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.

judge whether a medical problem is *present* or *absent*, and 10 clinical document types were annotated. However, this corpus was somewhat limited in that only medical problems and two kinds of entity assertions were annotated. To extract further information from clinical texts automatically, Roberts et al. [29] randomly chose 50 clinical narratives, 50 histopathology reports, and 50 imaging reports to annotate entities, relations, modifiers, co-references, and temporal information in the CLinical E-Science Framework (CLEF) project [40]. This was the first corpus that extended the number of entity types to six, and was the first attempt at annotating relations and temporal information in clinical texts. Moreover, an iterative approach was used to develop annotation guidelines, and this greatly inspired subsequent work to build high-quality corpora in the clinical domain. Besides, Savova et al. built a named entity corpus [6] that included disorder entities with attached UMLS concept unique identifiers (CUI) and assertions that are of the types *negated*, *current*, *history of*, *family history of*, and *possible*. This corpus has contributed towards the development of cTAKES system, which brings enormous benefits to subsequent clinical text studies. In 2010, Uzuner et al. [41] released a corpus that annotated concepts, assertions, and relations. Based on semantic types defined in UMLS, concepts are classified into medical problems, tests, and treatments; meanwhile, there are six types of assertions for medical problems and three groups of relations between concepts. Furthermore, the annotation guidelines [42–44] in this corpus are of great importance for corpus construction in the clinical domain. However, diseases and symptoms, which are treated differently in medical practice, are not subdivided in this corpus but are merged into medical problems. In fact, Uzuner et al. [45] split medical problems into diseases and symptoms in a study before the i2b2 2010 challenge. Considering the differences between disorders and symptoms in many medical applications, Albright et al. [24] annotated disorders as a semantic type independent of signs or symptoms, and built a corpus that annotated entities and their assertions. In 2015, to enhance NLP research in the clinical domain, Elhadad et al. [31] released a corpus that annotated disorders with various attributes in SemEval-2015 Task 14. The attributes of the disorders are beneficial for extracting deeper patient information in the clinical texts.

2.2.2. Current status in Chinese clinical texts

Referring to the concept annotation guidelines in the 2010 i2b2/VA challenge, Xu et al. [26] labeled medical problems, treatments, and tests in Chinese discharge summaries and added two more entity types, namely medication and anatomy. Medication is separated from treatment for further analysis on the usage and effectiveness of medications, and anatomy can help to locate positions of symptoms or tests. Similar to Xu et al.'s corpus on entities, Lei et al. [32] developed an entity corpus of discharge summaries and admission notes from Peking Union Medical College Hospital. Their entity categories differ from the 2010 i2b2/VA concept guidelines in that treatments are divided into procedures and medications. Moreover, Xu et al. [36] annotated medical terms in the "history of present illness" section in clinical texts, and proposed an effective rule-based method. Differ from the above research on Chinese clinical texts, Jia et al. [35] manually marked up negated information on medical terms. To our knowledge, this is the first entity corpus with assertion information annotated in Chinese clinical texts.

Moreover, research into the clinical texts of traditional Chinese medicine has gradually been taken into account. Wang et al. [33] conducted research in recognizing symptoms based on the chief complaints, but the text types and entity categories in their corpus were relatively few. Li et al. [37] proposed a network-based correlation analysis method to detect herb-symptom associations and built a dataset of herb-symptom records that annotated correlations between symptoms and herbs. This study is meaningful for research on relation extraction from Chinese clinical texts.

However, for the clinical texts on a particular disease, the existing classification standards of medical entities are too rough, and some important information has not been distinguished effectively. In order to identify tumor-related information from Chinese operation notes, Wang et al. [34] manually annotated 12 entity types on operations, which revealed operation details and correlated strongly with patients' pathological status. This study provides a good reference for research on information extraction for specific diseases.

2.3. Shortcomings of research on Chinese clinical texts

Corpus construction on English clinical texts is a mature field, and its annotation scheme and evaluation method are of great significance for Chinese clinical texts. Considering the research status described above, research on Chinese clinical texts has three shortcomings: first, research on low-level tasks is quite limited, and this may cause performance improvement of higher-level tasks to encounter a bottleneck; second, as far as we know, only negated assertion and symptom-herb correlation have been annotated, while other types of assertions or relations have not been annotated systematically; and third, guideline tuning and annotator training are needed in corpus construction, but descriptions of previous research efforts have not described these processing procedures. Based upon the above three points, along with the fact that no clinical corpus written in Chinese has been released to the public, it is imperative to build a comprehensive corpus that follows a complete annotation scheme.

By referring to the existing research on English clinical texts, we constructed a comprehensive corpus of Chinese clinical texts. In our annotation method, some existing well-developed guidelines were used and adapted into Chinese clinical texts in the process of annotation guideline tuning. Next, annotator training and various measures were conducted to ensure the quality of this corpus. Furthermore, according to the annotations in this corpus, corresponding automatic system modules were developed.

3. Materials and methods

3.1. Types of clinical text

Discharge summaries and progress notes employed in this work were randomly selected from clinical texts of the Second Affiliated Hospital of Harbin Medical University (a general hospital in China, with a distribution of doctor number in different departments shown in Fig. 1), and all identifying information was removed manually to protect patient privacy. These two types of clinical text are semi-structured documents, and free text in the document is divided into several sections, as listed in Fig. 2.

3.2. Annotation guidelines

Due to the diversity of clinical texts, there is no existing annotation schema widely applicable in the clinical domain [24]. Owing to different language features between Chinese and English, annotation guidelines for CTB [46–48] were chosen to develop guidelines for low-level tasks on Chinese clinical texts; meanwhile, annotation guidelines in the 2010 i2b2/VA challenge [42–44] were consulted to develop guidelines for higher-level tasks. According to the characteristics of Chinese clinical texts, we developed several modified annotation guidelines [49–52] for four low-level and three higher-level tasks. Fig. 3 shows an example of the annota-

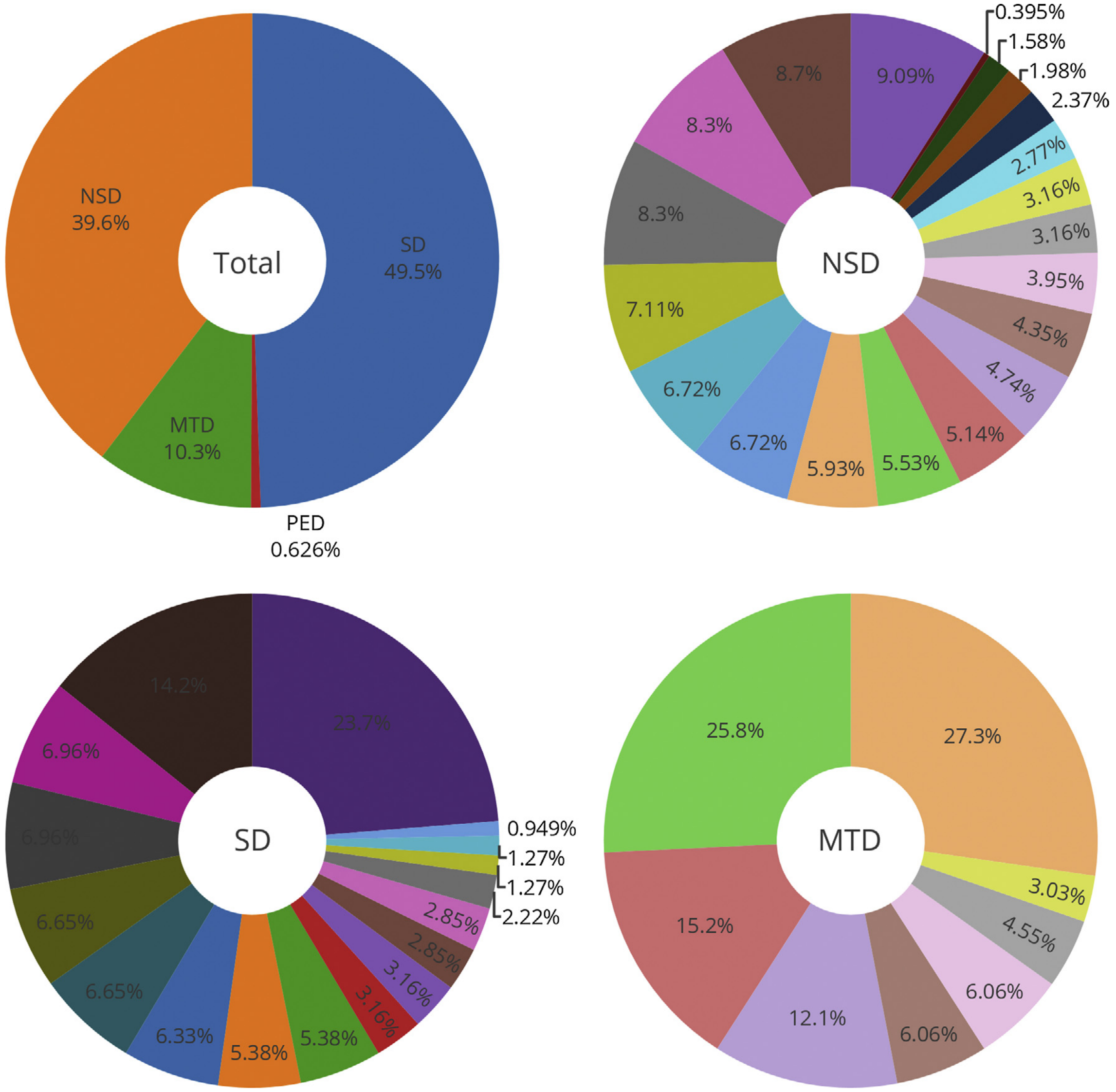


Fig. 1. Distribution of doctor number in different departments. Total = NSD + SD + MTD + PED; NSD, non-surgical departments; SD, surgical departments; MTD, medical technical departments; PED, physical examination department.

tions in a sentence from the case characteristics section of a progress note.

3.2.1. Guideline development for low-level tasks

3.2.1.1. Word segmentation. The segmentation guidelines for the Penn Chinese TreeBank (CTB) [46] cannot cover all the segmentation ambiguities in clinical texts, especially the segmentation of medical terms, abbreviations, and their combinations. In order to reduce these segmentation ambiguities, we summarized some adaptations as follows.

