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

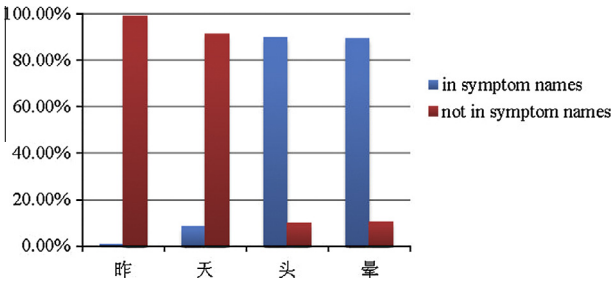

Fig. 1. An example of the differences between the distributions of the words used in symptom names and outside of symptom names.