grant No. LY12F02010.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2013.09.008.

References

[1] Pal SK. Complementary and alternative medicine: an overview. Curr Sci 2002;82(5):518–24.
[2] Barnes PM, Powell-Griner E, McFann K, Nahin RL. Complementary and alternative medicine use among adults: United States, 2002. Semin Integr Med 2004;2(2):54–71.
[3] Molassiots M, Fernadez-Ortega P, Pud D, Ozden G, Scott JA, Panteli V, et al. Use of complementary and alternative medicine in cancer patients: a European survey. Ann Oncol 2005;16:655–63.
[4] Tang J-L, Liu B-Y, Ma K-W. Traditional Chinese medicine. Lancet 2008;372 (9654):1938–40.
[5] Feng Y, Wu Z, Zhou X, Zhou Z, Fan W. Knowledge discovery in traditional Chinese medicine: state of the art and perspectives. Artif Intell Med 2006;38 (3):219–36.
[6] Zhou X, Chen S, Liu B, Zhang R, Wang Y, Li P, et al. Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support. Artif Intell Med 2010;48(2–3):139–52.
[7] Wang X, Qu H, Liu P. A self-learning expert system for diagnosis in traditional Chinese medicine. Expert Syst Appl 2004;26:557–66.
[8] Zhou X, Peng Y, Liu B. Text mining for traditional Chinese medicine knowledge discovery: a survey. J Biomed Inform 2010;43:650–60.
[9] Jurafsky D, Martin JH. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. 2nd ed. Englewood Cliffs (NJ): Pearson Prentice-Hall; 2008.
[10] Wu Y, Zhao J, Xu B, Yu H. Chinese named entity recognition based on multiple features. In: Proceedings of human language technology conference and conference on empirical methods in natural language processing (HLT/EMNLP); 2005. p. 427–34.
[11] Fu G, Luke KK. Chinese named entity recognition using lexicalized HMMs. SIGKDD Expl 2005;7(5):19–25.
[12] Gao J, Li M, Wu A, Huang C-N. Chinese word segmentation and named entity recognition: a pragmatic approach. Comput Linguist 2005;31(4):531–74.
[13] Wang Y, Yu Z, Jiang Y, Liu Y, Chen L, Liu Y. A framework and its empirical study of automatic diagnosis of traditional Chinese medicine utilizing raw free-text clinical records. J Biomed Inform 2012;45:210–23.
[14] Uzuner Ö, Solti I, Cadag E. Extracting medication information from clinical text. J Am Med Inform Assoc 2010;17:514–8.
[15] Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc 2011;18:552–6.
[16] Patrick J, Min L. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. J Am Med Inform Assoc 2010;17:524–7.
[17] Li Z, Cao Y, Antieau L, Agarwal S, Zhang Q, Yu H. Extracting medication information from patient discharge summaries. In: Proceedings of the third i2b2 workshop on challenges in natural language processing for clinical data; 2009.
[18] 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 Am Med Inform Assoc 2011;18(5):601–6.
[19] Kang N, Barendse RJ, Afzal Z, Singh B, Schuemie MJ, van Mulligen EM, et al. Erasmus MC approaches to the i2b2 challenge. In: Proceedings of the 2010 i2b2/VA workshop on challenges in natural language processing for clinical data. Boston, MA, USA: i2b2; 2010.
[20] Torii M, Wagholikar K, Liu H. Using machine learning for concept extraction on clinical documents from multiple data sources. J Am Med Inform Assoc 2011;18(5):580–7.
[21] Gurulingappa H, Hofmann-Apitius M, Fluck J. Concept identification and assertion classification in patient health records. In: Proceedings of the 2010 i2b2/VA workshop on challenges in natural language processing for clinical data. Boston, MA, USA: i2b2; 2010.
[22] Patrick JD, Nguyen DHM, Wang Y, Li M. A knowledge discovery and reuse pipeline for information extraction in clinical notes. J Am Med Inform Assoc 2011;18(5):574–9.
[23] Jonnalagadda S, Gonzalez G. Can distributional statistics aid clinical concept extraction? In: Proceedings of the 2010 i2b2/VA workshop on challenges in natural language processing for clinical data. Boston, MA, USA: i2b2; 2010.
[24] Sasaki Y, Ishihara K, Yamamoto Y, et al. TTI's systems for 2010 i2b2/VA challenge. In: Proceedings of the 2010 i2b2/VA workshop on challenges in natural language processing for clinical data. Boston, MA, USA: i2b2; 2010.
[25] Roberts K, Rink B, Harabagiu S. A flexible framework for deriving assertions from electronic medical records. J Am Med Inform Assoc 2011;18(5):568–73.
[26] Pai AK, Agichtein E, Post AR, et al. The emory system for extracting medical concepts at 2010 i2b2 challenge: integrating natural language processing and machine learning techniques. In: Proceedings of the 2010 i2b2/VA workshop on challenges in natural language processing for clinical data. Boston, MA, USA: i2b2; 2010.
[27] Chen MJ. A comparison of Chinese and English language processing. Adv Psychol 1993;103:97–117.
[28] Wang Y, Yu Z, Jiang Y, Xu K, Chen X. Automatic symptom name normalization in clinical records of traditional Chinese medicine. BMC Bioinform 2010;11:40.
[29] Wang Y, Liu Yi, Yu Z, Chen L, Jiang Y. A preliminary work on symptom name recognition from free-text clinical records of traditional Chinese medicine using conditional random fields and reasonable features. In: Proceedings of the 2012 workshop on biomedical natural language processing. Montreal, QC, Canada; 2012. p. 223–30.
[30] Sha F, Pereira F. Shallow parsing with conditional random fields. In: Proceedings of the 2003 conference of the North American chapter of the association of computer linguistics on human language technology. Edmonton, Canada; 2003. p. 134–41.
[31] Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 1989;77(2):257–86.
[32] McCallum A, Freitag D, Pereira F. Maximum entropy Markov models for information extraction and segmentation. In: Proceedings of the seventeenth international conference on machine learning; 2000. p. 591–8.
[33] Malouf R. A comparison of algorithm for maximum entropy parameter estimation. In: Proceedings of the sixth conference on natural language learning; 2002. p. 49–55.
[34] Byrd RH, Lu P, Nocedal J. A limited memory algorithm for bound constrained optimization. SIAM J Sci Stat Comput 1995;16:1190–208.
[35] Lafferty J, McCallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the eighteenth international conference on machine learning; 2001. p. 282–9.
[36] CRF++: Yet another CRF toolkit. <http://crfpp.googlecode.com/svn/trunk/doc/index.html> [validated 06.06.13].
[37] Nie J-J, Gao J, Zhang J, Zhou M. On the use of words and n-grams for Chinese information retrieval. In: Proceedings of the fifth international workshop on information retrieval with Asian languages; 2000. p. 141–8.
[38] Cohen's kappa on Wikipedia. http://en.wikipedia.org/wiki/Cohen's_kappa, validated on March 26, 2013.
[39] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159–74.
[40] Fleiss JL. Statistical methods for rates and proportions. 2nd ed. New York: John Wiley; 1981.
[41] Confusion matrix on Wikipedia. <http://en.wikipedia.org/wiki/Confusion_matrix> [validated 06.06.13].