An Empirical Investigation on Fine-Grained Syndrome Segmentation in TCM by Learning a CRF from a Noisy Labeled Data

Yaqiang Wang, Dan Tang, and Hongping Shu
College of Software Engineering, Chengdu University of Information Technology, Chengdu, China
Email: {yaqwang, tangdan, shp}@cuit.edu.cn

Chen Su
Sichuan Academy of Chinese Medicine Sciences, Chengdu, China
Email: coin_0308@foxmail.com

***Abstract*—Syndrome is an important component in Traditional Chinese Medicine (TCM), and it is also a distinctive concept in TCM compared with Western Medicine (WM). Clearly understand the TCM syndrome help researchers digest TCM regularities and bridge TCM and WM. Syndromes are often used in coarse-grained forms, however fine-grained medical information buried in the coarse-grained TCM syndromes would not be considered. In this paper, we empirically investigate Fine-Grained Syndrome Segmentation (FGSS) in TCM by a distantly supervised method to build a noisy labeled data for training CRFs for FGSS in TCM. The feasibility and effectiveness of the method are demonstrated based on a series of elaborate experiments. The best F1-score can reach 0.9177. To the best of our knowledge, our work is the first to focus on fine-grained information extraction in Chinese medical texts.**

***Index Terms*—information extraction, distant supervision, biomedical natural language processing, traditional Chinese medicine**

I. Introduction

Syndrome in Traditional Chinese Medicine (TCM) concludes the etiology, location of the disease, pathology as well as relationship between right and evil pathogen. It is a distinctive concept in TCM comparing to Western Medicine (WM). A clearly understanding of TCM syndromes will help researchers digest TCM regularities and bridge the TCM and WM [1], [2]. Recently, many researchers work on knowledge discovery of syndrome differentiation in TCM [3]-[5].

Syndromes in TCM literature and clinical records are in coarse-grained forms. For example, syndrome "外感风热" (exogenous wind hot) is a syndrome concluded when a TCM practitioner observes a medical problem description "发热 (fever), 恶寒 (aversion to cold), 头痛 (headache), 脉浮数 (rapid floating pulse), 苔薄白 (moss thin and white), 舌红 (red tongue), 口渴 (thirst)". It is composed of three fine-grained segments, including "外感 (表证)" (exogenous (syndrome)), "风(证)" (wind (syndrome)), and "热(证)" (hot (syndrome)). These fine-grained syndromes are inferred based on many parts of the medical problem description, according to the diagnosis theory of TCM. To fetch the fine-grained information, coarse-grained syndromes need to be segmented into their fine-grained forms [6].

Most of researches on knowledge discovery related to TCM syndromes did not consider above fine-grained segmentation problem. Consequently, in this paper, we focus our attention on automatic fine-grained TCM syndrome segmentation, FGSS for short. The results will support other precision medical researches of TCM [2], [7], such as TCM syndrome differentiation and syndrome-disease relationship mining, syndrome-gene network analysis, etc.

FGSS in TCM can be naturally treated as a Chinese word segmentation task, and fine-grained segments of TCM syndromes can be naturally regarded as Chinese words. Then this segmentation task can be solved by utilizing supervised sequence labeling models, e.g. Hidden Markov Models [8], Maximum Entropy Markov Models [9], [10], Conditional Random Fields (CRFs) [11], [12]. However, supervised models suffer from a labor-intensive problem to build a labeled training data. To address this deficiency, semi-supervised models [13]-[15] are proposed in general domain. Although these methods reduce much cost of manual labor, labeled data is still a necessity. It is still a challenging task to manually build a domain-specifically used dataset due to a requirement of cross-domain knowledge.

