A Preliminary Work on Symptom Name Recognition from Free-Text Clinical Records of Traditional Chinese Medicine using Conditional Random Fields and Reasonable Features

Yaqiang Wang, Yiguang Liu, Zhonghua Yu*, Li Chen
Department of Computer Science
Sichuan University
Chengdu, Sichuan 610064, China
yaq.wang@yahoo.com, lygpapers@yahoo.com.cn,
yuzhonghua@scu.edu.cn, cl@scu.edu.cn

Yongguang Jiang
Department of Preclinical Medicine
Chengdu University of TCM
Chengdu, Sichuan 610075, China
cdtcm@163.com

Abstract

A preliminary work on symptom name recognition from free-text clinical records (FCRs) of traditional Chinese medicine (TCM) is depicted in this paper. This problem is viewed as labeling each character in FCRs of TCM with a pre-defined tag ("B-SYC", "I-SYC" or "O-SYC") to indicate the character's role (a beginning, inside or outside part of a symptom name). The task is handled by Conditional Random Fields (CRFs) based on two types of features. The symptom name recognition F-Measure can reach up to 62.829% with recognition rate 93.403% and recognition error rate 52.665% under our experiment settings. The feasibility and effectiveness of the methods and reasonable features are verified, and several interesting and helpful results are shown. A detailed analysis for recognizing symptom names from FCRs of TCM is presented through analyzing labeling results of CRFs.

1 Introduction

Traditional Chinese medicine (TCM), a complementary medical theory to western medicine, provides a distinct way to view our human life (Pal, 2002; Barnes, et al., 2004; Molassiotis, et al., 2005). Moreover, it has shown that TCM knowledge, which is accumulated in clinical practice, has become one of the most important sources of modern biomedical research (Zhou, et al., 2010).

In recent years, Data Mining and Machine Learning have been more than ever before applied to TCM clinical research, such as establishing TCM diagnosis expert systems for supporting decision making (Wang, et al., 2004; Huang and Chen, 2007; Zhang, et al., 2008). However, most of the works are based on manually well-structured datasets.

Because of the high cost of manually structuring and maintaining free-text clinical records (FCRs) of TCM, large volume of such datasets has not been exploited effectively (Zhou, et al., 2010), although they are significant for discovering new knowledge or capturing medical regularities. Therefore, developing appropriate information extraction methods for handling FCRs of TCM is an urgent need to reduce the manual labor for researchers.

Automatically extracting meaningful information and knowledge from FCRs of TCM is challenging in Data Mining and Machine Learning fields (Zhou, et al., 2010). As the basis, symptom name recognition or extraction from FCRs of TCM is in an early stage. To the best of our knowledge, there has little work to solve this problem (Wang, et al., 2010; Wang, et al., 2012). Symptom name recognition from FCRs of TCM was firstly attempted in (Wang, et al., 2010) through normalizing the symptom names in clinical records based on literal similarity and remedy-based similarity methods but not directly recognizing original clinical symptom names from FCRs of TCM. In 2012, Wang, et al. proposed a framework of automatic diagnosis of TCM for practice. Symptom name recognition is one part of the framework and simp-

*Corresponding author

Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP 2012), pages 223–230,
Montréal, Canada, June 8, 2012.

ly attempted through a literal similarity method without detailed analysis (summarized procedures for the previous wok are shown in figure 1).

Wang, Y., et al., 2010:

-> Input FCR's clauses "昨日肠鸣", "失气多", "心中不适"

-> Output the standard symptom names that are most similar to these clauses (the similarity measured by literal similarity metrics, remedy-based similarity metrics or hybrid similarity metrics)

Wang, Y., et al., 2012:

(1) Dictionary-based method:

-> Input a FCR "昨日肠鸣, 失气多, 心中不适"

-> Based on symptom name dictionary matching symptom name

-> Output symptom names: "肠鸣", "失气多", "心中不适"

(2) Bigram-based method & literal similarity method:

-> Input a FCR "昨日肠鸣, 失气多, 心中不适"

-> Segment the FCR into bigrams "昨日, 日肠, 肠鸣, 失气, …, 不适"

-> Generating all possible bigram combination lists through combining and merging neighbor bigrams

-> Output the combination list that has the highest *CombValue* defined in the paper

Figure 1. Simple Conclusions of the Previous Work.