Nineteen specifications are listed in our word segmentation guidelines with examples from Chinese clinical texts, in which we added more detailed descriptions than CTB in some cases. For example, CTB only specifies 1 + 1 or 2 + 1 in the compound words of nouns, but segmentation of 1 + 2 ("脑实质 (brain parenchyma)" or "肌张力 (muscle tension)") or 3 + 1 ("高血压病 (hypertension)") are not specified, so we added the rule "to an N + N word whose length ranges from 2 to 4 and N1 modifies N2, if the length of N1 or N2 is 1, then treat N1 + N2 as one word".

Obviously, medical terms, if they are nouns or do not have the combinability attribute, are not normally split, such as "糖尿病 (diabetes)". Therefore, segmentation of non-nominal terms that have the combinability attribute is a major problem. Apart from the specifications mentioned before, an assistant word segmentation method was developed, as shown in Fig. 4a, and the following three word attributes were introduced into it:

1. *Combinability* [53], which means that a word can be separated into two sub-words and that each sub-word has its independent POS.
2. *Reducibility*, which indicates that, if a word is an abbreviation, then it can be reverted to its complete expanded form to clarify its description.
3. *Replaceability*, which denotes that one sub-word in a combined word can be replaced by another word with the same POS, and that the new combined word may still appear in clinical texts.

Fig. 4b illustrates the process of the segmentation of "抗凝 (anti-coagulation)". First, split "抗凝 (anti-coagulation)" into "抗$_{[anti]}$ 凝$_{[coagulation]}$". Then, revert the short form "凝$_{[coagulation]}$"

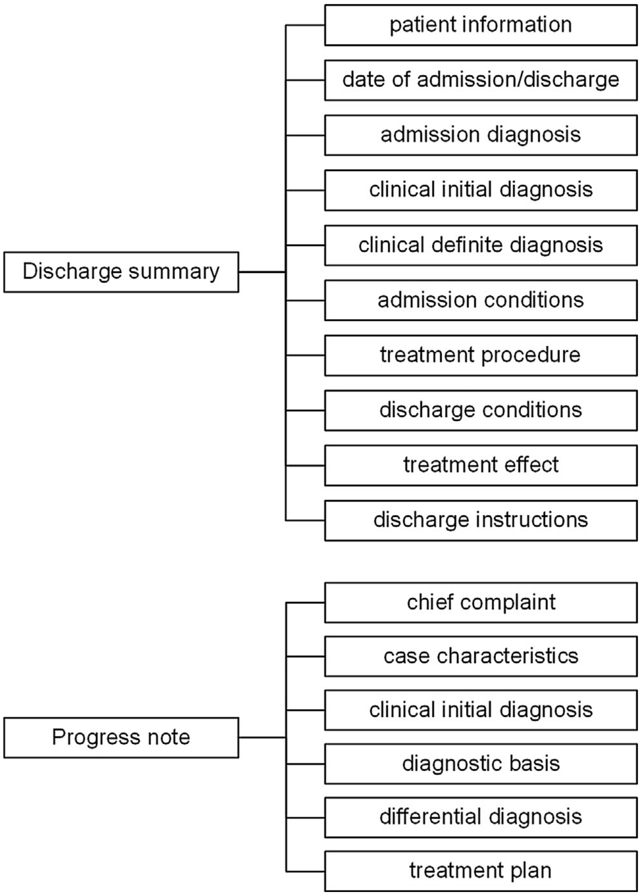


Fig. 2. Semi-structured sections in Chinese discharge summaries and progress notes.

to its corresponding word “凝血 (coagulation)”. We found that “凝血 (coagulation)” could be replaced with “发炎 (inflammation)” and that these two words have the same POS, and the combination of “抗$_{\text{[anti]}}$” and “炎$_{\text{[inflammation]}}$” (the short form of “发炎 (inflammation)”) is “抗炎 (anti-inflammation)”, which appears in clinical texts. Therefore, the word “抗凝 (anti-coagulation)” can be split into two sub-words.

To improve the usability of our word segmentation guidelines, examples extracted from clinical texts are listed, so do the following annotation guidelines.

3.2.1.2. POS tagging and parsing. To build POS tagging guidelines for Chinese clinical texts, the POS tag set and confusing parts of speech in the POS tagging guidelines for CTB were adapted using instances from clinical texts. Additionally, some adaptions of special cases in clinical texts were made, and three main problems and their solutions are described as follows:

1. Some usage of specific symbols that do not exist in CTB appear as abbreviations of certain words in clinical texts, such as“+” in “肌力4+级 (myodynamia level is 4+)” means “stronger” and “−” in “3–4次/分 (3–4 times per minute)” means “to”. Moreover, we tagged the POS of the specific symbol based on its meaning in context; hence, “+” and “−” in these examples should be tagged as a VA (predicative adjective) and CC (coordinating conjunction), respectively.
2. A verb-complement phrase will be utilized as an object to describe a patient’s symptom in clinical texts, but this usage does not appear in CTB, so POS tags of words in a verb-complement phrase come with some ambiguity. For instance, “视物 模糊 (blurred vision)” in “伴有 视物 模糊 (with blurred vision)” is a symptom and can be seen as a noun phrase, so “视物$_{\text{[see things]}}$” can be tagged as an NN (common nouns); but from the perspective of phrase structure, “视物$_{\text{[see things]}}$ 模糊$_{\text{[blurred]}}$” is a verb-complement phrase, and thus “视物$_{\text{[see things]}}$” should be tagged as a VV (other verbs). To solve this kind of ambiguity, POS tags are achieved according to the POS of the word itself, so “视物$_{\text{[see things]}}$” is tagged as a VV.
3. Annotation ambiguities caused by sentences with missing elements. For example, the POS tag of “左侧$_{\text{[left side]}}$” in “左侧肢体麻木 (numbness in the left limbs)” and “右肺呼吸音清左侧弱 (right lung breath sounds clear and the left is weak)” are different because the latter sentence omitted some words. In CTB, ambiguity between “NN” and “JJ” can usually be disambiguated by judging whether the word is the head of a noun phrase; however, neither occurrence of “左侧$_{\text{[left side]}}$” in the above two examples is the head of a noun phrase, so we need to complement omitted elements in the sentence. The former example has a normal grammatical structure in which “左侧$_{\text{[left side]}}$” modifies “肢体$_{\text{[limbs]}}$” and should be tagged as a JJ; but “左侧$_{\text{[left side]}}$” in the latter example is short for “左侧肺呼吸音 (left lung breath sounds)”, which is a noun phrase and should be tagged as an NN.

Furthermore, we simplified the bracketing guidelines for CTB [48] and adapted these annotation specifications to the clinical domain, providing clear guidance in annotating for the parsing (shallow parsing and full parsing) task in Chinese clinical texts.

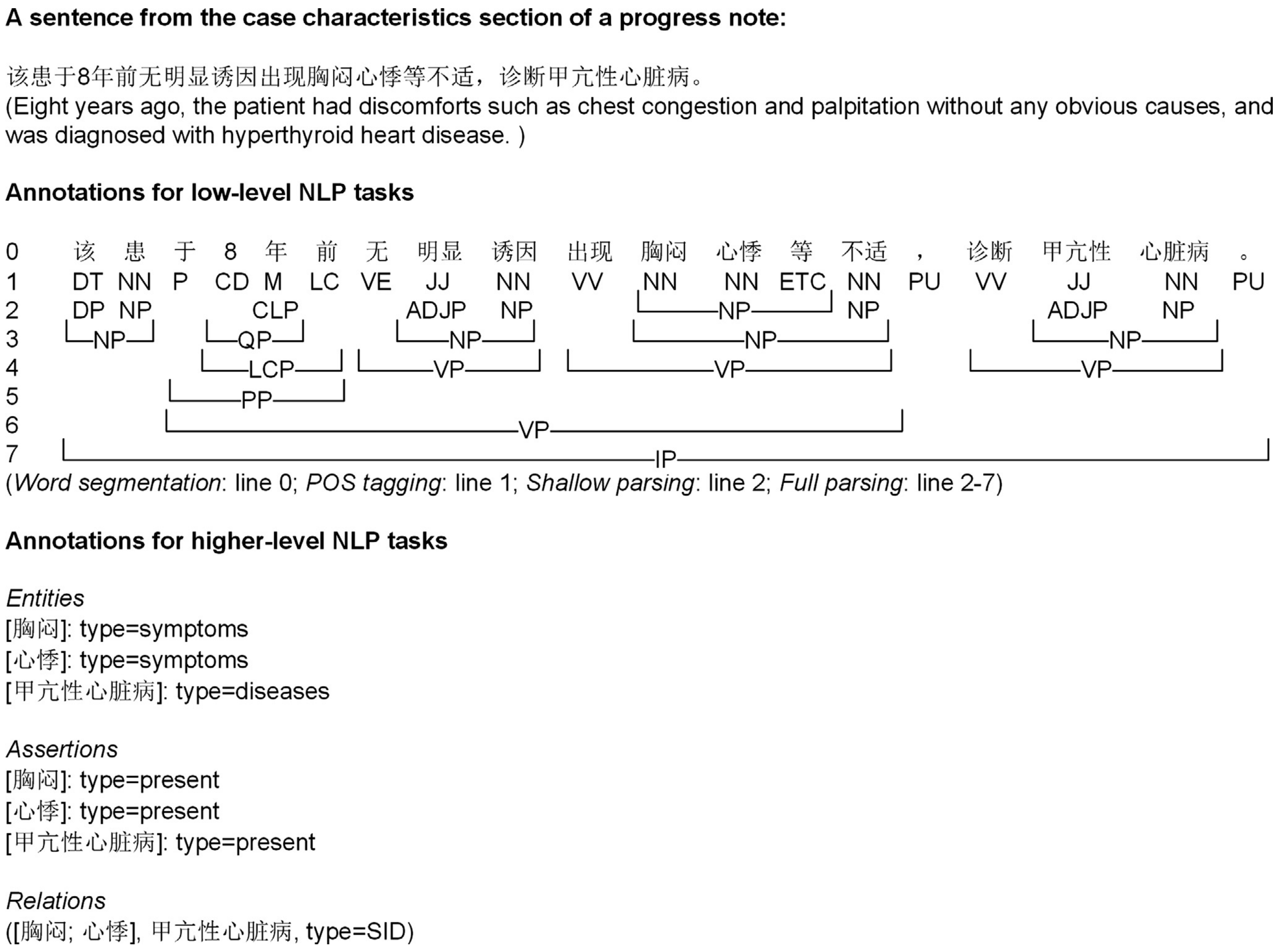
3.2.2. Guideline development for higher-level tasks

3.2.2.1. Entities and assertions. Concept annotation guidelines in the 2010 i2b2/VA challenge [42] include three categories of concept: medical problems, tests, and treatments. However, as Uzuner et al. [45] pointed out, patients’ medical problems can be represented as diseases and symptoms, and these two kinds of concept have separate UMLS semantic types; hence, we treated diseases and symptoms as two types of medical entity in our annotation guidelines, as shown in Table 2A.