In this paper, we propose a distantly supervised method for training CRFs for FGSS in TCM. The method takes advantages of naming and translating conventions of TCM syndromes. It firstly builds a noisy labeled training data through applying a probabilistic alignment model, which is used to exploit relations between Chinese characters and English words in a Chinese-English parallel corpus of TCM syndromes. This alignment model can infer high confident fine-grained segments in TCM syndromes by heuristically employing a forward

Manuscript received January 12, 2018; revised May 17, 2018.

doi: 10.12720/jait.9.2.45-50

direction alignment and a reversed direction check. Then, CRFs can be adopted and trained on this noisy labeled data. Experimental results demonstrate the feasibility and the effectiveness of the proposed method. The highest F1-score can reach 0.9177, which is a competitive result comparing to the result (0.939) obtained by a CRF trained based on a manually segmented data.

II. The Probabilistic Alignment Method

In this section, we propose a simple and effective probabilistic alignment method to confidently align Chinese characters in TCM syndromes to English words in the corresponding translations. The Chinese character(s) aligned to the same English word would form a fine-grained segment in a TCM syndrome. The alignment results are confident, because they passed a bidirectional confirmation process, i.e. a forward direction aligning and a reversed direction checking, which would be introduced in following sections, and the results of these two methods are heuristically combined together in order to infer the fine-grained segments.

A. Forward Direction Aligning Method

Forward direction aligning method completes a process that probabilistically aligning Chinese characters c_i in a TCM syndrome $\boldsymbol{c}$ with English words e_j in its translation $\boldsymbol{e}$. The alignment probability of e_j given c_i can be directly defined by a conditional probability with a Laplace smoothing [16] that is

$$p(e_j|c_i) = \frac{c(e_j,c_i) + \alpha}{c(c_i) + \alpha|V_c|} \quad (1)$$

where $c(e_j,c_i)$ is the co-occurrence frequency of e_j and c_i, $c(c_i)$ is the count of c_i, and $|V_c|$ is the amount of unique Chinese characters in the corpus. $\alpha >= 0$ is a smoothing parameter. A higher $p(e_j|c_i)$ indicates that compared to other English words in $\boldsymbol{e}$, e_j has a higher strength aligning to c_i.

B. Reversed Direction Checking Method

Reversed direction checking method is designed to inspect results of the forward direction aligning method, whether e_j aligning to c_i with high confidence. It means that compared to other Chinese characters in $\boldsymbol{c}$, c_i should also have a higher strength aligned with e_j. It can be defined as

$$p(c_i|e_j) = \frac{c(c_i,e_j) + \beta}{c(e_j) + \beta|V_e|} \quad (2)$$

In the equation, $c(c_i,e_j)$ has the same meaning as $c(e_j,c_i)$ in equation (1), $c(e_j)$ is the count of e_j, and $|V_e|$ is the amount of unique English words in the corpus. It is also smoothed by the Laplace, and $\beta >= 0$ is the smoothing parameter.

C. The Heuristic Inferece Method

The heuristic inference method is used to find the best fine-grained segmentation $\hat{\boldsymbol{s}}$ of $\boldsymbol{c}$ with the highest confidence. $\hat{\boldsymbol{s}}$ consists of $\hat{s}_k$, and $\hat{s}_k$ is composed by Chinese character(s), which are aligned with the same e_j in $\boldsymbol{e}$ with the highest confident alignment probability. The alignment probability is measured by a combination of the forward direction aligning probability and the reversed direction checking probability, and it is defined as follows.

$$\hat{a}_i = \arg\max_{a_i} (1-\lambda)\, p\left(c_i \middle| e_{a_i}\right) + \lambda\, p\left(e_{a_i} \middle| c_i\right) \quad (3)$$

where λ is the parameter that is used to balance the contributions of $p(c_i|e_j)$ and $p(e_j|c_i)$. If there is a tie in the inferring process, a heuristic strategy would be used that $\hat{a}_i$ will be set by the same value as previously adjacent alignment result $\hat{a}_{i-1}$, and if current position is the first place of $\boldsymbol{c}$ in the TCM syndrome, $\hat{a}_i$ will be empirically set to 1.