Named Entity Recognition (NER) has been widely studied. There have been lots of methods for Chinese NER (Zhang, et al., 2003; Wu, et al., 2003; Gao, et al., 2005; Fu and Luke, 2005; Zhou, 2006; Duan and Zhang, 2011). However, these methods cannot be directly applied on symptom name recognition from FCRs of TCM due to big differences of characteristics of the corpus (Wang, et al., 2012). There are also several related work on English NER, but Chinese NER has more challenges because of the distinct characteristics of Chinese (Wu, et al., 2003).

In this paper, the task of symptom name recognition from FCRs of TCM is studied. The symptom names are recognized through finding their description boundaries from FCRs of TCM, and the method is described in section 2. Several reasonable and helpful features are introduced for CRFs to label the characters in FCRs of TCM with pre-defined boundary tags to indicate their roles (a beginning, inside or outside part of a symptom name) (presented in section 3). At last, several interesting and valuable experimental results are shown in section 4 and a conclusion is given in section 5.

2 Symptom Name Recognition from FCRs of TCM

The task of symptom name recognition from FCRs of TCM can be treated as detecting the boundaries of the symptom name descriptions in the sentences of FCRs of TCM. Therefore, this task can be viewed as labeling each tagging unit (e.g. word) in the sentences with a pre-defined tag indicating whether the unit is a beginning, inside, or outside part of a symptom name.

Generally, the tagging unit is word (Ramshaw and Marcus, 1995). However, there is no natural segmentation for words in Chinese sentences. Therefore, Chinese word segmentation problem has to face up firstly (Gao, et al., 2005). Because of the characteristics of FCRs of TCM (Wang, et al., 2012), automatically segmenting FCRs of TCM into words is not trivial and common Chinese word segmentation methods are not suitable. In order to tackle this problem, Chinese character is settled as the basic tagging unit. An example sentence of the labeling task is shown in figure 2.

昨	日	肠	鸣	，	失	气	多	，	心	中	不	适
O-SYC	O-SYC	[B-SYC	I-SYC]		[B-SYC	I-SYC	I-SYC]		[B-SYC	I-SYC	I-SYC	I-SYC]

Figure 2. An Example Sentence of the Symptom Name Recognition Task.

In figure 2, each character is labeled with a pre-defined tag ("B-SYC", "I-SYC" or "O-SYC"). The meaning of each tag is defined in table 1.

Tag	Meaning
B-SYC	Beginning of a TCM symptom name
I-SYC	Inside a TCM symptom name
O-SYC	Outside the TCM symptom names

Table 1. Meanings of the Pre-defined Tags.

Consequently, a recognized symptom name should start with a character labeled with "B-SYC" and end before the character whose corresponding label changes from "I-SYC" to "B-SYC" or "O-SYC" for the first time. The labeling task can be formulated as follows:

Given a FCR $\mathbf{x} = x_1, x_2, ..., x_n$, where x_i is a Chinese character, the goal is to build a annotator p to accurately label $\mathbf{x}$ with the credible corresponding tag sequence $\mathbf{y} = p(\mathbf{x})$, where $\mathbf{y} = y_1, y_2, ..., y_n$ and $y_n \in \{\text{B-SYC,I-SYC,O-SYC}\}$. This task can be effectively done by CRFs (Sha and Pereira, 2003) based on a training dataset which is consisted of pairs of sequences $(\mathbf{x},\mathbf{y})$.

3 Conditional Random Fields for Symptom Name Recognition

3.1 Conditional Random Fields

A Conditional Random Field can be defined as an undirected graphical model (see figure 3) which consists of a sequence of vertices representing random variables $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ and edges representing conditional dependencies, conditioned on $\mathbf{X} = (X_1, X_2, ..., X_n)$. The random variable Y_i only has edges with its predecessor Y_{i-1} and successor Y_{i+1}, thus, random variables $Y_1, Y_2, ..., Y_n$ obey the Markov property and form a linear Markov chain.

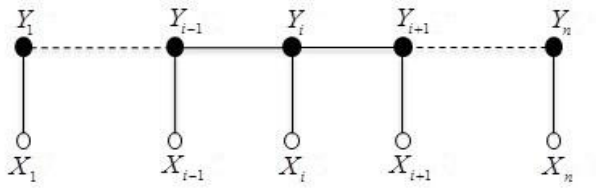

Figure 3. An Undirected Graphical Structure for a Conditional Random Field.