In the 2010 i2b2/VA challenge, only assertions of medical problems were annotated, and each medical problem was assigned one of the six assertion types [43]. In our work, we did not find any *hypothetical* entity in Chinese clinical texts, but observed a relatively frequent assertion in the default category *present*, so we deleted the *hypothetical* assertion type and separated the additional kind of assertion *occasional* from *present*. We assigned six assertion types to diseases and symptoms in Chinese clinical texts. Furthermore, because the statuses of treatments administered in patients are important references for clinical diagnoses, we annotated three types of assertion in treatments: *present*, *absent*, and *historical*. Table 2B lists the assertions of medical entities with their examples.

3.2.2.2. Relations. Based on relations in the 2010 i2b2/VA challenge [44], we extended the relation types into five main categories and 15 subcategories in Chinese clinical texts, as shown in Table 3. All these relationships are bounded by sentences, and entity assertions are not considered when labeling relationships.

In Chinese clinical texts, entities of the same type usually appear one after the other in a sentence, and there are commonly concurrent relationships between these entities; for example, some treatments are administered for a disease or a disease causes some symptoms. Additionally, these entities may have the same type of relationship with an entity of a different type in the sentence, but one-to-one relationships between one of the former entities and the latter entity may not be clearly indicated, causing difficulty in the annotation of one-to-one relationships. To avoid annotating fuzzy one-to-one relationships, we referred to the def-



A sentence from the case characteristics section of a progress note:

该患于8年前无明显诱因出现胸闷心悸等不适，诊断甲亢性心脏病。
(Eight years ago, the patient had discomforts such as chest congestion and palpitation without any obvious causes, and was diagnosed with hyperthyroid heart disease.)

Annotations for low-level NLP tasks

0 该 患 于 8 年 前 无 明显 诱因 出现 胸闷 心悸 等 不适 ， 诊断 甲亢性 心脏病 。
1 DT NN P CD M LC VE JJ NN VV NN NN ETC NN PU VV JJ NN PU
2 DP NP CLP ADJP NP NP NP ADJP NP
3 NP QP NP NP NP
4 LCP VP VP VP
5 PP
6 VP
7 IP

(*Word segmentation*: line 0; *POS tagging*: line 1; *Shallow parsing*: line 2; *Full parsing*: line 2-7)

Annotations for higher-level NLP tasks

Entities
[胸闷]: type=symptoms
[心悸]: type=symptoms
[甲亢性心脏病]: type=diseases

Assertions
[胸闷]: type=present
[心悸]: type=present
[甲亢性心脏病]: type=present

Relations
([胸闷; 心悸], 甲亢性心脏病, type=SID)

Fig. 3. An example of the annotations in a sentence from a progress note. NLP, natural language processing; DT, determiner; NN, common nouns; P, prepositions; CD, cardinal numbers; M, measure word; LC, localizer; VE, you3 as the main verb; JJ, noun-modifier other than nouns; VV, other verbs; ETC, tags for deng3 and deng3deng3 in coordination phrases; PU, punctuation; DP, determiner phrase; NP, noun phrase; CLP, classifier phrase; ADJP, adjective phrase; QP, quantifier phrase; LCP, phrase formed by "phrase + LC"; VP, verb phrase; PP, preposition phrase; IP, simple clause; POS, part-of-speech; SID, symptom indicates disease.

inition of a narrative container used in temporal relations [54], and proposed the concept of an "entity group" to assist in the relation annotation task in Chinese clinical texts.

Entities of the same type in a sentence are combined into an entity group if they satisfy the following two conditions: (1) simultaneity, which means that these entities appear at the same time during a medical activity of the patient, indicating a concurrent relationship between entities; (2) these entities have the same type of relationship with an entity of a different type in the sentence.

According to the definition of an entity group, one-to-one relationships can be developed into a relationship between an entity and an entity group, or a relationship between an entity group and another entity group. In the example shown in Fig. 3, the patient had symptoms of "胸闷 (chest congestion)" and "心悸 (palpitation)", and was diagnosed with "甲亢性心脏病 (hyperthyroid heart disease)", so a relationship between entity group "[胸闷; 心悸]" and entity "甲亢性心脏病" was annotated.

The introduction of entity groups may weaken one-to-one relationships between entities, but solves the problem of fuzzy relationships. Besides, the definition of an entity group can also be explained by doctors' habits of clinical diagnosis and treatment: when a doctor makes a diagnosis based on the patient's current symptoms, the diagnosis is not based on one symptom but on a comprehensive judgment of a group of symptoms, and several tests or treatments are applied cooperatively to the patient.

3.3. Annotation method

Referring to the annotation methods in English clinical texts [24,25], annotation guideline development and corpus construction for each NLP task were executed in three major stages (as shown in Fig. 5):

1. Building the draft guidelines: Annotation guidelines for CTB [46–48] and annotation guidelines in the 2010 i2b2/VA challenge [42–44] were chosen as the basis for developing guidelines for NLP tasks on Chinese clinical texts. With the help of two resident physicians (P1 and P2), four annotators with backgrounds in computational linguistics (CL1 and CL2 for low-level tasks, CL5 and CL6 for higher-level tasks) summarized the characteristics of Chinese clinical texts and drafted annotation guidelines adapted for them. In these guidelines, a large number of annotated examples are listed, and annotation ambiguities are analyzed in detail to make the annotation work easier.
2. Training the annotators and updating the guidelines: An iterative method was proposed to train the annotators and update the guidelines. In each round, a certain number of clinical documents were randomly sampled from the unannotated dataset. To accelerate the annotation progress as well as to ensure annotation quality, different strategies were implemented during the double-annotation period of different tasks: (1) automated

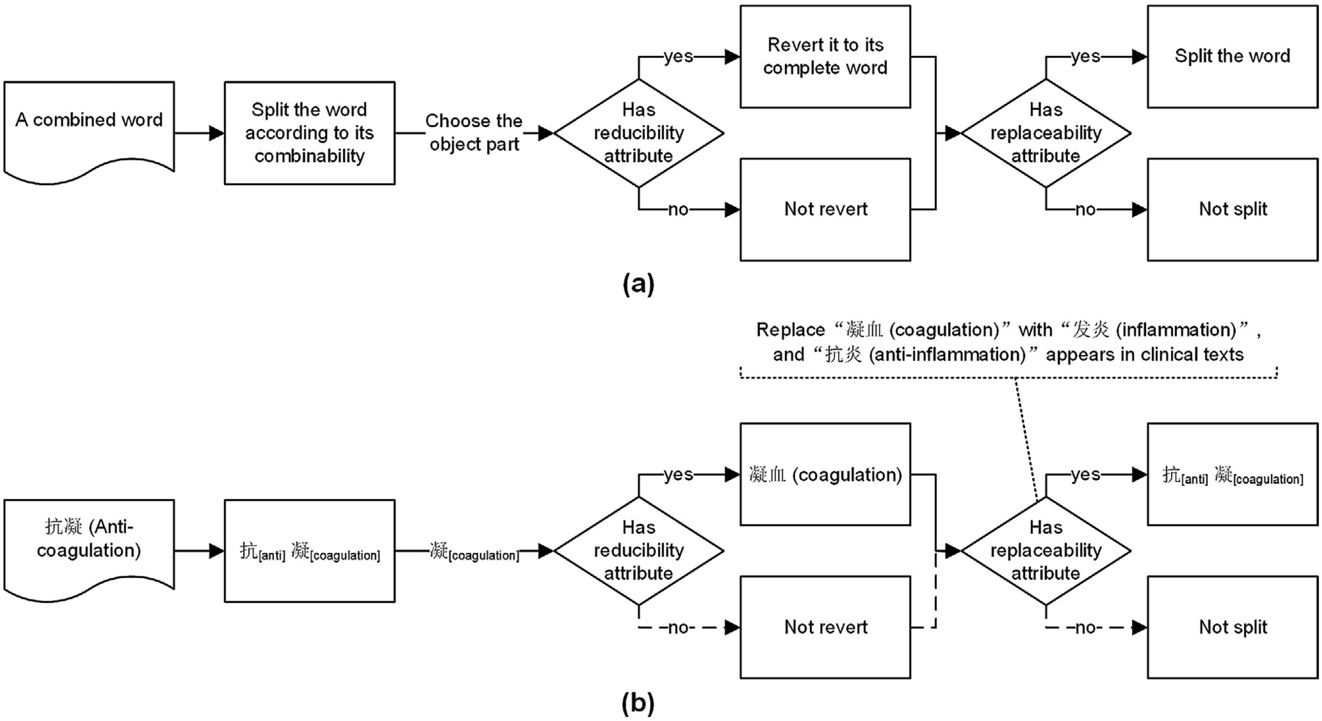

A combined word
Split the word according to its combinability
Choose the object part
Has reducibility attribute
yes
no
Revert it to its complete word
Not revert
Has replaceability attribute
yes
no
Split the word
Not split
(a)
Replace "凝血 (coagulation)" with "发炎 (inflammation)", and "抗炎 (anti-inflammation)" appears in clinical texts
抗凝 (Anti-coagulation)
抗[anti] 凝[coagulation]
凝[coagulation]
Has reducibility attribute
yes
no
凝血 (coagulation)
Not revert
Has replaceability attribute
yes
no
抗[anti] 凝[coagulation]
Not split
(b)

Fig. 4. A word segmentation method for non-nominal terms that have the combinability attribute.

tools trained in the general domain [55–57] were applied in the pre-tagging of low-level annotations, and four annotators with backgrounds in computational linguistics were divided into two groups (CL1 and CL3 in annotator group 1, CL2 and CL4 in annotator group 2) to conduct double verification and correction of the automatically added annotations (the annotators in each group accomplish the work collaboratively); annotation disagreements were then adjusted by a physician (P1); (2) since annotations of entities and assertions require professional medical knowledge, we had two physicians (P1 in annotator group 3, P2 in annotator group 4) annotate documents in parallel from the beginning; (3) in the relation annotation task, the documents were double-annotated by two annotator groups (CL5 and CL7 in annotator group 5, CL6 and CL8 in annotator group 6), and a physician (P2) was also assigned to resolve the annotation differences. IAA was then calculated to measure the quality of annotator training, and inconsistent cases were discussed to update the annotation guidelines.