III. CRFs for FGSS in TCM

In Section II, we introduced a heuristic method, which help us recognize optimal fine-grained segments of TCM syndromes with the help of natural separators between English words in the translations of the TCM syndromes. The resulted segments could, consequently, make up a noisy labeled data for training CRFs for FGSS in TCM. In general, there are three key points with applying CRFs to a segmentation task, including segment representation [17], feature definition [18], and the parameter settings (or its implementation). The detailed information of these key points in this paper is present bellow.

A. The Segment Representation

The task of FGSS in TCM can be naturally defined as a sequence labeling process that labels each Chinese character with a label, which is used to represent which part of a fine-grained segment the Chinese character belongs to. There are five representative types of labels that are commonly used for indicating a specific part of a segment, including "B" (beginning of a segment), "I" (inside of a segment), "O" (outside of segments), "E" (end of a segment), and "S" (the single Chinese character type segment).

Specifically, in this paper, segment representation would like to include "B", "I" and "S". Firstly, there is no redundant content in TCM syndromes. Therefore, the segment representation "O" would be excluded in this paper. Secondly, owing to the concise style of descriptions of TCM syndromes, lengths of fine-grained segments in TCM syndromes are less than 3, and in most cases they are less than 2. Consequently, identifying the "end of fine-grained segments" individually would make little sense to improve the labeling performance, and sometimes, it would reduce the labeling performance due to the difficulty of rare class identification. Moreover, it has been verified that individually identifying single Chinese character type segments is helpful to lift labeling performance [17], [18], especially in our task many single Chinese character type segments are contained.

B. Feature Definition

The CRFs is a sophisticated discriminative model for sequence labeling tasks. Its good labeling performance is dependent on a well-tailored feature definition [19]. In this paper we investigated two types of features,

including n-gram features, specifically the unigram, bigram and trigram features, and n-gram features with a gap (i.e. the skip n-gram features).

N-gram features: It is not a trivial task to segment short and brief texts into words. Thus n-grams are usually used as their alternative to be the features. In this paper, we investigate the unigram, bigram and trigram features. Their usefulness has been demonstrated in the named entity recognition task in TCM [19]. For convenience, in this paper, we set an exact window size 3 to extract the features, because lengths of TCM syndromes are often short, usually shorter than 6.

Jumping n-gram features: Intuitively, bigram and trigram features with a gap may be helpful to FGSS in TCM. For example, a trigram feature with a gap in it "肾*虚" ("kidney * deficiency") could be used to determine that the Chinese character at the position of "*" would be a fine-grained segment, such as the fine-grained segments "阴" (yin) and "阳" (yang) in "肾阴虚" (kidney yin deficiency) and "肾阳虚" (kidney yang deficiency), respectively.

The templates of above introduced features are listed in Table I.

TABLE I. DEFINITIONS OF FEATURE TEMPLATES

Feature Type	Template
Unigram Feature (U)	c_{i-3}; c_{i-2}; c_{i-1}; c_i; c_{i+1}; c_{i+2}; c_{i+3}
Bigram Feature (B)	$c_{i-3}c_{i-2}$; $c_{i-2}c_{i-1}$; $c_{i-1}c_i$; c_ic_{i+1}; $c_{i+1}c_{i+2}$; $c_{i+2}c_{i+3}$
Trigram Feature (T)	$c_{i-3}c_{i-2}c_{i-1}$; $c_{i-2}c_{i-1}c_i$; $c_{i-1}c_ic_{i+1}$; $c_ic_{i+1}c_{i+2}$; $c_{i+1}c_{i+2}c_{i+3}$
Skip Bigram Feature (JB)	$c_{i-3}c_{i-1}$; $c_{i-2}c_i$; $c_{i-1}c_{i+1}$; c_ic_{i+2}; $c_{i+1}c_{i+3}$
Skip Trigram Feature (JT)	$c_{i-3}c_{i-2}c_i$; $c_{i-2}c_{i-1}c_{i+1}$; $c_{i-1}c_ic_{i+2}$; $c_ic_{i+1}c_{i+3}$; $c_{i-3}c_{i-1}c_i$; $c_{i-2}c_ic_{i+1}$; $c_{i-1}c_{i+1}c_{i+2}$; $c_ic_{i+2}c_{i+3}$;

C. CRFs Implementation

In our experiments, the CRF++ tool is adopted, which provides an efficient implementation for CRFs by using the limited-memory quasi-Newton algorithm for training the models [11], [20]. Our aim is to demonstrate the feasibility and effectiveness of the proposed method to build a noisy labeled data for training CRFs for FGSS in TCM, and the usefulness of the segment representations and features described previously. So the default settings of the CRF++ tool are directly used for simplicity.

