Journal of Biomedical Informatics 48 (2014) 130–136

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Extracting important information from Chinese Operation Notes with natural language processing methods

Hui Wang [a,b,c], Weide Zhang [d], Qiang Zeng [e], Zuofeng Li [e], Kaiyan Feng [f], Lei Liu [a,b,e,*]

[a] Shanghai Public Health Clinical Center, Institutes of Biomedical Sciences, and Key laboratory of Medical Molecular Virology, Ministry of Education and Health, Fudan University, Shanghai, China
[b] Key Laboratory of Medical Imaging Computing and Computer Assisted Intervention of Shanghai, Fudan University, Shanghai, China
[c] Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China
[d] Zhongshan Hospital, Fudan University, Shanghai, China
[e] Shanghai Center for Bioinformatics Technology, Shanghai, China
[f] BGI-Shenzhen, Shenzhen, China

ARTICLE INFO

Article history:
Received 4 January 2013
Accepted 13 December 2013
Available online 31 January 2014

Keywords:
Clinical operation notes
Information extraction
Chinese EMR
Conditional random fields

ABSTRACT

Extracting information from unstructured clinical narratives is valuable for many clinical applications. Although natural Language Processing (NLP) methods have been profoundly studied in electronic medical records (EMR), few studies have explored NLP in extracting information from Chinese clinical narratives. In this study, we report the development and evaluation of extracting tumor-related information from operation notes of hepatic carcinomas which were written in Chinese. Using 86 operation notes manually annotated by physicians as the training set, we explored both rule-based and supervised machine-learning approaches. Evaluating on unseen 29 operation notes, our best approach yielded 69.6% in precision, 58.3% in recall and 63.5% *F*-score.

1. Introduction

Clinical documents contain a wealth of information for medical study. It has been advocated that electronic medical record (EMR) adoption is a key to solving problems related to quality of care, clinical decision support, and reliable information flow among individuals and departments participating in patient care [1]. But a large part of EMRs' data is saved in an unstructured textual format (such as discharge summaries and progress reports) which presents a big challenge for automated text mining. Manually annotating such narratives by domain experts is definitely a time consuming and error-prone process. Therefore, extracting relevant data elements from clinical narratives constitutes a basic enabling technology to unlock the knowledge within and support more advanced reasoning applications such as diagnosis explanation, disease progression modeling, and intelligent analysis of the effectiveness of treatment [2].

Recent years we have seen rapid adoption of hospital information system across China and medical documents in Chinese are accumulating fast. Realizing the potential of Chinese EMRs is a burgeoning field of research. A lot of approaches have been developed for English medical language processing, but studies focusing on Chinese are relatively limited. In this paper we have tried both rule-based method and sequential labeling algorithm which have been applied on English successfully. The structured representation of medical concepts and values could enable physicians a quick abstraction of patients' pathological status and also offers great convenience for large scale analysis. The results can also provide us with a lot of insights on how to design extracting models according to Chinese own characteristics.

2. Background

In the past 20 years, a number of tools and systems have been developed specifically for information extraction from clinical documents [3–10]. These systems and methods have been applied to many different tasks such as adverse events detection [11], abstraction of family history from discharge summaries [12], medication information extraction [13,14], etc. Due to the casualty and conciseness of clinical narratives, methods for fine-grained demand such as negation detection [15], coreference resolution [16], and ontology techniques [17,18] are also explored. Meystre et al. [19] presented a detailed review for extracting information from textual documents in EMRs.

The i2b2 [20] organizers have held a series of shared tasks focusing on biomedical informatics since 2006. The fourth i2b2/VA

* Corresponding author at: Shanghai Public Health Clinical Center, Institutes of Biomedical Sciences, and Key laboratory of Medical Molecular Virology, Ministry of Education and Health, Fudan University, Shanghai, China.
E-mail address: liulei@fudan.edu.cn (L. Liu).

http://dx.doi.org/10.1016/j.jbi.2013.12.017
1532-0464/

shared-task and workshop [21] dealing with extraction of medical concepts, assertions and relations in clinical text was quite informative for our study. A number of novel approaches [22–24] were proposed by worldwide participants. Studies on Chinese NLP started in recent 10 years. Research groups at Stanford University have developed several software focusing on Chinese word segmentation [25,26]. Their tools rely on a linear-chain conditional random filed (CRF) model, which treats word segmentation as a binary decision task. ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) [27] is an integrated Chinese lexical analysis system that uses an approach based on multi-layer HMM. ICTCLAS includes word segmentation, Part-Of-Speech tagging and unknown words recognition. Both precision and recall rates reached above 90%. Topics like Chinese named entity recognition has also been investigated, Jiang and his colleagues did a preliminary work on symptom recognition from traditional Chinese medicine records and proposed some reasonable features for machine learning models [28]. Zhao et al. reported their findings on which kind of tokens that should be taken as the graininess in NER task, characters or words [29]. Group from Zhejiang University has conducted several projects on extracting temporal relation [30] from Chinese narrative medical records and terms and negation detection [31]. Based on CRF model, they explored different templates with 63 annotated documents and achieved 86.94% accuracy in extracting temporal attributes and almost 100% accuracy in detecting negations. Researchers from Microsoft Research Asia established an annotated corpus of Chinese discharge summaries and conducted word segmentation and named entity recognition [32]. They improved the performance of both tasks by using combined techniques called dual decomposition.