Then the conditional probability of a label sequence given an input sequence can be defined as:

$$p_{\boldsymbol{\lambda}}(\mathbf{y}, \mathbf{x}) = \frac{\exp \boldsymbol{\lambda} \cdot \sum_{i=1}^{n} \boldsymbol{f}(\mathbf{y}, \mathbf{x}, i)}{Z_{\boldsymbol{\lambda}}(\mathbf{x})}$$

Where $\boldsymbol{f}$ is a *global feature vector* (Sha and Pereira, 2003) and each element of $\boldsymbol{f}$ is an arbitrary feature selection function f_k ($k \in [1, K]$, where K is the number of feature functions). $\boldsymbol{\lambda}$ is a weight vector comprised by the learned weight λ_k for each feature function. More detailed description is that,

$$p(\mathbf{y} \mid \mathbf{x}) = \frac{\exp\left(\sum_{i=1}^{n} \sum_{k=1}^{K} \lambda_k f_k(\mathrm{y}_{i-1}, \mathrm{y}_i, \mathbf{x}, i) \right)}{Z(\mathbf{x})}$$

$Z(\mathbf{x})$ in the equation is a normalization factor which is the sum over all possible label sequences $\boldsymbol{S}$:

$$Z(\mathbf{x}) = \sum_{\boldsymbol{S}} \exp\left(\sum_{i=1}^{n} \sum_{k=1}^{K} \lambda_k f_k(\mathrm{y}_{i-1}, \mathrm{y}_i, \mathbf{x}, i) \right)$$

The most likely label sequence for an input sequence $\mathbf{x}$ is:

$$\mathbf{y} = \arg\max_{\mathbf{y}} p(\mathbf{y} \mid \mathbf{x})$$

It can be found with the Viterbi algorithm. We use the CRF++ tool in the experiments, which provides an efficient implementation for CRFs by using the limited-memory quasi-Newton algorithm for training the models (Sha and Pereira, 2003; Lafferty, et al., 2001) and the default settings of CRF++ are used.

3.2 Features for Labeling

It is difficult to analyze the syntactic structure of the content in FCRs of TCM which has narrative form, concise style and nonstandard description characteristics. Therefore, no higher level syntactic features, such as POS tags or NP chunks, can be used at the moment. Through analyzing FCRs of TCM, two types of representative and reasonable features (i.e. literal features and positional features) are exploited. The features are introduced and their reasonableness is explained by examples as follows.

Literal Features: the simplest and the most obvious features for determining the boundaries of symptom name descriptions are literal features. For example, according to the observation that after a word which is used to specify time (e.g. “昨日” (yesterday)) there would usually follow a symptom name description, such as “肠鸣” (borborygmus).

The best approach to get such features is to divide the content of FCRs of TCM into words. However, as described before, Chinese word segmentation is not trivial work. Fortunately, segmenting the content into n-grams is considerable and reasonable, because the indicating words would be mixed in the n-gram segments and could be helpful to determine the boundaries of symptom name descriptions.

Furthermore, the FCRs of TCM have a concise style, i.e. the length of the clauses in FCRs of TCM is short and words are usually used in their brief form. Therefore, the n-grams as the literal features need not be too long. In general, the average length of a Chinese word approximates 2 (Nie, et al., 2000). Consequently, the value of *n* should set to range from 1 to 3. Moreover, according to the intuition that “the distance between current character and its related n-grams in FCRs of TCM would not be too far”, the context window size, which is the fragment scope picking up literal features (i.e. n-

grams (see examples in table 2)) in FCRs of TCM, would not be too large. Otherwise it would bring about noisy information, thereby reducing the labeling precision. The context window size in our experiment is specified smaller then 4.