IV. EXPERIMENTS

In this section, the proposed method for building a noisy labeled data for training a CRF for FGSS in TCM is verified. The results would present the performance of the proposed probabilistic alignment method and how noisy the built data is. Moreover, the performance of CRFs trained on the noisy labeled training data for FGSS in TCM under different settings is also validated. The results would illustrate the applicability of the noisy labeled training data and, at the same time, investigate the effectiveness of the announced segment representation and feature definition.

A. Experimental Datasets

In this paper, a standard Chinese-English parallel corpus of TCM syndromes is used as the basis data for building the noisy labeled training data for CRFs for FGSS in TCM. It contains 606 Chinese-English translation pairs of TCM syndromes. In order to evaluate the proposed probabilistic alignment method, the Chinese characters and English words in these translation pairs are manually aligned in advance. This dataset is named as "AD" (i.e., the alignment data).

TABLE II. DEFINITIONS OF FEATURE TEMPLATES

	AD	SD
Size	606	260
Number of Unique Chinese Characthers	187	109
Number of Unique Fine-Grained Segments	212	117
Number of Common Unique Fine-Grained Segments	80	

To illustrate the applicability of the noisy labeled training data and investigate the effectiveness of the aforementioned segment representation and feature definition, 260 clinical TCM syndromes are extracted from a clinical record dataset. The clinical record dataset contains 6722 clinical records, which are accumulated during the routine diagnostic work of TCM practitioners. None of these extracted clinical TCM syndromes exists in the parallel corpus. Moreover, TCM experts are hired to pre-segment these clinical TCM syndromes into their fine-grained forms manually. This dataset is named as "SD" (i.e., segmentation data).

More detailed information of the experimental datasets is listed in Table II.

B. Evaluation Metrics

We define three types of metrics for evaluating the proposed methods as follows.

Alignment a**C**curacy (**AC**): this metric is used for evaluating the alignment performance of the proposed probabilistic alignment method. In this paper, an alignment is correct, if and only if the aligned Chinese character is or is a part of a segment that the English word in the translation of the TCM syndrome corresponds to. Higher AC is achieved; better alignment performance would be obtained. The AC is defined as follows.

$$AC = \frac{c_{ca}}{c_a} \tag{4}$$

where c_{ca} is the count of the correctly aligned Chinese character and English word pairs, and c_a is the amount of aligned Chinese character and English word pairs.

Segmentation **P**recision (**SP**), **R**ecall (**SR**) and **F**1-score (**SF**): these metrics are used to validate the applicability of the resulted noisy labeled data for training CRFs for FGSS in TCM by the proposed probabilistic alignment method and the appropriateness of the segment representation and the defined features. If and only if the boundaries of a fine-grained segment are both recognized

accurately, the fine-grained segment is considered to be a correct result. Higher SP and higher SR are achieved; better global segmentation performance would be gotten. SP, SR and SF are formulated below.

$$SP = \frac{c_{sc}}{c_s} \tag{5}$$

$$SR = \frac{c_{sc}}{c_{sr}} \tag{6}$$

$$SF = \frac{2 \cdot SP \cdot SR}{SP + SR} \tag{7}$$

where c_{sc} is the count of segments segmented correctly, c_s is the count of segments segmented, and c_{sr} is the count of gold segments in the test data.