3. Corpus construction: The iterative process in stage 2 continued until IAA was consistently high in the latest three iterations, showing that annotators reached an agreement on annotation guidelines. After the iterative annotator training process, two annotator groups in each task labeled different datasets separately to reduce the consumption of time and money. During this period, three measures were taken to ensure annotation quality: (1) duplicate documents were assigned to two annotator groups for the IAA evaluation of stage 3; (2) annotators recorded uncertain annotations, whose final results were achieved after discussion; (3) sampling inspection was carried out and at least one third of the annotations were checked, and the conflicts with the latest guidelines were then modified after further discussion.

3.4. Inter-annotator agreement

To evaluate the annotation quality of our corpus, IAA was calculated using the F_1 measure. The annotations of one annotator group were seen as the gold standard, and were used to calculate the precision, recall, and F_1 measure of the second annotator group, as described in the following equations [58]:

$$Precision = \mathrm{AgreedNumber}(AG_1, AG_2)/\mathrm{AnnotationNumber}(AG_2), \tag{1}$$

$$Recall = \mathrm{AgreedNumber}(AG_1, AG_2)/\mathrm{AnnotationNumber}(AG_1), \tag{2}$$

$$F = (1 + \beta^2) * Precision * Recall/(\beta^2 * Precision + Recall), \tag{3}$$

where $\mathrm{AgreedNumber}(x, y)$ means the number of the consistent annotations between x and y, $\mathrm{AnnotationNumber}(x)$ means the annotation number of x, AG_i means annotator group i, and $\beta = 1$ in our work.

For parsing annotations, Evalb [59] was utilized to calculate the IAA of the parsing trees. Since entities and their assertions were annotated simultaneously to accelerate annotation progress, we merged these two IAA evaluations into one, in which the agreement should satisfy the condition that the extent, type, and assertion of an entity are consistent. Considering the existence of entity groups in entity relations, two types of IAA for relations were computed: the first measured the IAA of relation annotations that preserve entity groups in the relationship; the second separated entity groups into entities and then calculated the IAA of the one-to-one relationships.

4. Results

4.1. Annotation consistency

As shown in Table 4, the IAA values of these annotation tasks show an increasing trend in the latest three annotator training iterations, indicating that an annotator's mastery of the annotation guidelines improves continually. Furthermore, on account of the fact that the IAA values of relation annotations in the training stage are relatively lower, we added duplicate documents in the corpus

Table 2
Entities and their assertions annotated in Chinese clinical texts.

Part A

Entity type	Example
Diseases	行支气管镜检查示: 小细胞肺癌 (Bronchoscopy showed: small cell lung cancer)
Symptoms	疼痛时伴右下肢活动受限 (pain accompanied by the right lower extremity activity limitation)
Tests	行支气管镜检查示: 小细胞肺癌 (Bronchoscopy showed: small cell lung cancer)
Treatments	注射胰岛素控制血糖 (injection of insulin to control blood glucose)

Part B

Entity type	Assertion type	Description	Example
Diseases symptoms	Present	Disease or symptom exists in the patient	头CT示: 双侧多发腔梗 (head CT ***showed***: bilateral multiple lacunar infarct)
	Absent	Disease or symptom does not exist in the patient	双下肢无浮肿 (***no*** edema in both lower limbs)
	Possible	Disease or symptom may exist in the patient	右肺下叶考虑创伤性湿肺 (right lung lower lobe ***consider*** traumatic wet lung)
	Conditional	Disease or symptom occurs in the patient under certain conditions	. . .胸闷、气短, 常于饮酒后出现 (. . .chest tightness, shortness of breath, commonly occurs ***after drinking***)
	Not associated with the patient	Disease or symptom exists in the patient's relatives	患者父母均患有糖尿病 (***parents*** of the patient suffer from diabetes)
	Occasional	Disease or symptom exists in the patient occasionally	时有胸闷气短 (there are chest tightness and shortness of breath ***sometimes***)
Treatments	Present	The patient is experiencing or will experience the treatment	右侧胸部见引流管 (drainage tube in the right side of the chest)
	Absent	The patient does not experience the treatment	分娩前无镇静 剂 (***no*** sedative before childbirth)
	Historical	The patient experienced the treatment in the past	***18***年前剖宫产手术 (cesarean section ***18 years ago***)

In the examples, entities are underlined and indicators of the assertions are highlighted in bold and italics.

Table 3
Relations between medical entities annotated in Chinese clinical texts.

Entity pair	Relation type	Description	Example
Treatments, diseases	TrID	Treatment improves disease	. . .诊断[贫血]$_{D}$,给予[输血]$_{Tr}$后好转 (. . .was diagnosed with [anemia]$_{D}$, and improved after giving [blood transfusion]$_{Tr}$)
	TrWD	Treatment worsen disease	[高血压病]$_{D}$口服[替米沙坦]$_{Tr}$控制,但血压控制不佳 (oral [Telmisartan]$_{Tr}$ to control [hypertensive disease]$_{D}$, but poorly controlled blood pressure)
	TrCD	Treatment causes disease	[电除颤]$_{Tr}$后:[III度房室传导阻滞]$_{D}$ (after [electric defibrillation]$_{Tr}$: [three degree atrioventricular block]$_{D}$)
	TrAD	Treatment is administered for disease	. . .被诊断为[结肠癌]$_{D}$,行[右半结肠癌根治术]$_{Tr}$ (. . .was diagnosed with [colon cancer]$_{D}$, and [right hemi-colonic carcinoma radical operation]$_{Tr}$ was administered)
Treatments, symptoms	TrIS	Treatment improves symptom	. . .服用[钙剂]$_{Tr}$等治疗后,[后背部疼痛]$_{S}$显著缓解 (. . .after taking [calcium]$_{Tr}$ and other treatments, [back pain]$_{S}$ was significantly alleviated)
	TrWS	Treatment worsen symptom	. . .发现[血糖升高]$_{S}$,口服[拜糖平]$_{Tr}$及[二甲双胍]$_{Tr}$8天,血糖控制欠佳 (. . .found that [blood glucose rose]$_{S}$, oral [acarbose]$_{Tr}$ and [metformin]$_{Tr}$ eight days, poorly controlled blood glucose)
	TrCS	Treatment causes symptom	. . .应用长效干扰素[派罗欣]$_{Tr}$后出现[体力下降]$_{S}$,[周身不适]$_{S}$ (. . .after application of [Pegasys]$_{Tr}$, appeared [physical decline]$_{S}$ and [general malaise]$_{S}$)
	TrAS	Treatment is administered for symptom	. . .于医院查[肌酐增高]$_{S}$,给与患者[改善肾血流]$_{Tr}$等相关治疗 (. . .checked out [creatinine increased]$_{S}$ in the hospital, and the patient was given [improvement of renal blood flow]$_{Tr}$ and other related treatments)
	TrNAS	Treatment is not administered because of symptom	. . .发现[转氨酶高]$_{S}$,停用[达那唑]$_{Tr}$. . . (. . .found [high transaminase]$_{S}$, stopped taking [danazol]$_{Tr}$. . .)
Tests, diseases	TeRD	Test reveals disease	[头CT]$_{Te}$示:[双侧多发腔梗]$_{D}$ ([head CT]$_{Te}$ showed: [bilateral multiple lacunar infarct]$_{D}$)
	TeCD	Test conducted to investigate disease	患者病情尚不除外[脑炎]$_{D}$,建议[腰穿]$_{Te}$. . . ([encephalitis]$_{D}$ was not excepted in the patient's conditions, suggest [lumbar puncture check]$_{Te}$. . .)
Tests, symptoms	TeRS	Test reveals symptom	. . .[头CT检查]$_{Te}$显示[颅内多发低密度病灶]$_{S}$ (. . .[head CT examination]$_{Te}$ showed [intracranial multiple low density lesions]$_{S}$)
	TeAS	Test is administered because of symptom	. . .出现[发热]$_{S}$,[鼻出血]$_{S}$,当地查[血常规]$_{Te}$. . . (. . .appeared [fever]$_{S}$, [epistaxis]$_{S}$, and checked [blood routine]$_{Te}$ in local. . .)
Diseases, symptoms	DCS	Disease causes symptom	3年前[脑梗死]$_{D}$遗留[说话含糊不清]$_{S}$,[走路拖沓]$_{S}$. . . ([cerebral infarction]$_{D}$ three years ago, now presenting with [muffled speech]$_{S}$, [walk procrastination]$_{S}$. . .)
	SID	Symptom indicates disease	. . .出现[胸闷]$_{S}$[心悸]$_{S}$等不适,诊断[甲亢性心脏病]$_{D}$ (. . .had discomforts such as [chest congestion]$_{S}$ and [palpitation]$_{S}$, and was diagnosed with [hyperthyroid heart disease]$_{D}$)

In the examples, entities are in brackets followed by the abbreviation of the entity type. D, diseases; S, symptoms; Te, tests; Tr, treatments.

construction stage of higher-level tasks. The last column of Table 4 shows that the IAA of these documents remained at a relatively high level, indicating that annotators have the ability to accomplish these annotation tasks with acceptable consistencies.

4.2. Data analysis of annotations for low-level tasks

Annotations for low-level tasks cover 72 Chinese discharge summaries and 66 progress notes, including 2612 full parsing trees.