Feature Type	Context Window Size (CWS)	Literal feature examples under different CWS
Unigram Features (Uni)	1	C_{i-1}, C_i, C_{i+1}
	2	C_{i-2}, C_{i-1}, C_i, C_{i+1}, C_{i+2}
	3	C_{i-3}, C_{i-2}, C_{i-1}, C_i, C_{i+1}, C_{i+2}, C_{i+3}
	4	C_{i-4}, C_{i-3}, C_{i-2}, C_{i-1}, C_i, C_{i+1}, C_{i+2}, C_{i+3}, C_{i+4}
Bigram Features (Big)	1	$C_{i-1}C_i$, C_i C_{i+1}
	2	$C_{i-2}C_{i-1}$, $C_{i-1}C_i$, C_iC_{i+1}, $C_{i+1}C_{i+2}$
	3	$C_{i-3}C_{i-2}$, $C_{i-2}C_{i-1}$, $C_{i-1}C_i$, C_iC_{i+1}, $C_{i+1}C_{i+2}$, $C_{i+2}C_{i+3}$
	4	$C_{i-4}C_{i-3}$, $C_{i-3}C_{i-2}$, $C_{i-2}C_{i-1}$, $C_{i-1}C_i$, C_iC_{i+1}, $C_{i+1}C_{i+2}$, $C_{i+2}C_{i+3}$, $C_{i+3}C_{i+4}$
Trigram Features (Tri)	1	$C_{i-1}C_iC_{i+1}$
	2	$C_{i-2}C_{i-1}C_i$, $C_{i-1}C_iC_{i+1}$, $C_iC_{i+1}C_{i+2}$
	3	$C_{i-3}C_{i-2}C_{i-1}$, $C_{i-2}C_{i-1}C_i$, $C_{i-1}C_iC_{i+1}$, $C_iC_{i+1}C_{i+2}$, $C_{i+1}C_{i+2}C_{i+3}$
	4	$C_{i-4}C_{i-3}C_{i-2}C_{i-1}$, $C_{i-3}C_{i-2}C_{i-1}C_i$, $C_{i-2}C_{i-1}C_iC_{i+1}$, $C_{i-1}C_iC_{i+1}C_{i+2}$, $C_iC_{i+1}C_{i+2}C_{i+3}$, $C_{i+1}C_{i+2}C_{i+3}C_{i+4}$

Table 2. Literal Feature Examples Used in the Experiments. C_i is the character at current position i in one clause.

Positional Features: positions of characters in FCRs of TCM are also helpful. They are assistant features to determine the boundaries of symptom name descriptions.

The start of a sentence would be usually a common character (i.e. its corresponding label is "O-SYC") rather than the beginning of a symptom name description. On the contrary, the starting positions of the following clauses have higher probabilities to be labeled with "B-SYC". Taking the FCR "昨日肠鸣, 失气多, 心中不适" (Yesterday, the patient had borborygmus and more farting, and his/her heart was uncomfortable) as an example, it starts with a common word "昨日" (yesterday) followed by a symptom name "肠鸣" (borborygmus). And at the same time, following clauses all start with symptom name descriptions.

The example of positional features is shown in figure 4.

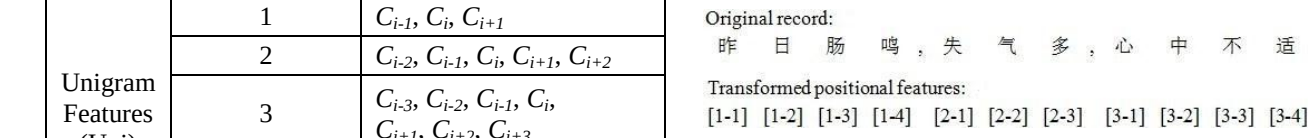

Figure 4. Example of Positional Features.

In figure 4, one "[*SubSID-POS*]" represents a positional feature, and *SubSID* is the index of current clause in a FCR and *POS* indicates the position of a character in current clause.

4 Experiments

In this section, the proposed method for symptom name recognition from TCM FCRs is evaluated, and the usefulness of the introduced features is verified based on a TCM clinical dataset. The results are depicted bellow.

4.1 Experimental Datasets

In this paper, a clinical record dataset (CRD) is used. It contains 11613 FCRs of TCM and was collected by TCM doctors during their routine diagnostic work. The Chinese characters in FCRs of CRD are annotated with tags "B-SYC", "I-SYC", and "O-SYC". The number of each type of tags is 69193, 104243 and 142860, respectively. There are 4235 unique symptom names in CRD, and the amount of annotated symptom names is 69193.