Labeling **P**recision (**LP**), **R**ecall (**LR**) and **F**1-score (**LF**): these metrics are utilized to assess the labeling performance of CRFs based FGSS in TCM in order to give a detailed analysis for investigating the appropriateness of the segment representations and the defined features. Higher LP and LR are achieved; better global labeling performance LF would be gotten. LP, LR and LF are defined by equations (8), (9) and (10), respectively.

$$LP_t = \frac{c_{lc,t}}{c_{l,t}} \tag{8}$$

$$LR_t = \frac{c_{lc,t}}{c_{lr,t}} \tag{9}$$

$$LF_t = \frac{2 \cdot LP_t \cdot LR_t}{LP_t + LR_t} \tag{10}$$

where the subscript t indicates which class label it is, e.g. "B", "I", "E", or "S". $c_{lc,t}$ is the count of Chinese characters correctly labeled with the class label t, $c_{l,t}$ is the count of Chinese characters labeled with the class label t, and $c_{lr,t}$ is the count of gold class label t in the test data.

C. Evaluation of the Alignment Results

In order to investigate the sensitivity of the proposed probabilistic alignment method to the smoothing parameter, we evaluate the forward direction aligning (FDA) results and the reversed direction checking (RDC) results separately under different smoothing parameter settings. The results are listed in Table III. It shows clearly that FDA prefers a larger smoothing parameter value than RDC, and the best parameter values are 1.0 and 0.1, respectively. Furthermore, the results demonstrate that the smoothing would be helpful to improve the accuracy of the pro-posed probabilistic alignment method.

To validate the performance of the proposed probabilistic alignment method, we verify the alignment accuracy of the heuristic inference method (HIM) described in Section II under different λ settings when the smoothing parameters α and β are set to 1.0 and 0.1, respectively, which are the parameter settings let FDA and RDC methods achieve the best alignment accuracy results. The results obtained by the HIM are listed in Table IV. It shows that a better result can be obtained through balancing the contributions of $p(c_i|e_j)$ and $p(e_j|c_i)$ by changing the value of λ, and the best result can reach 0.932, when λ = 0.1. Moreover, the results illustrate that RDC is important to the alignment task in this paper, and these results also conveys that the heuristically inferred segment data contains noisy.

TABLE III. RESULTS OBTAINED BY FDA AND RDC METHODS RESPECTIVELY UNDER DIFFERENT α AND β SETTINGS

α values	AC of FDA	β values	AC of RDC
0.0	0.4464	**0.0**	0.8017
0.1	0.6602	**0.1**	**0.9308**
0.2	0.6864	**0.2**	0.9271
0.3	0.6864	**0.3**	0.8884
0.4	0.6858	**0.4**	0.8853
0.5	0.6852	**0.5**	0.884
0.6	0.6908	**0.6**	0.8828
0.7	0.6926	**0.7**	0.8749
0.8	0.6926	**0.8**	0.8747
0.9	0.6914	**0.9**	0.8716
1.0	**0.6933**	**1.0**	0.8709

TABLE IV. COPYRIGHT FORMS AND RE RESULTS OBTAINED BY HIM UNDER DIFFERENT λ SETTINGS RESPECTIVELY WHEN α AND β ARE SET TO 1.0 AND 0.1

λ values	AC of HIM
0.0	0.9308
0.1	**0.932**
0.2	0.899
0.3	0.8953
0.4	0.8953
0.5	0.8797
0.6	0.861
0.7	0.8479
0.8	0.8074
0.9	0.7431
1.0	0.6933

D. Evaluation of the Alignment Results

The FGSS results obtained by CRFs trained on the noisy labeled data under different segment representation and feature definition settings are shown in Fig. 1. The best SF can reach 0.9177, when "BIS" segment representation and "UB" features are used. It shows that the noisy labeled data can be used for training CRFs for FGSS in TCM.