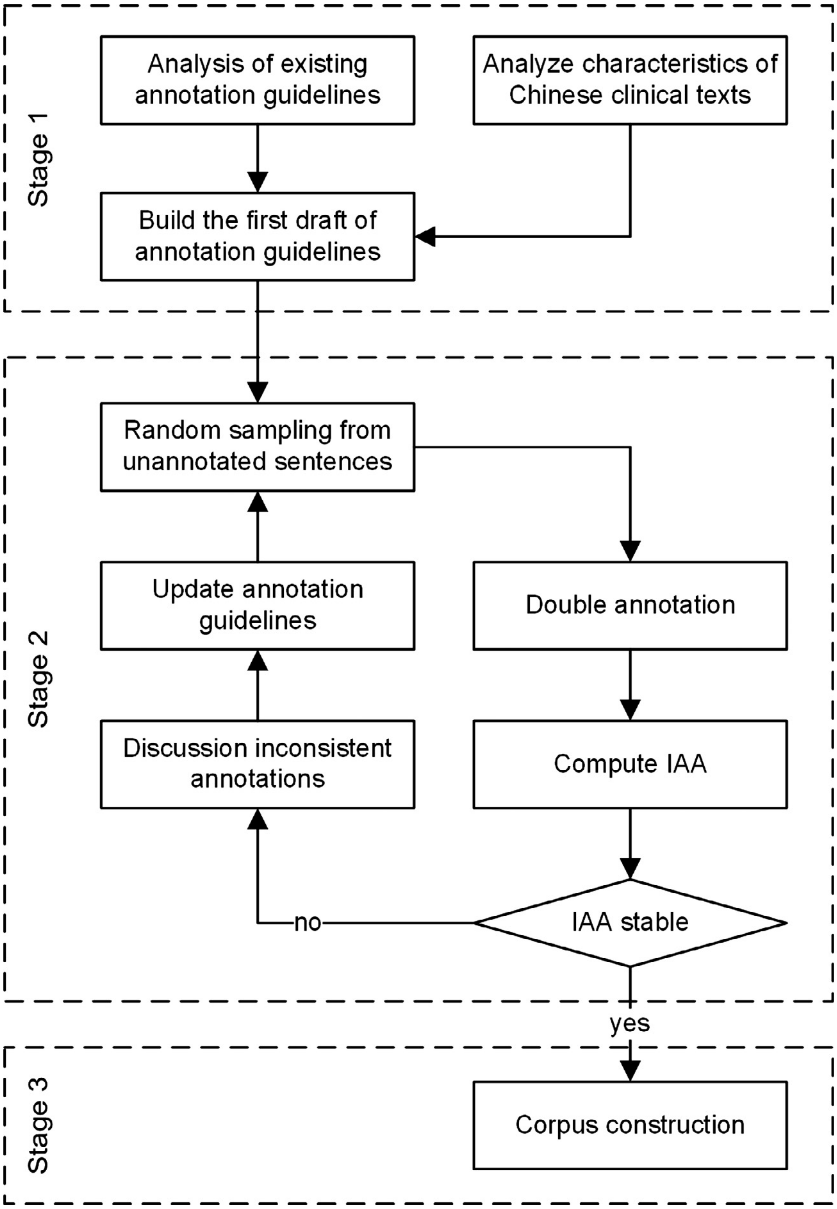

Stage 1
Analysis of existing annotation guidelines
Analyze characteristics of Chinese clinical texts
Build the first draft of annotation guidelines
Stage 2
Random sampling from unannotated sentences
Update annotation guidelines
Double annotation
Discussion inconsistent annotations
Compute IAA
IAA stable
no
yes
Stage 3
Corpus construction

Fig. 5. Iterative annotation method for guideline development and corpus construction. IAA, inter-annotator agreement.

Table 4
Inter-annotator agreement in the latest three annotator training iterations and corpus construction stage (F_1 measure).

	IAA			
	Training [-3]	Training [-2]	Training [-1]	Corpus construction
Word segmentation	0.965	0.979	0.983	–
POS tagging	0.893	0.952	0.956	–
Shallow parsing	0.956	0.969	0.970	–
Full parsing	0.805	0.840	0.865	–
Entity (span, type, assertion)	0.848	0.920	0.927	0.922
Relation (entity group preserved)	0.765	0.781	0.843	0.772
Relation (one-to-one)	0.742	0.774	0.805	0.755

"–" means not evaluated. IAA, inter-annotator agreement; POS, part-of-speech.

Tables 5 and 6 list POS and syntactic tag counts in our annotated clinical texts. There are 47,426 tokens in this corpus, and its average sentence length is 18.16 tokens, which is much shorter than the 27.09 in CTB 5.0. Within clinical texts, the average sentence length of discharge summaries is shorter than that of progress notes (14.13 vs. 22.50) because sentences in some sections of discharge summaries are quite short, especially in cases where only one token exists in the "treatment effect" section. As statistical results of the special cases described in Section 3.2.1.2 show, in our corpus, case 1 and case 2 appeared 137 and 54 times, respectively, and the proportion of grammar rules with omitted elements is about twice that in CTB. This higher proportion is closely related to the concise-

Table 5
POS tag counts in our annotated clinical texts.

Annotation type	Description	Counts
NN	Common nouns	14,765
PU	Punctuation	10,755
VV	Other verbs	5889
CD	Cardinal numbers	3484
VA	Predicative adjective	2788
JJ	Noun-modifier other than nouns	2088
AD	Adverbs	1759
M	Measure word (including classifiers)	1736
VE	You3 as the main verb	1160
P	Prepositions (excluding ba3 and bei4)	627
LC	Localizer	595
NT	Temporal nouns	584
CC	Coordinating conj	470
DT	Determiner	251
OD	ordinal numbers	232
ETC	Tags for deng3 and deng3deng3 in coordination phrases	74
NR	Proper nouns	53
VC	Copula shi4	44
PN	Pronouns	26
DEG	Associative de5	16
MSP	Some particles	8
CS	Subordinating conj	7
DEC	De5 for relative-clause etc.	6
SB	Bei4 in short bei-construction	5
BA	Ba3 in ba-const	1
LB	Bei4 in long bei-construction	1
AS	Aspect marker	1
FW	Foreign words	0
SP	Sentence-final particle	0
DER	De5 in V-de const. and V-de-R	0
DEV	De5 as the head of DVP	0
IJ	Interjection	0
ON	Onomatopoeia	0

POS, part-of-speech.

Table 6
Syntactic tag counts in our annotated clinical texts.

Annotation type	Description	Counts
NP	Noun phrase	17240
VP	Verb phrase	14686
IP	Simple clause	9621
QP	Quantifier phrase	2699
ADJP	Adjective phrase	2117
ADVP	Adverbial phrase	1754
CLP	Classifier phrase	1736
LST	List marker	1104
PP	Preposition phrase	662
LCP	Phrase formed by “phrase + LC”	598
FRAG	Fragment	341
DP	Determiner phrase	251
VCD	Coordinated verb compound	164
PRN	Parenthetical	106
VSB	Verb compounds formed by a modifier + a head	97
VRD	Verb resultative compound	35
UCP	Unidentical coordination phrase	27
DNP	Phrase formed by “phrase + DEG”	23
CP	Clause headed by C (complementizer)	6
VPT	Potential form V-de-R or V-bu-R	1
VNV	Verb compounds formed by A-not-A or A-one-A	1
VCP	Verb compounds formed by VV + VC	1
DVP	Phrase formed by “phrase + DEV”	0

LC, localizer; DEG, associative de5; VV, other verbs; VC, copula shi4; DEV, de5 as the head of DVP.

ness of physicians’ idioms. Moreover, Figs. 6 and 7 give a detailed comparison between tag distributions in Chinese clinical texts and CTB; items with an asterisk (*) symbol indicate statistical significance ($p < 0.05$) of the difference between the two corpora.

Compared with CTB, the POS tag distribution in clinical texts is relatively concentrated, and some tags are rare, such as NR (proper nouns), VC (copula shi4), PN (pronouns), DEG (associative de5), and DEC (de5 for relative-clause, etc.). The low percentage of NR in clinical texts is due to the de-identification of patients. Furthermore, the 22.68% of PU (punctuation) in clinical texts is much higher than the 15.29% in CTB because phrase structures, which are separated by punctuations, appear frequently in clinical texts to describe patients’ conditions. Moreover, some of the test results in clinical texts are described in the form of a numerical value, resulting in the percentage of CD (cardinal numbers) much higher than that in CTB.

As shown in Fig. 7, syntactic tag distribution in clinical texts is quite different from that of CTB, and this is closely related to the sublanguage properties of clinical texts. Some syntactic tags are rare in clinical texts, such as DNP (phrase formed by “phrase + DEG”) and CP (clause headed by complementizer). Moreover, the low proportion of DNP can be attributed to the same low percentage of DEG in POS tags. Furthermore, some sections in clinical texts, such as case characteristics and treatment plans, are detailed in the form of a list. For this reason, the 2.07% of LST (list marker) in clinical texts is understandably higher than the 0.03% in CTB.

4.3. Data analysis of annotations for higher-level tasks

Annotations for higher-level tasks contain 500 discharge summaries and 492 progress notes, including 39,511 entities and 7693 one-to-one relations. Tables 7 and 8 list entity and relation type counts in this corpus. Compared with discharge summaries, entities and relations contained in progress notes occur in larger quantities, accounting for three fifths and four fifths of the total numbers, respectively. Figs. 8 and 9 show entity and relation type distributions in these discharge summaries and progress notes, items with an asterisk (*) symbol indicate statistical significance ($p < 0.05$) of the difference between the two text types.

Discharge summaries and progress notes have similar distributions of the four entity types, as shown in Fig. 8. Symptoms account for nearly half of the total entities in discharge summaries and progress notes, respectively, and almost three fifths of these symptoms are *absent*, which can be used by physicians to distinguish patients’ conditions. In addition to these approximate distributions, the proportions of some assertion types in discharge summaries and progress notes show some differences. In discharge summaries, admission diagnosis results in more *possible* diseases, while clinical definite diagnosis leads to more *present* diseases; however, case characteristics describe the patient’s medical history, leading to many more *absent* diseases and *historical* treatments in progress notes.

In Fig. 9, relation type distributions in discharge summaries and progress notes are quite different for some relation types, especially disease-symptom relations, and this is closely related to the content emphasis of different clinical text types. In progress notes, present illness history is presented in the section of case characteristics, including patients’ conditions, tests, and relevant diagnoses, so the proportion of TeAS (test is administered because of symptom) and SID (symptom indicates disease) are much higher than those in discharge summaries.