	Training Data	Test Data
Number of Unique Symptom Names	1813	3463
Amount of Symptom Names	17339	51854
Number of Each Type of Tags ("B-SYC", "I-SYC", "O-SYC")	17339, 25738, 35995	51854, 78505, 106865

Table 3. Detailed Information of the Training and Test Datasets.

CRD is divided into two sub-datasets (i.e. a training dataset (3483 FCRs, 25% of CRD) and a test dataset (8130 FCRs, 75% of CRD)). For con-

venience, all numbers (e.g. integers, decimals and fractions, etc.) in CRD are uniformly replaced by a English character "*N*" in advance. Detailed information of training and test datasets is listed in table 3.

4.2 Evaluation Metrics

A new method for symptom name recognition from FCRs of TCM is proposed and two types of features are introduced. To evaluate the feasibility and effectiveness of the method and features, two groups of evaluation metrics are designed: (1) for assessing the ability of symptom name recognition, symptom name recognition rate, recognition error rate and recognition F-Measure are defined; (2) for giving a detailed analysis, the labeling precision, recall, and F-Measure are exercised. The detailed explanations of these metrics are described below.

Symptom name recognition rate (RR_{det}), recognition error rate (RER_{det}) and recognition F-Measure (RFM_{det}): these metrics are designed for assessing capability of the proposed method for symptom name recognition from TCM FCRs. If and only if the boundary of a symptom name is labeled accurately (i.e. starting with "B-SYC" and ending with the first change from "I-SYC" to "B--SYC" or "O-SYC"), the recognized symptom name is correct. Higher RR_{det} and lower RER_{det} are achieved; better symptom name recognition performance RFM_{det} would be obtained. RR_{det}, RER_{det} and RFM_{det} are formulated as follows.

$$RR_{det} = \frac{|NSDC|}{|NCS|}$$

$$RER_{det} = \frac{|SD| - |NSDC|}{|SD|}$$

$$RFM_{det} = \frac{2 \cdot DR_{det} \cdot (1 - DER_{det})}{DR_{det} - DER_{det} + 1}$$

Where $|NSDC|$ is the number of symptom name recognized correctly from the test dataset, $|NCS|$ is the number of clinical symptom names in the test dataset, and $|SD|$ is the number of symptom name recognized.

Labeling precision (Pre_{lab}), recall (Rec_{lab}) and F-Measure (FM_{lab}): the metrics (Pre_{lab}, Rec_{lab} and FM_{lab}) are used to evaluate the performance of labeling Chinese character sequences of FCRs of TCM for giving a detailed analysis. They are defined below.

$$Pre_{lab} = \frac{|NCLC|}{|NCL|}$$

$$Rec_{lab} = \frac{|NCL|}{|NC|}$$

$$FM_{lab} = \frac{2 \cdot Pre_{lab} \cdot Rec_{lab}}{Pre_{lab} + Rec_{lab}}$$

Where $|NCLC|$ is the number of characters labeled correctly with their corresponding tags, $|NCL|$ is the number of characters labeled with tags, and $|NC|$ is the number of characters should be labeled.

4.3 Evaluation of Symptom Name Recognition Ability

Comprehensive evaluations of symptom name recognition ability using CRFs with reasonable features are shown in figure 5, 6 and 7. These figures show that CRFs with reasonable features for symptom name recognition from FCRs of TCM is feasible. The best RFM_{det} 62.829% (RR_{det} 93.403% and RER_{det} 52.665%) is achieved under settings $CWS = 3$ and features Uni+Big+Tri used.

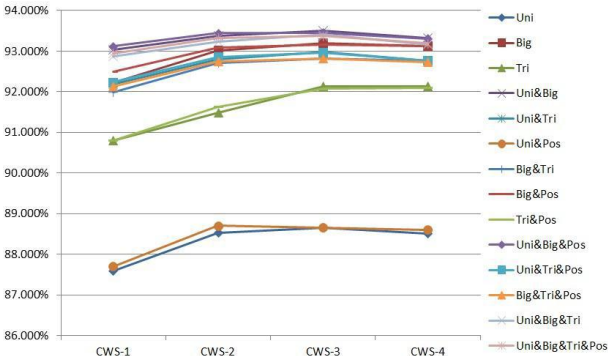

Figure 5. Symptom Name Recognition Rate.