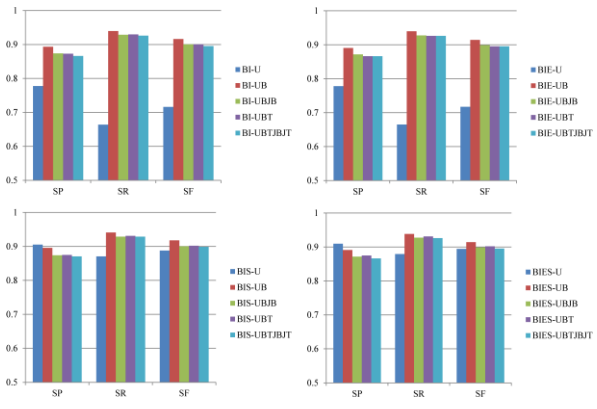


Figure 1. Results of SP, SR and SF obtained by CRFs trained on the noisy labeled training data under different segment representation and feature definition settings.

In Fig. 1, we can also see that "BIS" is the most suitable segment representation to CRFs based FGSS in TCM. Under the same feature definition setting, "BIS" can achieve better results than other type of segment representations in most cases, except under the "U" (worse than "BIES") and "UBT" (equal to "BIES") feature definitions.

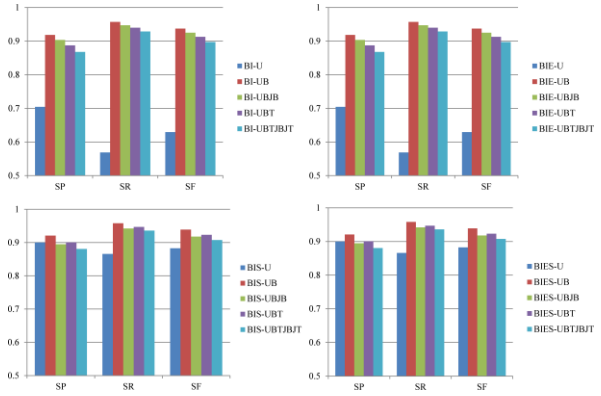


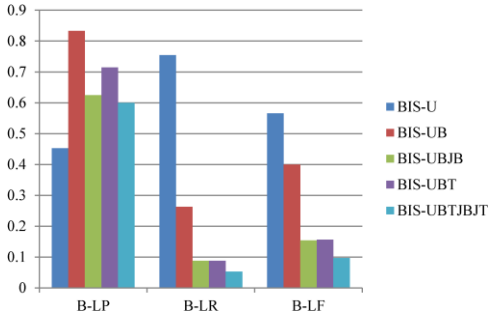
Figure 2. Results of SP, SR and SF obtained by CRFs trained on the labeled training data generated by the gold aligned data under different segment representation and feature definition settings.

It is also shown in Fig. 1 that, to the CRFs based FGSS in TCM, it is enough to just using the unigram and bigram features in the discriminative process. Under the same segment representation, when combining unigram and bigram features, the results are always better than other conditions.

In order to further validate the usability of the noisy labeled training data to learn a CRF for FGSS in TCM, we run the same experiment again based on the labeled training data generated by the gold aligned data, and the results are shown in Fig. 2. We can see that, under the same settings, the results obtained by the CRFs trained on the noisy labeled data can achieve competitive SFs with the results obtained by the CRFs trained on the labeled data generated by the gold aligned data. More importantly, our method can learn a CRF for FGSS in TCM without additional manual labor for building a training data.

E. Detailed Analysis of the FGSS Results

Due to the space limitation, we just show the detailed labeling results, which are obtained by the CRFs trained on the noisy labeled data under the "BIS" segment representation and different feature definition settings, in Fig. 3. The conclusions on these results obtained under other types of the segment representations are similar to the results shown by these figures.



