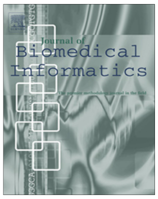
Journal of Biomedical Informatics 69 (2017) 203–217

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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



Building a comprehensive syntactic and semantic corpus of Chinese clinical texts

Bin He [a], Bin Dong [b], Yi Guan [a,*], Jinfeng Yang [c], Zhipeng Jiang [a], Qiubin Yu [d], Jianyi Cheng [a], Chunyan Qu [a]

[a] *School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China*
[b] *Ricoh Software Research Center (Beijing), Beijing, China*
[c] *School of Software, Harbin University of Science and Technology, Harbin, China*
[d] *Medical Records Room, The Second Affiliated Hospital of Harbin Medical University, Harbin, China*

ARTICLE INFO

Article history:
Received 7 November 2016
Revised 6 April 2017
Accepted 7 April 2017
Available online 9 April 2017

Keywords:
Chinese clinical texts
Corpus construction
Guideline development
Annotation method
Natural language processing

ABSTRACT

Objective: To build a comprehensive corpus covering syntactic and semantic annotations of Chinese clinical texts with corresponding annotation guidelines and methods as well as to develop tools trained on the annotated corpus, which supplies baselines for research on Chinese texts in the clinical domain.
Materials and methods: An iterative annotation method was proposed to train annotators and to develop annotation guidelines. Then, by using annotation quality assurance measures, a comprehensive corpus was built, containing annotations of part-of-speech (POS) tags, syntactic tags, entities, assertions, and relations. Inter-annotator agreement (IAA) was calculated to evaluate the annotation quality and a Chinese clinical text processing and information extraction system (CCTPIES) was developed based on our annotated corpus.
Results: The syntactic corpus consists of 138 Chinese clinical documents with 47,426 tokens and 2612 full parsing trees, while the semantic corpus includes 992 documents that annotated 39,511 entities with their assertions and 7693 relations. IAA evaluation shows that this comprehensive corpus is of good quality, and the system modules are effective.
Discussion: The annotated corpus makes a considerable contribution to natural language processing (NLP) research into Chinese texts in the clinical domain. However, this corpus has a number of limitations. Some additional types of clinical text should be introduced to improve corpus coverage and active learning methods should be utilized to promote annotation efficiency.
Conclusions: In this study, several annotation guidelines and an annotation method for Chinese clinical texts were proposed, and a comprehensive corpus with its NLP modules were constructed, providing a foundation for further study of applying NLP techniques to Chinese texts in the clinical domain.

1. Introduction

Electronic medical records (EMRs) represent the storage of all healthcare data and information in electronic formats [1] and constitute core data in the implementation of health care services. These services are undergoing enormous changes with increasing health awareness and demand for medical services. The situation is becoming more urgent for China, a country with the largest population but limited medical resources. In facing these challenges, the Chinese Ministry of Health (MOH) has issued a series of relevant regulations since 2010 to standardize EMR systems and their intelligent support [2–4]. With the rapid popularization of EMRs, the development of healthcare services has a solid data foundation.

Clinical texts, an important type of patient data within EMRs, are free-text documents that contain large amounts of information about patients' medical activities. In recent years, natural language processing (NLP) techniques on English clinical texts have been widely used [5,6] and many resources have been established for the development of these techniques. For example, the Unified Medical Language System (UMLS) [7], an integrated knowledge base of biomedical concepts, is widely applied in medical informatics research. Moreover, challenges organized by Informatics for Integrating Biology & the Bedside (i2b2) have released various kinds of annotated data for medical information extraction tasks, and enable clinical researchers to employ these clinical corpora for discovery research [8].

[*] Corresponding author at: Room 803, Zonghe Building, Harbin Institute of Technology, No. 92, West Dazhi Street, Harbin, Heilongjiang, China.
E-mail address: guanyi@hit.edu.cn (Y. Guan).

http://dx.doi.org/10.1016/j.jbi.2017.04.006
1532-0464/© 2017 Published by Elsevier Inc.

However, due to the lack of an annotated corpus, NLP research on Chinese clinical texts is still at a preliminary stage. Chinese clinical text has sublanguage features [9] that make it difficult for research on general-domain texts to be applied directly to clinical texts. In this study, we focus our efforts on conducting syntactic and semantic annotations of Chinese clinical texts, involving two resident physicians (P1 and P2) and eight annotators with backgrounds in computational linguistics (CL1-CL8). To our knowledge, this is the first comprehensive Chinese clinical corpus that includes several types of syntactic and semantic annotations, making it possible to develop effective NLP techniques for application to Chinese texts in the clinical domain.