It obviously shows in figures 5, 6 and 7 that literal features and positional features are helpful to symptom name recognition from FCRs of TCM. More types of features are used; better recognition performance would be obtained in most cases. When $CWS = 1$ and referred features changed from unigram literal features to the combination of unigram and bigram literal features, the highest growth about 3.925% of RFM_{det} is achieved (the

RR_{det} increases from 87.586% to 93.034% and the RER_{det} decreases from 56.173% to 53.118%).

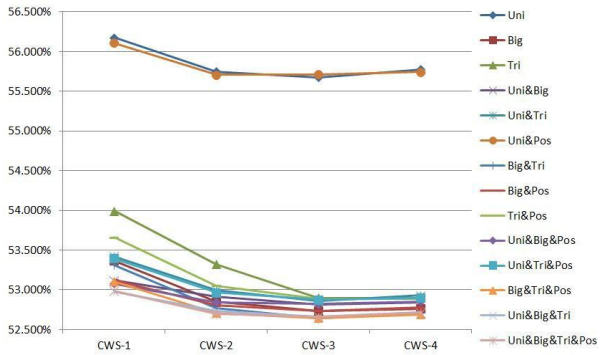

Figure 6. Symptom Name Recognition Error Rate.

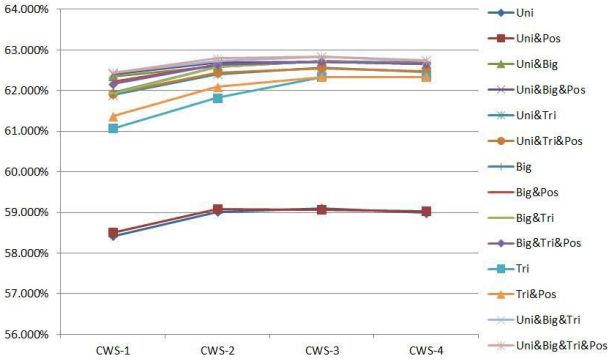

Figure 7. Symptom Name Recognition F-Measure.

As described previously, the context information is helpful to symptom name recognition. However, the context window size should not be too large. In figures 5, 6 and 7, it clearly shows that when CWS increase RR_{det} and RFM_{det} are improved and RFM_{det} is reduced. When CWS grows too large (larger than 3 here), RR_{det} and RFM_{det} begin, nevertheless, to descend and RER_{det} is raised in most every cases.

Moreover, positional features are complementary features to literal features for symptom name recognition from FCRs of TCM. It vividly shows in figures 5, 6 and 7 that RR_{det} and RFM_{det} would be improved and RER_{det} would be reduced more or less when literal features combined with positional features. The highest elevation can reach 0.297% if the combination features of trigram literal features and positional features are used and $\mathrm{CWS}=1$.

4.4 Evaluation of Labeling Performance and Detailed Analysis for Symptom Name Recognition

In this part, firstly, an evaluation for labeling performance is given, and then a detailed analysis for symptom name recognition from FCRs of TCM using CRFs with reasonable features would be described.

The results of Pre_{lab} and FM_{lab} under different situations are shown in figure 8 and 9, respectively. The Rec_{lab} here are all 100%. It can be seen from these figures that the FM_{lab} can reach nearly up to 97.596% with corresponding Pre_{lab} 95.305%. The results can also demonstrate the feasibility of the proposed method for symptom name recognition from FCRs of TCM and the worth of the representative and reasonable features introduced in this paper. The properties of literal features and positional features, which are just described in section 4.3, are also reflected in figures 8 and 9.

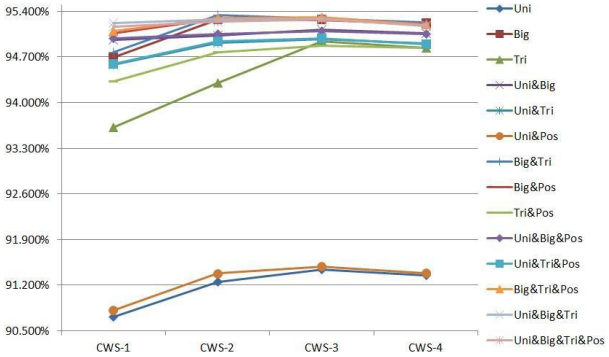

Figure 8. Results of Pre_{lab} under Different Situations.