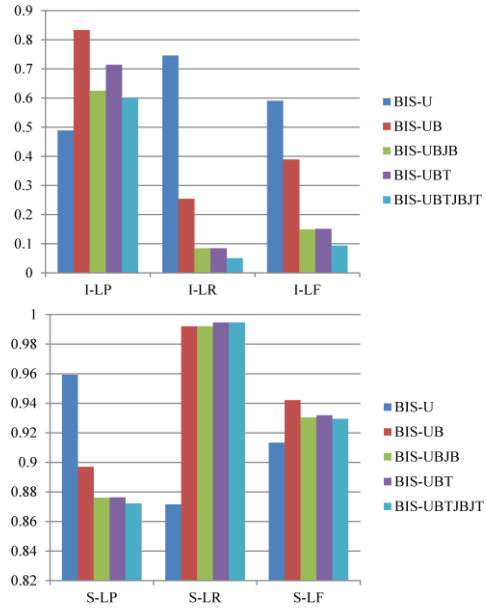


Figure 3. Results of LP, LR and LF of "B", "I" and "S" obtained by CRFs trained on the noisy labeled training data under the condition of "BIS" as the segment representation and different feature definition used.

It clearly shows in Fig. 3 that the results of LP, LR and LF of "S" are all better than the performance achieved by "B" and "I". It may be due to the cause that long fine-grained segments are greatly less than the single Chinese character type of fine-grained segments. Thus the distribution of labels is skewed, and it will result in the difficulty of recognizing "B" and "I".

Furthermore, in Fig. 3 we can also see that LP results of "B" and "I" are always higher than the LR results under the same condition. In contrast, the results of "S" have an opposite conclusion. It reminds that we should pay more attention on improving LR results for recognizing "B" and "I" and, at the same time, enhancing LR results for "S" in future work.

V. Experiments

In this paper, we empirically investigated the problem of fine-grained segmentation of TCM syndromes. A distantly supervised method is proposed to learn CRFs for FGSS in TCM. A series of experiments have done, and the feasibility and effectiveness of the proposed method were preliminarily proved in this paper. In future work, in order to further verify its practicability, the method would be applied on a larger scale real world data in different situations. Moreover, if we do not have an available standard parallel corpus, could we use large-scale translation software generated noisy data as the input and how to deal with it? Many other new research problems would produce when fine-grained information extracted, and how to generate clinical syndromes of TCM based on the fine-grained syndrome segments in TCM could be explored in future.

Acknowledgment

Authors are pleased to acknowledge the National Natural Science Foundation of China (Grant No.

61501063), the Scientific Research Foundation of Science and Technology Department of Sichuan Province (Grant No. 2016JY0240), the Young Talent Research Start-up Foundation of Chengdu University and Information Technology (Grant No. KYTZ201638) and the Young Academic Leader Research Foundation of Chengdu University and Information Technology (Grant No. J201705); Corresponding author: Yaqiang Wang, Email: yaqwang@ cuit.edu.cn.

REFERENCES

[1] W. Wang, H. Zhao, J. Chen, J. Chen, and G. Xi, "Bridge the gap between syndrome in traditional Chinese medicine and proteome in western medicine by unsupervised pattern discovery algorithm," in *Proc. IEEE International Conference on Networking, Sensing and Control*, 2008, pp. 1–750.

[2] P. Gu and H. Chen, "Modern bioinformatics meets traditional Chinese medicine," *Briefings in Bioinformatics*, vol. 15, no. 6, pp. 984–1003, 2013.

[3] X. Zhou, Y. Peng, and B. Liu, "Text mining for traditional Chinese medical knowledge discovery: A survey," *Journal of Biomedical Informatics*, vol. 43, no. 4, pp. 650–660, 2010.

[4] Y. Wang, Z. Yu, Y. Jiang, Y. Liu, L. Chen, and Y. Liu, "A framework and its empirical study of automatic diagnosis of traditional Chinese medicine utilizing raw free-text clinical records," *Journal of Biomedical Informatics*, vol. 45, no. 2, pp. 210–223, 2012.

[5] H. Wang, X. Liu, B. Lv, F. Yang, and Y. Hong, "Reliable multi-label learning via conformal predictor and random forest for syndrome differentiation of chronic fatigue in traditional Chinese medicine," *PloS ONE*, vol. 9, no. 6, p. e99565, 2014.

[6] M. Jiang, C. Lua, C. Zhang, J. Yang, Y. Tan, A. Lu, and K. Chan, "Syndrome differentiation in modern research of traditional Chinese medicine," *Journal of Ethnopharmacology*, vol. 140, no. 3, pp. 634–642, 2012.