4.4. System development

To verify the usefulness of our annotated corpus, we developed a Chinese clinical text processing and information extraction system (CCTPIES) that consisted of a word segmenter, POS tagger, shallow parser, full parser, named entity recognizer, and relation extractor [60], and the performance of these modules was evalu-

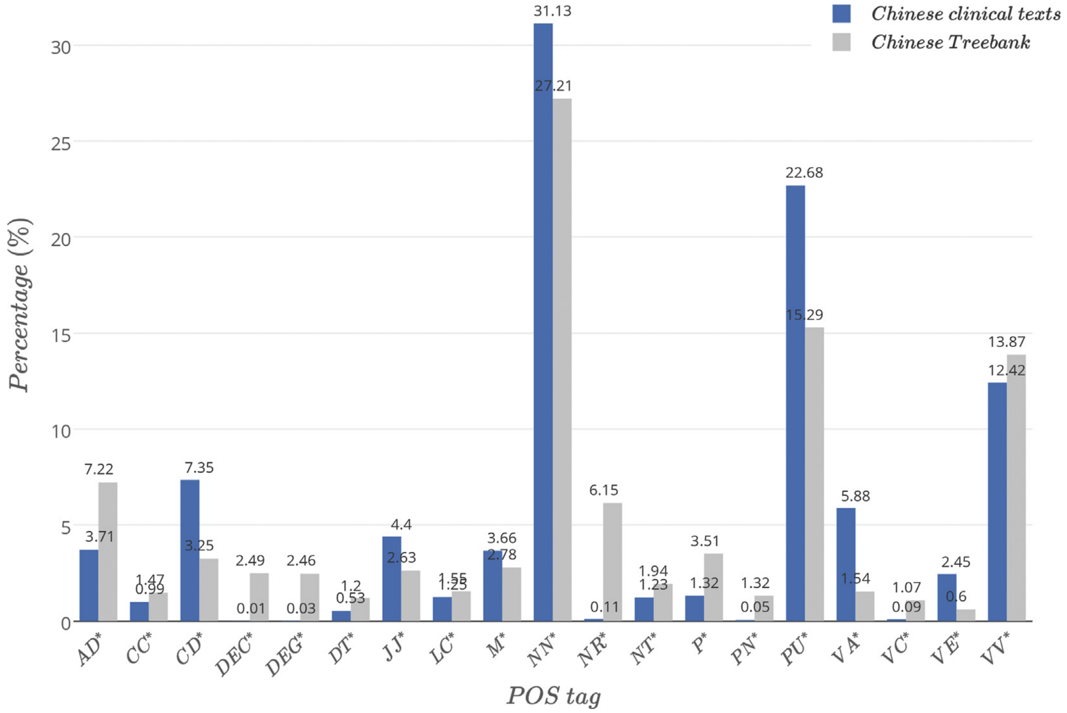


Fig. 6. POS tag distribution in Chinese clinical texts and CTB 5.0. The tags, whose percentages in clinical texts and CTB are both below 1%, are not listed in this figure. CTB, Chinese Treebank; AD, adverbs; CC, coordinating conj; CD, cardinal numbers; DEC, de5 for relative-clause, etc.; DEG, associative de5; DT, determiner; JJ, noun-modifier other than nouns; LC, localizer; M, measure word; NN, common nouns; NR, proper nouns; NT, temporal nouns; P, prepositions; PN, pronouns; PU, punctuation; VA, predicative adjective; VC, copula shi4; VE, you3 as the main verb; VV, other verbs; POS, part-of-speech. * indicates significant difference with $p < 0.05$.

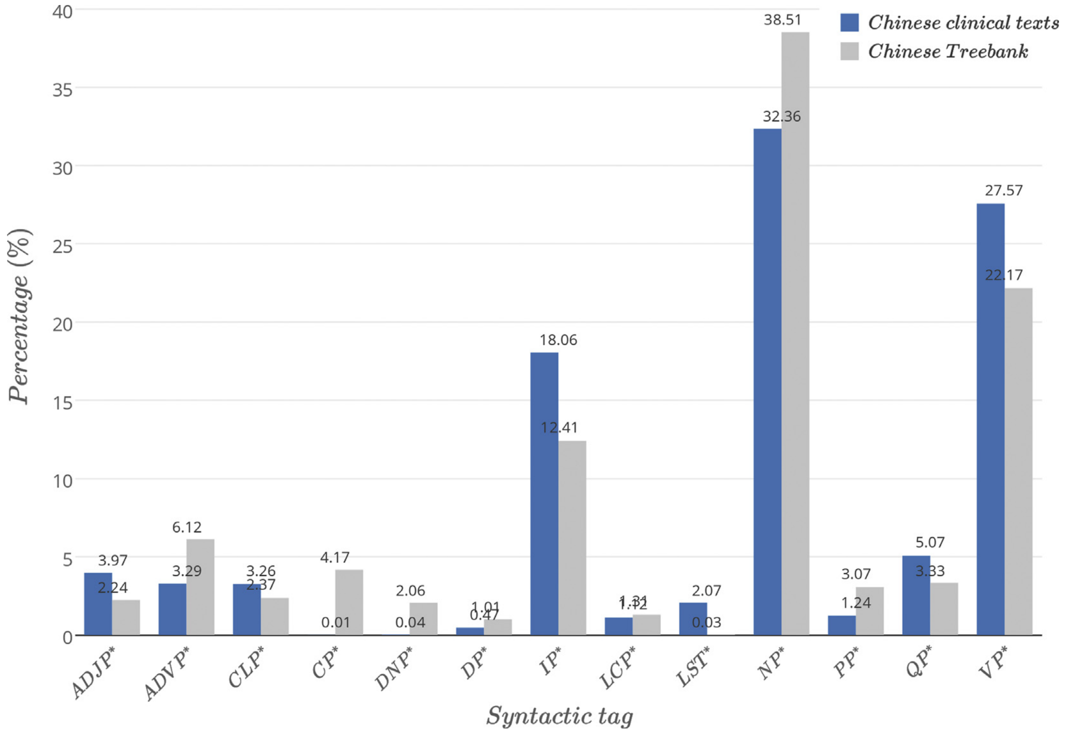


Fig. 7. Syntactic tag distribution in Chinese clinical texts and CTB 5.0. The tags, whose percentages in clinical texts and CTB are both below 1%, are not listed in this figure. CTB, Chinese Treebank; ADJP, adjective phrase; ADVP, adverbial phrase; CLP, classifier phrase; CP, clause headed by complementizer; DNP, phrase formed by "phrase + DEG"; DP, determiner phrase; IP, simple clause; LCP, phrase formed by "phrase + LC"; LST, list marker; NP, noun phrase; PP, preposition phrase; QP, quantifier phrase; VP, verb phrase. * indicates significant difference with $p < 0.05$.

ated by 10-fold cross validation on the annotated corpus; results are shown in Table 9.

We used a sequence-labeling method to train statistical models for word segmentation, POS tagging, shallow parsing, and named entity recognition. CRF++ [61], an open-source implementation of the conditional random fields algorithm, was used to train these models. As shown in Table 9, the evaluation results of these modules trained by CRF++ are quite excellent in that all of them

Table 7
Entity type distribution in our annotated clinical texts.

Annotation type	Counts	% in the corresponding entity type
Diseases: possible	3255	39.09
Diseases: present	2685	32.24
Diseases: absent	2352	28.25
Diseases: not associated with the patient	35	0.42
Diseases: conditional	0	0.00
Diseases: occasional	0	0.00
Diseases: total	*8327*	*100.00*
Symptoms: absent	12,070	63.69
Symptoms: present	6426	33.91
Symptoms: conditional	257	1.36
Symptoms: occasional	153	0.81
Symptoms: possible	41	0.22
Symptoms: not associated with the patient	5	0.03
Symptoms: total	*18,952*	*100.00*
Treatments: present	3703	70.63
Treatments: historical	1414	26.97
Treatments: absent	126	2.40
Treatments: total	*5243*	*100.00*
Tests: total	*6989*	*100.00*

Table 8
Relation type distribution in our annotated clinical texts.

Annotation type	Description	Counts	% in the corresponding entity pair
TrAD	Treatment is administered for disease	393	58.66
TrID	Treatment improves disease	201	30.00
TrWD	Treatment worsen disease	70	10.45
TrCD	Treatment causes disease	6	0.90
R(Tr, D)		*670*	*100.00*
TrAS	Treatment is administered for symptom	614	30.37
TrIS	Treatment improves symptom	566	27.99
TrWS	Treatment worsen symptom	540	26.71
TrCS	Treatment causes symptom	298	14.74
TrNAS	Treatment is not administered because of symptom	4	0.20
R(Tr, S)		*2020*	*100.00*
TeRD	Test reveals disease	581	99.49
TeCD	Test conducted to investigate disease	3	0.51
R(Te, D)		*584*	*100.00*
TeRS	Test reveals symptom	1239	53.31
TeAS	Test is administered because of symptom	1085	46.69
R(Te, S)		*2324*	*100.00*
SID	Symptom indicates disease	1663	79.46
DCS	Disease causes symptom	430	20.54
R(D, S)		*2093*	*100.00*

R(entity1, entity2), relation between entity1 and entity2; D, diseases; S, symptoms; Te, tests; Tr, treatments.

achieved the level of practical application. To build a full parsing model, we trained the Stanford parser and the Berkeley parser [62] on our annotated corpus; results showed that both parsers were satisfactory, but that the Berkeley parser was slightly better. However, there were some null outputs in the Berkeley parser, so we used the corresponding outputs in the Stanford parser to replace them. This improvement further enhanced the evaluation of the full parser, and we chose this combined parser as our full parser. Moreover, similarly to most relation extraction research on English clinical texts, we used the support vector machines (SVM) algorithm to train models on our annotated Chinese clinical texts, and LIBSVM [63] was selected as the training tool.

5. Discussion

5.1. Contributions of this work

In this study, we constructed a comprehensive syntactic and semantic corpus of Chinese clinical texts, covering annotations for word segmentation, POS tagging, shallow parsing, full parsing, NER, assertion classification, and relation extraction.

Because extensive medical knowledge exists in clinical texts, we referred to annotation guidelines from the general domain and the clinical domain, and developed annotation guidelines for Chinese clinical texts with the help of physicians. As described in the guideline development section, many improvements were proposed to adapt to the characteristics of Chinese clinical texts.

Before building the corpus, annotators kept training by following annotation guidelines until their annotation consistency remained at a relatively high level. During the annotation period, existing open-source tools were used for pre-labeling, and significantly reduced the burden on annotators.

As is widely known, double annotation improves corpus quality; however, as the corpus scale grows, annotation costs in terms of time and money can be a challenge. Therefore, we balanced these factors and proposed an annotation method: double annotation was adopted in the annotator training stage; then, annotators were allowed to annotate separately in the corpus construction stage, using certain annotation quality assurance measures. The annotation consistency shows that our annotated corpus is of good quality.

Moreover, a Chinese clinical text processing and information extraction system was developed, and its modules can be seen as baselines for research in the clinical domain. To our knowledge, some of these modules described here are introduced into Chinese texts in the clinical domain for the first time, including the POS tagger, shallow parser, and full parser.

5.2. Limitations and future work

Although our annotated corpus makes a contribution to research on Chinese texts in the clinical domain, there are some limitations in our study. Because of limited annotation resources, the syntactic corpus only covers two departments within the Second Affiliated Hospital of Harbin Medical University. There are differences in medical terminologies from different hospital departments, which may weaken the adaptability of some NLP techniques across different departments.

As future work, some explorations will be conducted. First, transfer learning approaches will be introduced to solve the adaptation problem among different hospital departments. Second, some additional types of clinical text should be annotated to improve the practicability of NLP techniques developed based on this corpus. Third, active learning methods will be explored to reduce the annotation burden on annotators by filtering redundant samples from unlabeled data while selecting undertrained samples for the annotators. Finally, algorithms used to improve the performance of NLP systems for clinical texts will be developed.