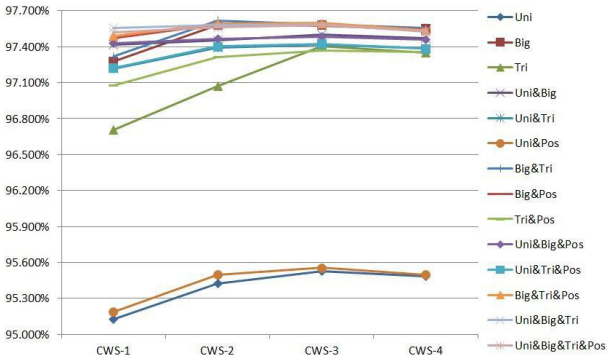

Figure 9. Results of FM_{lab} under Different Situations.

Although RR_{det} can achieve a very high performance, however, RER_{det} is also too high. In figures 8 and 9, high labeling results was gotten. It implies that the probable position of the symptom name can be found in TCM FCRs, but the exact boundaries of the symptom name descriptions cannot be detected accurately yet.

More careful results are listed in table 4. In this table, the average labeling Pre_{lab} of labels "B-

SYC" and "O-SYC" are always higher than the global average precision, but the average Pre_{lab} of "I-SYC" is lower than the global average precision. It implies that the performance of labeling the end position of a symptom name description is worse than the other position's. In other words, the judgment on whether "I-SYC" or "O-SYC" is more difficult. Therefore, as the future work, how to accurately determine the end of a symptom name description should be paid more attention to.

		CWS = 1	CWS = 2	CWS = 3	CWS = 4
Global P		94.186%	94.526%	94.616%	94.540%
B	P	95.184%	95.472%	95.519%	95.429%
	R	94.135%	94.243%	94.238%	94.113%
	F	94.656%	94.853%	94.873%	94.765%
I	P	93.085%	93.586%	93.772%	93.713%
	R	93.791%	94.181%	94.267%	94.201%
	F	93.434%	93.879%	94.016%	93.953%
O	P	94.533%	94.781%	94.819%	94.738%
	R	94.501%	94.916%	95.056%	94.996%
	F	94.514%	94.845%	94.934%	94.864%

Table 4. Detailed Results of Average Pre_{lab}, Rec_{lab} and FM_{lab} for Each Type of Labels. "B", "I" and "O" are short forms of "B-SYC", "I-SYC" and "O-SYC", respectively.

5 Conclusion

In this paper, a preliminary work on symptom name recognition from FCRs of TCM is described, and a feasible method based on CRFs with reasonable features is investigated. Through the experiments, the specialties, usage and effectiveness of the introduced features are verified.

In future, particular syntactic structure and grammatical rules for FCRs of TCM need to be defined and studied based on the characteristics of FCRs of TCM. On the one hand, they can help the TCM doctors and researchers to understand the clinical records deeper (Spasic, et al., 2005; Zhou, et al., 2010), and on the other hand, technically, they are good for filtering and reducing feature size and providing basics and adequate evidence for symptom name normalization process and automatic diagnosis procedure.

Acknowledgments

The authors would like to thank M.S. Xuehong Zhang and M.S. Shengrong Zhou for their helpful suggestions to this work and their valuable work on manually structuring the clinical records for us. The authors are grateful to Ms. Fang Yu and B.S. Yuheng Karen Chen for their helpful paper revising. The authors are also pleased to acknowledge the National Natural Science Foundation of China (Grant No. 61173182 and 61179071), the Provincial Science and Technology Foundation of Sichuan Province (Grant No. 2008SZ0049), the Specialized Research Fund for the Doctoral Program (Grant No. 20090181110052), and the New Century Excellent Talents Fund (Grant No. NCET-08-0370) for their supporting to this work.

References

P.M. Barnes, E. Powell-Griner, K. McFann, R.L. Nahin. 2004. Complementary and alternative medicine use among adults: United States, 2002. *Seminars in Integrative Medicine*, 2(2):54-71.

H. Duan, Y. Zheng. 2011. A study on features of the CRFs-based Chinese Named Entity Recognition. *International Journal of Advanced Intelligence*, 3(2):287-294.