[7] Y. Feng, Z. Wu, X. Zhou, Z. Zhou, and W. Fan, "Knowledge discovery in traditional Chinese medicine: State of the art and perspectives," *Artificial Intelligence in Medicine*, vol. 38, no. 3, pp. 219–236, 2006.

[8] W. J. Teahan, Y. Wen, R. McNab, and I. H. Witten, "A compression-based algorithm for Chinese word segmentation," *Computational Linguistics*, vol. 26, no. 3, pp. 375–393, 2000.

[9] A. McCallum, D. Freitag, and F. Pereira, "Maximum entropy markov models for information extraction and segmentation," in *Proc. 17th International Conference on Machine Learning*, 2000.

[10] N. Xue and L. Shen, "Chinese word segmentation as LMR tagging," in *Proc. Second SIGHAN Workshop on Chinese Language Processing*, 2003, pp. 176–179.

[11] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. International Conference on Machine Learning*, 2001, pp. 282–289.

[12] H. Zhao, C. N. Huang, and M. Li, "An improved Chinese word segmentation system with conditional random field," in *Proc. Fifth SIGHAN Workshop on Chinese Language Processing*, 2006, pp. 162–165.

[13] F. Jiao, S. Wang, C. H. Lee, R. Greiner, and D. Schuurmans, "Semisupervised conditional random fields for improved sequence segmentation and labeling," in *Proc. 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 2006, pp. 209–216.

[14] W. Sun and J. Xu, "Enhancing Chinese word segmentation using unlabeled data," in *Proc. Conference on Empirical Methods in Natural Language Processing*, 2011, pp. 970–979.

[15] X. Zeng, D. F. Wong, L. S. Chao, and I. Trancoso, "Graph-based semi-supervised model for joint chinese word segmentation and part-of-speech tagging," in *Proc. 51st Annual Meeting of the Association for Computational Linguistics*, 2013, pp. 770–779.

[16] Q. Yuan, G. Cong, and N. M. Thalmann, "Enhancing naive Bayes with various smoothing methods for short text classification," in *Proc. 21st International Conference on World Wide Web*, 2012, pp. 645–646.

[17] H. C. Cho, N. Okazaki, M. Miwa, and J. Tsujii, "Named entity recognition with multiple segment representations," *Information Processing and Management*, vol. 49, no. 4, pp. 954–965, 2013.

[18] H. Zhao, C. N. Huang, and B. L. Lu, "A unified character-based tagging framework for Chinese word segmentation," *ACM Transactions on Asian Language Information Processing*, vol. 9, no. 2, pp. 1–32, 2010.

[19] Y. Wang, Y. Liu, Z. Yu, L. Chen, and Y. Jiang, "A preliminary work on symptom name recognition from free-text clinical records of traditional Chinese medicine using conditional random fields and reasonable features," in *Proc. Workshop on Biomedical Natural Language Processing*, 2012, pp. 223–230.

[20] F. Sha and F. Pereira, "Shallow parsing with conditional random fields," in *Proc. Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 2003, pp. 134–141.

Yaqiang Wang received his PhD degree in 2013 from College of Computer Science, Sichuan University. His MS and BS degree were obtained from Sichuan University in 2010 and CUIT in 2007, respectively.
Yaqiang is now a Lecture of Chengdu University of Information Technology. He elected as a Master advisor of Chengdu University of Information Technology in 2016. He is a member of ACL and CCF. He received Outstanding Reviewer award twice from Journal of Biomedical Informatics in 2014 and 2017, respectively. Currently, his research interests are natural language processing, machine learning, and biomedical informatics.

Dan Tang is a Professor of Chengdu University of Information Technology. Currently, his research interests are data mining and information security.

Hongping Shu is a Professor of Chengdu University of Information Technology. Currently, his research interests are big data mining and software engineering.

Chen Su is an Associate Professor of Sichuan Academy of Chinese Medicine Sciences. Currently, her research interests are big data mining and biomedical informatics.