6. Conclusions

In this paper, we described the construction of a corpus of Chinese clinical texts using an iterative annotation method. By following the annotation guidelines developed in this study, good levels of annotation consistency were achieved. Moreover, a CCTPIES

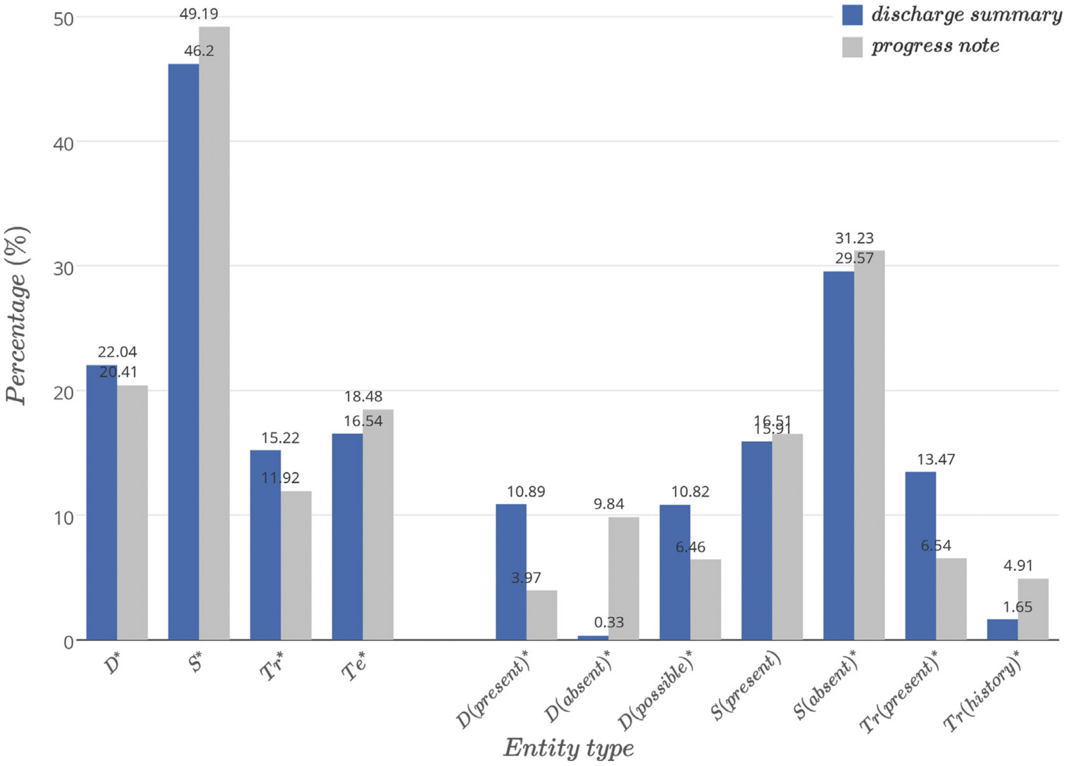


Fig. 8. Entity type distribution in Chinese discharge summaries and progress notes. The types, whose percentages in discharge summaries and progress notes are both below 1%, are not listed in this figure. D, diseases; S, symptoms; Tr, treatments, Te, tests. * indicates significant difference with $p < 0.05$.

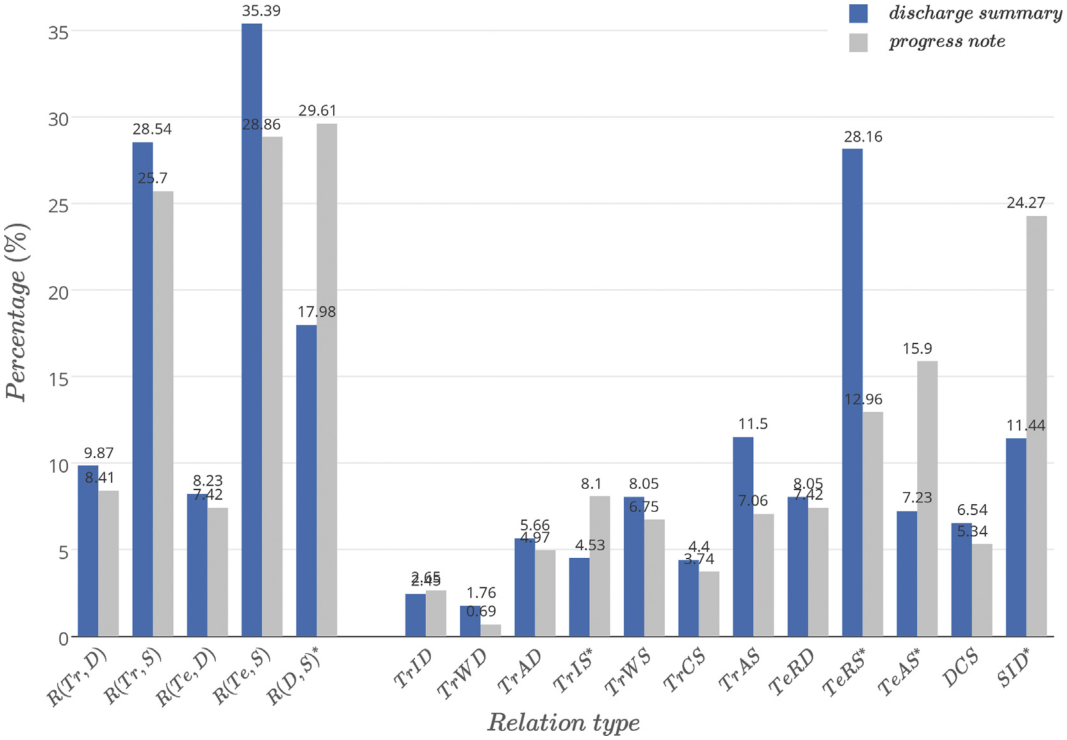


Fig. 9. Relation type distributions in Chinese discharge summaries and progress notes. The types, whose percentages in discharge summaries and progress notes are both below 1%, are not listed in this figure. D, diseases; S, symptoms; Tr, treatments, Te, tests; R (entity1, entity2), relation between entity1 and entity2; TrID, treatment improves disease; TrWD, treatment worsen disease; TrAD, treatment is administered for disease; TrIS, treatment improves symptom; TrWS, treatment worsen symptom; TrCS, treatment causes symptom; TrAS, treatment is administered for symptom; TeRD, test reveals disease; TeRS, test reveals symptom; TeAS, test is administered because of symptom; DCS, disease causes symptom; SID, symptom indicates disease. * indicates significant difference with $p < 0.05$.

Table 9
Performance of system modules trained on our annotated clinical texts.

Module	Precision	Recall	F1
Word segmenter	0.981	0.979	0.980
POS tagger	0.966	0.964	0.965
Shallow parser	0.946	0.949	0.948
Full parser	0.845	0.841	0.843
Named entity recognizer	0.923	0.902	0.912
Relation extractor	0.784	0.691	0.735

was developed to verify the usefulness of the corpus, which achieved excellent performance. To the best of our knowledge, this corpus is the first comprehensive annotated corpus of Chinese texts in the clinical domain, laying a solid foundation for future research. The related annotation resources are available at http://github.com/WILAB-HIT/Resources. The annotation guidelines, annotation tool, and data samples are now available for download. NLP shared tasks will be held after approval by the Medical Ethics Committee of the Second Affiliated Hospital of Harbin Medical University, and the whole corpus will then be released.

Author contributions

This work was a collaboration of all the authors. BH, JY, ZJ, and CQ developed the annotation guidelines and took part in corpus construction. BH, BD, YG, and QY performed corpus analysis. BH, ZJ, and JC developed system modules and evaluated their performance. All authors contributed to drafting, revision, and final approval of this manuscript.

Funding

This work was supported by the Ricoh Software Research Center (Beijing).

Conflict of interest

The authors have no conflicts of interest to declare.

Acknowledgements

We would like to thank the medical records room of the Second Affiliated Hospital of Harbin Medical University for providing the Chinese clinical texts used in this study. We are grateful to our physicians, YZ and YL, and the following annotators with backgrounds in computational linguistics: FZ, XD, TQ, and YP.

References

[1] T.J. Hannan, Electronic medical records, Heal Inf. Overv. (1996) 133–148.

[2] Electronic medical records basic specifications (trial), 2010, http://www.moh.gov.cn/mohyzs/s3585/201003/46174.shtml (Accessed 12.06.16).

[3] Measurement and standard of the capability level of electronic medical record system (trial), 2010, http://www.moh.gov.cn/mohyzs/s3586/201111/53274.shtml (Accessed 12.06.16).

[4] Functional specification of electronic medical record system (trial), 2010, http://www.moh.gov.cn/mohyzs/s3585/201012/50229.shtml (Accessed 12.06.16).

[5] D. Demner-Fushman, W.W. Chapman, C.J. McDonald, What can natural language processing do for clinical decision support?, J Biomed. Inform. 42 (2009) 760–772.

[6] G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, C.G. Chute, Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications, J. Am. Med. Inform. Assoc. 17 (2010) 507–513.

[7] Unified Medical Language System (UMLS), http://www.nlm.nih.gov/research/umls/ (Accessed 12.06.16).

[8] Informatics for Integrating Biology & the Bedside (i2b2), http://www.i2b2.org/ (Accessed 12.06.16).

[9] J. Yang, Q. Yu, Y. Guan, Z. Jiang, An overview of research on electronic medical record oriented named entity recognition and entity relation extraction, Acta Autom. Sin. 40 (2014) 1537–1562.

[10] P.M. Nadkarni, L. Ohno-Machado, W.W. Chapman, Natural language processing: an introduction, J. Am. Med. Inform. Assoc. 18 (2011) 544–551.

[11] M.P. Marcus, M.A. Marcinkiewicz, B. Santorini, Building a large annotated corpus of English: the penn treebank, Comput. Linguist. 19 (1993) 313–330.

[12] B. Santorini, Part-of-speech tagging guidelines for the Penn Treebank Project (3rd revision) (1990).

[13] A. Bies, M. Ferguson, K. Katz, R. MacIntyre, V. Tredinnick, G. Kim, M.A. Marcinkiewicz, B. Schasberger, Bracketing guidelines for Treebank II style Penn Treebank project, Univ. Pa. 97 (1995) 100.

[14] K.S. Tjong, Erik F, F. De Meulder, Introduction to the CoNLL-2003 shared task language-independent named entity recognition, Conf. Nat. Language Learning Hlt-Naacl (2003) 142–147.

[15] R. Grishman, D. Westbrook, A. Meyers, Nyu's English ace 2005 system description, J. Satisf. 51 (2005) 1927–1938.