G. Fu, K.K. Luke. 2005. Chinese named entity recognition using lexicalized HMMs. *SIGKDD Explorations*, 7(1):19-25.

J. Gao, M. Li, A. Wu, C.-N. Huang. 2005.Chinese word segmentation and named entity recognition: a pragmatic approach. *Computational Linguistics*, 31(4):531-574.

M. Huang, M. Chen. 2007. Integrated design of the intelligent web-based Chinese medical system (CMDS)-systematic development for digestive health. *Expert System with Applications*, 32:658-673.

J. Lafferty, A. McCallum, F. Pereira. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. *In Proceedings of the 18th International Conference on Machine Learning*.

D. Li, K. Kipper-Schuler, G. Savova. 2008. Conditioal Random Fields and Support Vector Machine for disorder named entity recognition in clinical texts. *In BioNLP 2008: Current Trends in Biomedical Natural Language Processing*, pp:94-95.

A. McCallum, W. Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. *In Proceedings of the 7th Conference on Natural Language Learning (CoNLL) at HLT-NAACL*.

M. Molassiotis, P. Fernadez-Ortega, D. Pud, G. Ozden, J.A. Scott, V. Panteli, A. Margulies, M. Browall, M.

Magri, S. Selvekerova, E. Madsen, L. Milovics, I. Bruyns, G. Gudmundsdottir, S. Hummerston, A. M.-A. Ahmad, N. Platin, N. Kearney, E. Pariraki. 2005. Use of complementary and alternative medicine in cancer patients: a European survey. *Annals of Oncology*, 16(4):655-663.

J.-J. Nie, J. Gao, J. Zhang, M. Zhou. 2000. 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*.

S.K. Pal. 2002. Complementary and alternative medicine: an overview. *Current Science*, 82(5):518-524.

L.A. Ramshaw, M.P. Marcus. 1995. Text chunking using transformation-based learning. *In Proceedings of the Third Workshop on Very Large Corpora. ACL*.

F. Sha, F. Pereira. 2003. Shallow parsing with conditional random fields. Proceedings of the 2003 Conference of the North American Chapter of the Association of Computer Linguistics on Human Language Technology.

I. Spasic, S. Ananiadou, J. McNaught, A. Kumar. 2005. Text mining and ontologies in biomedicine: making sense of raw text. *Briefings in Bioinformatics*, 6(3):239-251.

X. Wang, H. Qu, P. Liu. 2004. A self-learning expert system for diagnosis in traditional Chinese medicine. *Expert System with Applications*, 26:557-566.

Y. Wang, Z. Yu, Y. Jiang, K. Xu, X. Chen. 2010. Automatic symptom name normalization in clinical records of traditional Chinese medicine. *BMC Bioinformatics*, 11:40.

Y. Wang, Z. Yu, Y. Jiang, Y. Liu, L. Chen, Y. Liu. 2012. A framework and its empirical study of automatic diagnosis of traditional Chinese medicine utilizing raw free-text clinical records. *Journal of Biomedical Informatics*, 45:210-223.

Y. Wu, J. Zhao, B. Xu. 2003. Chinese named entity recognition combining a statistical model with human knowledge. *In Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-Language Named Entity Recognition* (*MultiNER'03*), pp:65-72.

K. Yoshida, J. Tsujii. 2007. Reranking for biomedical named-entity recognition. *In BioNLP 2007: Biological, translational, and clinical language processing*, pp:209-216.

H.-P. Zhang, Q. Liu, H.-K. Yu, X.-Q. Cheng, S. Bai. 2003. Chinese named entity recognition using role model. Computational Linguistics and Chinese Language Processing, 8(2):29-60.

N.L. Zhang, S. Yuan, Y. Wang. 2008. Latent tree models and diagnosis in traditional Chinese medicine. *Artificial Intelligence in Medicine*, 42:229-245.

J. Zhou, L. He, X. Dai, J. Chen. 2006. Chinese named entity recognition with a multi-phase model. *In Proceedings of the fifth Workshop on Chinese Language Processing*, pp:213-216.

X. Zhou, Y. Peng, B. Liu. 2010. Text mining for traditional Chinese medical knowledge discovery: a survey. *Joural of Biomedical Informatics*, 43:650-660.

G.D. Zhou, J. Su. 2002. Named entity recognition using an HMM-based Chunk Tagger. *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.