This paper has six sections and is organized as follows: background on NLP research on clinical texts is summarized in Section 2; we then describe the development of annotation guidelines, annotation method, and annotation quality measurement in Section 3; next, Section 4 presents inter-annotator agreement (IAA) scores, data analysis of the annotations, and system development based on this corpus; in Section 5, we describe the contributions of this work and identify further improvements for future work.

2. Background

NLP tasks can be divided into low-level tasks and higher-level tasks: low-level tasks include sentence boundary detection, tokenization, word segmentation, part-of-speech (POS) tagging, shallow parsing, and so on; based on low-level tasks, higher-level tasks include named entity recognition (NER), negation identification, relationship extraction, etc. [10]. The annotated corpus is one of the fundamental points to the development of these NLP techniques. In the general domain, some publicly available corpora have a considerable effect, such as Penn Treebank [11–13], the CoNLL 2003 corpus [14], the ACE 2005 dataset [15], and the SemEval-2010 Task 8 dataset [16]. Similarly, there are a number of annotated corpora in the biomedical domain, such as the GENIA corpus [17], the PennBioIE corpus [18], the Yapex corpus [19], the GENETAG corpus [20], the CRAFT corpus [21], the BioText data [22], and the ITI TXM corpus [23]. Moreover, Table 1 summarizes some major annotated corpora in the clinical domain. In this study, our goal is to build a comprehensive corpus of clinical texts, therefore, the corpora listed in Table 1 will be described in detail below.

2.1. Annotated clinical text corpus for low-level tasks

2.1.1. Current status in English clinical texts

The Mayo Clinic's cTAKES system aims at comprehensive processing of clinical texts and covers various NLP techniques [6]. In this work, a linguistic corpus annotated for POS tagging and shallow parsing was accomplished by three linguistic experts via extending the Penn TreeBank (PTB) annotation guidelines [12,13] to the clinical domain. Additionally, Albright et al. [24] constructed a corpus involving annotations of POS tags and syntactic trees, and its advantage is that multilayer annotations are carried out in each sentence, which is beneficial in training joint models. As Albright et al. pointed out, the sentences in clinical texts contain numerous patterns that do not appear in the bracketing guidelines for the PTB [13], and clinical texts have sublanguage properties [38,39]. Therefore, Fan et al. [25] developed annotation guidelines for parsing clinical texts and annotated a syntactic corpus of progress notes from the University of Pittsburgh Medical Center (UPMC).

2.1.2. Current status in Chinese clinical texts

Word segmentation is an initial processing step in low-level tasks on Chinese texts. Xu et al. [26] found that out-of-vocabulary words and resolving ambiguities in clinical texts brought great challenges to word segmentation and that a state-of-the-art Chinese word segmenter trained by a general corpus would have poor performance in the clinical domain. Therefore, they manually annotated a corpus of segmented words in discharge summaries to improve the performance of word segmenters in Chinese clinical texts. Analogously, Zhang et al. [27] constructed similar corpus to achieve better word-embedding features.

2.2. Annotated clinical text corpus for higher-level tasks

2.2.1. Current status in English clinical texts

In 2006, Meystre and Haug [28] constructed an entity corpus involving 80 different medical problems with their assertions to

Table 1
Clinical text corpora for research on low-level and higher-level NLP tasks.

Part A

Author	Year	Language	Scale	Chinese word segmentation	POS tagging	Shallow parsing	Full parsing
Savova et al. [6]	2010	English	273 documents	–	√	√	–
Albright et al. [24]	2013	English	13,091 sentences	–	√	√	√
Fan et al. [25]	2013	English	1100 sentences	–	√	√	√
Xu et al. [26]	2014	Chinese	336 documents	√	–	–	–
Zhang et al. [27]	2016	Chinese	100 documents	√	–	–	–

Part B

Author	Year	Language	Scale	Entities	Assertions	Relations
Meystre et al. [28]	2006	English	160 documents	√	√	–
Roberts et al. [29]	2009	English	150 documents	√	√	√
Savova et al. [6]	2010	English	160 documents	√	√	–
Uzuner et al. [30]	2011	English	826 documents	√	√	√
Albright et al. [24]	2013	English	13,091 sentences	√	√	–
Elhadad et al. [31]	2015	English	531 documents	√	√	–
Xu et al. [26]	2014	Chinese	336 documents	√	–	–
Lei et al. [32]	2014	Chinese	800 documents	√	–	–
Wang et al. [33]	2014	Chinese	11 613 CCs	√	–	–
Wang et al. [34]	2014	Chinese	115 documents	√	–	–
Jia et al. [35]	2014	Chinese	30 documents	√	√	–
Xu et al. [36]	2015	Chinese	500 HPIs	√	√	–
Li et al. [37]	2015	Chinese	1000 documents	√	–	√

"√" means annotated, and "–" means unannotated. POS, part-of-speech; CC, chief complaint; HPI, history of present illness.