[16] I. Hendrickx, S.N. Kim, Z. Kozareva, P. Nakov, D. Aghdha, Sebastian, M. Pennacchiotti, L. Romano, S. Szpakowicz, SemEval-2010 task 8 multi-way classification of semantic relations between pairs of nominals, The Workshop on Semantic Evaluations: Recent Achievements and Future Directions (2009) 94–99.

[17] J.D. Kim, T. Ohta, Y. Tateisi, J. Tsujii, GENIA corpus—a semantically annotated corpus for bio-textmining, Bioinformatics 19 (Suppl 1) (2003) i180–i182.

[18] S. Kulick, A. Bies, M. Liberman, M. Mandel, R. Mcdonald, M. Palmer, A. Schein, L. Ungar, S. Winters, P. White, Integrated annotation for biomedical information extraction (2004).

[19] K. Franzén, G. Eriksson, F. Olsson, L. Asker, P. Lidén, J. Cöster, Protein names and how to find them, Int. J. Med. Inform. 67 (2003) 49–61.

[20] L. Tanabe, N. Xie, L.H. Thom, W. Matten, W.J. Wilbur, GENETAG: a tagged corpus for gene/protein named entity recognition, BMC Bioinf. 6 (2005) 1–7.

[21] M. Bada, M. Eckert, D. Evans, K. Garcia, K. Shipley, D. Sitnikov, W.A.B. Jr, K.B. Cohen, K. Verspoor, J.A. Blake, Concept annotation in the CRAFT corpus, BMC Bioinf. 13 (2012) 1–20.

[22] B. Rosario, M.A. Hearst, Classifying semantic relations in bioscience text, in: Meeting of the Association for Computational Linguistics, 21–26 July, 2004, Barcelona, Spain, 2004, pp. 430–437.

[23] B. Alex, C. Grover, B. Haddow, M. Kabadjov, E. Klein, M. Matthews, S. Roebuck, R. Tobin, X. Wang, The ITI TXM corpora tissue expressions and protein-protein interactions (2008).

[24] D. Albright, A. Lanfranchi, A. Fredriksen, W.F. Styler, C. Warner, J.D. Hwang, J.D. Choi, D. Dligach, R.D. Nielsen, J. Martin, Towards comprehensive syntactic and semantic annotations of the clinical narrative, J. Am. Med. Inform. Assoc. 20 (2013) 922–930.

[25] J.-W. Fan, E.W. Yang, M. Jiang, R. Prasad, R.M. Loomis, D.S. Zisook, J.C. Denny, H. Xu, Y. Huang, Syntactic parsing of clinical text: guideline and corpus development with handling ill-formed sentences, J. Am. Med. Inform. Assoc. 20 (2013) 1168–1177.

[26] Y. Xu, Y. Wang, T. Liu, J. Liu, Y. Fan, Y. Qian, J. Tsujii, E.I. Chang, Joint segmentation and named entity recognition using dual decomposition in Chinese discharge summaries, J. Am. Med. Inform. Assoc. 21 (2014) e84–e92.

[27] S. Zhang, T. Kang, X. Zhang, D. Wen, N. Elhadad, J. Lei, Speculation detection for Chinese clinical notes: impacts of word segmentation and embedding models, J. Biomed. Inform. 60 (2016) 334–341.

[28] S. Meystre, P.J. Haug, Natural language processing to extract medical problems from electronic clinical documents: performance evaluation, J. Biomed. Inform. 39 (2006) 589–599.

[29] A. Roberts, R. Gaizauskas, M. Hepple, G. Demetriou, Y. Guo, I. Roberts, A. Setzer, Building a semantically annotated corpus of clinical texts, J. Biomed. Inform. 42 (2009) 950–966.

[30] Ö. Uzuner, B.R. South, S. Shen, S.L. DuVall, I2b2/VA challenge on concepts, assertions, and relations in clinical text, J. Am. Med. Inform. Assoc. 18 (2011) (2010) 552–556.

[31] N. Elhadad, S. Pradhan, W. Chapman, S. Manandhar, G. Savova, SemEval-2015 task 14: analysis of clinical text, proc of workshop on semantic evaluation, Assoc. Comput. Linguistics (2015) 303–310.

[32] J. Lei, B. Tang, X. Lu, K. Gao, M. Jiang, H. Xu, A comprehensive study of named entity recognition in Chinese clinical text, J. Am. Med. Inform. Assoc. 21 (2014) 808–814.

[33] Y. Wang, Z. Yu, L. Chen, Y. Chen, Y. Liu, X. Hu, Y. Jiang, Supervised methods for symptom name recognition in free-text clinical records of traditional Chinese medicine: an empirical study, J. Biomed. Inform. 47 (2014) 91–104.

[34] H. Wang, W. Zhang, Q. Zeng, Z. Li, K. Feng, L. Liu, Extracting important information from Chinese operation notes with natural language processing methods, J. Biomed. Inform. 48 (2014) 130–136.

[35] Z. Jia, H. Li, M. Ju, Y. Zhang, Z. Huang, C. Ge, H. Duan, A finite-state automata based negation detection algorithm for chinese clinical documents, in: Progress in Informatics and Computing (PIC), 2014 International Conference on, 2014, pp. 128–132.

[36] D. Xu, M. Zhang, T. Zhao, C. Ge, W. Gao, J. Wei, K.Q. Zhu, Data-driven information extraction from Chinese electronic medical records, PLoS One 10 (2015) e0136270.

[37] Y.-B. Li, X.-Z. Zhou, R.-S. Zhang, Y.-H. Wang, Y. Peng, J.-Q. Hu, Q. Xie, Y.-X. Xue, L.-L. Xu, X.-F. Liu, Detection of herb-symptom associations from traditional Chinese medicine clinical data, Evid. Complement. Altern. Med. 2015 (2015).

[38] L. Hirschman, N. Sager, Automatic information formatting of a medical sublanguage, Sublanguage Stud. Lang. Restricted Semant. (1982) 27–80.

[39] C. Friedman, P. Kra, A. Rzhetsky, Two biomedical sublanguages: a description based on the theories of Zellig Harris, J. Biomed. Inform. 35 (2002) 222–235.

[40] A. Rector, J. Rogers, A. Taweel, D. Ingram, D. Kalra, J. Milan, P. Singleton, R. Gaizauskas, M. Hepple, D. Scott, CLEF joining up healthcare with clinical and post-genomic research (2003).

[41] Fourth i2b2/VA Shared-Task and Workshop Challenges in Natural Language Processing for Clinical Data, 2010, http://www.i2b2.org/NLP/Relations/ (Accessed 12.06.16).

[42] 2010 i2b2/VA Challenge Evaluation Concept Annotation Guidelines, 2010, http://www.i2b2.org/NLP/Relations/assets/Concept%20Annotation%20Guideline.pdf (Accessed 12.06.16).

[43] 2010 i2b2/VA Challenge Evaluation Assertion Annotation Guidelines, 2010, http://www.i2b2.org/NLP/Relations/assets/Assertion%20Annotation%20Guideline.pdf (Accessed 12.06.16).

[44] 2010 i2b2/VA Challenge Evaluation Relation Annotation Guidelines, 2010, http://www.i2b2.org/NLP/Relations/assets/Relation%20Annotation%20Guideline.pdf (Accessed 12.06.16).

[45] O. Uzuner, J. Mailoa, R. Ryan, T. Sibanda, Semantic relations for problem-oriented medical records, Artif. Intell. Med. 50 (2010) 63–73.

[46] The Segmentation Guidelines for the Penn Chinese Treebank (3.0), 2000, http://www.cis.upenn.edu/~chinese/segguide.3rd.ch.pdf (Accessed 28.10.15).

[47] The part-of-speech tagging guidelines for the penn Chinese treebank (3.0), 2000, http://repository.upenn.edu/cgi/viewcontent.cgi?article=1039&context=ircs_reports (Accessed 27.06.16).

[48] The bracketing guidelines for the penn Chinese treebank (3.0), 2000, http://www.cis.upenn.edu/~chinese/parseguide.3rd.ch.pdf (Accessed 28.10.15).

[49] The segmentation annotation guidelines on Chinese clinical texts, 2016, http://github.com/WILAB-HIT/Resources/blob/master/segmentation_pos_parsing/annotation_guidelines/Seg.pdf (Accessed 24.10.16).

[50] The part-of-speech tagging annotation guidelines on Chinese clinical texts, 2016, http://github.com/WILAB-HIT/Resources/blob/master/segmentation_pos_parsing/annotation_guidelines/POS.pdf (Accessed 24.10.16).

[51] The bracketing annotation guidelines on Chinese clinical texts, 2016, http://github.com/WILAB-HIT/Resources/blob/master/segmentation_pos_parsing/annotation_guidelines/Bracketing.pdf (Accessed 24.10.16).

[52] The entity-assertion-relation annotation guidelines on Chinese clinical texts, 2016, http://github.com/WILAB-HIT/Resources/blob/master/entity_assertion_relation/annotation_guidelines/Entity_Assertion_Relation.pdf (Accessed 24.10.16).

[53] Y. Shi-wen, D. Hui-ming, Z. Xue-feng, S. Bin, The basic processing of contemporary Chinese corpus at peking university, J. Chin. Inf. Process. 16 (2002) 51–66.

[54] J. Pustejovsky, A. Stubbs, Increasing informativeness in temporal annotation, in: Proceedings of the 5th Linguistic Annotation Workshop, 2011, pp. 152–160.

[55] NLPIR/ICTCLAS2016, http://github.com/NLPIR-team/NLPIR/tree/master/NLPIR%20SDK/NLPIR-ICTCLAS (Accessed 12.06.16).

[56] Stanford POS Tagger, http://nlp.stanford.edu/software/tagger.shtml (Accessed 27.06.16).

[57] Stanford Parser, http://nlp.stanford.edu/software/lex-parser.shtml (Accessed 27.06.16).

[58] G. Hripcsak, A.S. Rothschild, Agreement, the f-measure, and reliability in information retrieval, J. Am. Med. Inform. Assoc. 12 (2005) 296–298.

[59] Evalb, http://nlp.cs.nyu.edu/evalb/ (Accessed 27.06.16).

[60] The Chinese Clinical Text Processing and Information Extraction System, 2016, http://wi.hit.edu.cn/cemr/ (Accessed 24.10.16).

[61] CRF++, http://taku910.github.io/crfpp/ (Accessed 28.10.15).

[62] Berkeley Parser, http://github.com/slavpetrov/berkeleyparser (Accessed 27.06.16).

[63] LIBSVM, http://www.csie.ntu.edu.tw/~cjlin/libsvm/ (Accessed 28.10.15).