Our problem is not much alike named entity recognition. What we want to extract is values for our pre-defined concepts or attributes. For example, attributes like tumor size recorded in operation note is very important to clinicians, so we expect to get (attribute: value) pair like (tumor size: $2 * 3 * 3$ cm) automatically when given a paragraph of plain text.

These attributes are questions frequently inquired by doctors. MedLEE [33] and MedKATp [34] have similar functions as our method has, but their approaches are rule-based. No machine learning method has ever been applied to solve the problems mentioned above. Our unique contribution in this paper is an extracting strategy based on keyword search, extracting answers by sequential tagging results. With this method, doctors can effectively obtain structured data from free-text operation notes.

3. Methods

3.1. Data collection and preprocessing

Clinical documents we used for developing and testing our approaches were operation notes of hepatic carcinomas. We obtained a total of 115 electronic medical records from Zhongshan Hospital affiliated to Fudan University. They came from 115 individual patients who were admitted between July and November in 2008. The original EMRs from the database of the hospital contained all information of patients such as basic information, operation notes, and discharge summaries and so on. We converted the de-identified EMRs into a format of XML (each content has a tag as identifier) and isolated the operation notes from other contents. Then, 86 samples were randomly selected for training extracting models and the rest 29 operation notes were left for evaluation. Fig. 1 shows a sample of operation notes we used.

3.2. Data elements to extract and annotation method

After extensive consultation with medical researchers and doctors from hepatic department of Zhong Shan hospital, we identified 12 data elements doctors wanted to get from a free-text operation note. These data were key information of operation details and usually highly related to patients' pathological status. They would be of great value for clinical studies of hepatic carcinoma if they can be automatically processed into structured format. These data elements were targets of our extracting system and they are presented as 12 questions shown in Table 2.

Two doctors were recruited to annotate these 12 elements manually in all 115 samples. We used Protégé [35] (version 3.3.1) to establish ontology for each clinical entity to be extracted from the operation notes. A plug-in called Knowtator [36] recorded each entity's location in the documents. A third clinical researcher was in charge of dealing with inconsistence between the two annotations. Then the annotated dataset were used as gold standard for constructing and validating our models. Table 1 lists the inter-annotator agreement results compared to the gold standard.

There were 961 entities annotated, 704 in training dataset and 257 in test dataset. For each questions, there were at least 20 samples for training, and 20 for testing. An annotation sample is shown in Fig. 2.

3.3. Extraction strategy

When extracting knowledge manually from EMRs, people usually search for some keywords which related to (or indicate) the concepts they concern. For example, when looking up whether the patient has tumor thrombus, one will first locate the word "thrombus" and then scan for answers from its context, neighboring words of this keyword. Our extracting strategy is to simulate this procedure by computer, so our method is a two-step process. The whole pipeline of our system can be found in Fig. 3.

First we locate the attributes by identifying related keywords. According to doctors' suggestions we picked one to three keywords manually for each question. These keywords were generated purely based on doctors' clinical knowledge and experience and

麻醉成功后，患者右侧抬高 45 度位，术野皮肤常规消毒铺巾，右侧肋缘下弧形切口逐层切开进腹。探查腹腔无腹水，胃、肠、胆囊、胰、脾及盆腔脏器无异常，肝门淋巴结无肿大，门静脉主干无栓子。肝硬化结节 0.3cm，肿瘤分别位于右叶 VI 段、VIII 段，直径为 5cm 和 2cm，术中 B 超证实肿瘤位置。　　上肝拉钩，切断肝圆韧带，断端结扎＋缝扎，切断肝镰状韧带、右冠状韧带、三角韧带、肝胃韧带。电刀标记预切线，沿切线血管钳法分离肝组织，断面所遇一切管道均予切断、结扎必要时缝扎，逐渐深入直至将肿瘤完整切除，缝扎肝断面出血点，冲洗，检查肝断面无胆漏，肝缝线关闭肝断面。同理切除位于 VIII 段肝肿瘤，血管缝线缝合肝静脉分支破口，缝扎肝断面出血点，冲洗，检查肝断面无胆漏，涂抹生物胶水，上覆止血纱布。清理腹腔无活动性出血，敷料器械完整无缺少，于右膈下放置乳胶管引流一枚于腹壁另口引出，逐层关腹术终。　　手术顺利，术中出血 600ml，术中无输血，术中肝门未阻断，病人安返病房，切除组织送病理检查。

Fig. 1. An operation note sample written in Chinese free-text.