

Fall 12-2014

# Named Entity Recognition in Chinese Clinical Text

Jianbo Lei

*University of Texas Health Science Center at Houston, Jianbo.Lei@uth.tmc.edu*

Follow this and additional works at: [http://digitalcommons.library.tmc.edu/uthshis\\_dissertations](http://digitalcommons.library.tmc.edu/uthshis_dissertations)

 Part of the [Bioinformatics Commons](#), and the [Medicine and Health Sciences Commons](#)

---

## Recommended Citation

Lei, Jianbo, "Named Entity Recognition in Chinese Clinical Text" (2014). *UT SBMI Dissertations (Open Access)*. 31.  
[http://digitalcommons.library.tmc.edu/uthshis\\_dissertations/31](http://digitalcommons.library.tmc.edu/uthshis_dissertations/31)

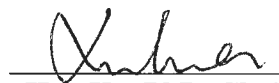
This is brought to you for free and open access by the School of Biomedical Informatics at DigitalCommons@TMC. It has been accepted for inclusion in UT SBMI Dissertations (Open Access) by an authorized administrator of DigitalCommons@TMC. For more information, please contact [laurel.sanders@library.tmc.edu](mailto:laurel.sanders@library.tmc.edu).

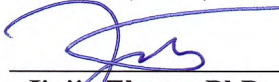
# Named entity recognition in Chinese clinical text

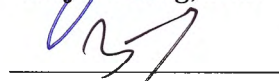
By

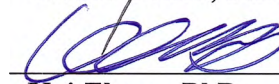
Jianbo Lei, M.D. M.S. M.A.

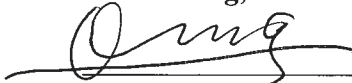
APPROVED:

  
Hua Xu, PhD, Chair

  
Jiajie Zhang, PhD

  
Trevor Cohen, PhD

  
Kai Zheng, PhD

  
Qun Meng, PhD

Date approved: 12-19-14

Named entity recognition in Chinese clinical text

A  
Dissertation

Presented to the Faculty of  
The University of Texas  
Health Science Centre at Houston  
School of Biomedical Informatics  
in Partial Fulfilment of the Requirements for the Degree of  
  
Doctor of Philosophy

By  
  
Jianbo Lei, M.D. M.A. M.S.

University of Texas Health Science Centre at Houston

2014

Dissertation Committee:

Hua Xu, PhD<sup>1</sup>, Advisor  
Jiajie Zhang, PhD<sup>1</sup>  
Trevor Cohen, PhD<sup>1</sup>  
Kai Zheng, PhD<sup>2</sup>  
Qun Meng, PhD<sup>3</sup>

<sup>1</sup>The School of Biomedical Informatics, University of Texas Health Science Centre at Houston

<sup>2</sup>The School of Information, University of Michigan

<sup>3</sup>The Centre for Statistics and Information, China National Health and Family Planning Commission

Copyright by

Jianbo Lei

2014

## **Dedication**

Dedicated to my motivation to improve the health outcomes of the population by using the state of the art technologies and my dream to steer the research and development of biomedical informatics as a discipline in China.

## **Acknowledgements**

I would like to thank my advisor, Dr. Hua Xu for guiding and supporting me over the years. You have set an example of excellence as a researcher, mentor, instructor, and role model. You especially assisted me in quickly transiting and completing my research idea to a systematic research study, which is amazing. I would also like to thank my dissertation committee members for all of their guidance through this process; your discussion, ideas, and feedback have been absolutely invaluable. Your broad knowledge, rigorous research method and unselfish support have made me produce a number of projects and papers in different fields in such a short time. I'd like to thank my fellow graduate students, research technicians, collaborators, and the multitude of undergraduates who contributed to this research. I am very grateful to all of you.

## Abstract

**Objective:** Named entity recognition (NER) is one of the fundamental tasks in natural language processing (NLP). In the medical domain, there have been a number of studies on NER in English clinical notes; however, very limited NER research has been done on clinical notes written in Chinese. The goal of this study is to develop corpora, methods, and systems for NER in Chinese clinical text.

**Materials and methods:** To study entities in Chinese clinical text, we started with building annotated clinical corpora in Chinese. We developed an NER annotation guideline in Chinese by extending the one used in the 2010 i2b2 NLP challenge. We randomly selected 400 admission notes and 400 discharge summaries from Peking Union Medical College Hospital (PUMCH) in China. For each note, four types of entities including clinical problems, procedures, labs, and medications were annotated according to the developed guideline. In addition, an annotation tool was developed to assist two MD students to annotate Chinese clinical documents. A comparison of entity distribution between Chinese and English clinical notes (646 English and 400 Chinese discharge summaries) was performed using the annotated corpora, to identify the important features for NER. In the NER study, two-thirds of the 400 notes were used for training the NER systems and one-third were used for testing. We investigated the effects of different types of features including bag-of-characters, word segmentation, part-of-speech, and section

information, with different machine learning (ML) algorithms including Conditional Random Fields (CRF), Support Vector Machines (SVM), Maximum Entropy (ME), and Structural Support Vector Machines (SSVM) on the Chinese clinical NER task. All classifiers were trained on the training dataset, evaluated on the test set, and micro-averaged precision, recall, and F-measure were reported.

**Results:** Our evaluation on the independent test set showed that most types of features were beneficial to Chinese NER systems, although the improvements were limited. By combining word segmentation and section information, the system achieved the highest performance, indicating that these two types of features are complementary to each other. When the same types of optimized features were used, CRF and SSVM outperformed SVM and ME. More specifically, SSVM reached the highest performance among the four algorithms, with F-measures of 93.51% and 90.01% for admission notes and discharge summaries respectively.

**Conclusions:** In this study, we created large annotated datasets of Chinese admission notes and discharge summaries and then systematically evaluated different types of features (e.g., syntactic, semantic, and segmentation information) and four ML algorithms including CRF, SVM, SSVM, and ME for clinical NER in Chinese. To the best of our knowledge, this is one of the earliest comprehensive effort in Chinese clinical NER research and we believe it will provide valuable insights to NLP research in Chinese clinical text. Our results suggest that both word segmentation and section information improves NER in Chinese clinical text, and SSVM, a recent sequential labelling algorithm, outperformed CRF and other classification algorithms. Our best system



achieved F-measures of 90.01% and 93.52% on Chinese discharge summaries and admission notes, respectively, indicating a promising start on Chinese NLP research.

## **Vita**

1993.....M.D., Clinical Medicine, West China  
..... University of Medical Sciences, P.R. China

2002.....M.S., Computer Sciences, Columbia  
..... University in the City of New York

2005.....M.A., Medical Informatics, Columbia  
..... University in the City of New York

2012 to present.....Graduate student, School of Biomedical  
..... Informatics, University of Texas Health  
.....Sciences Centre at Houston

## **Publications**

- [1]. Jianbo Lei, Buzhou Tang, Xueqin Lu, Kaihua Gao, Min Jiang, Hua Xu. A comprehensive study of named entity recognition in Chinese clinical text. J Am Med Inform Assoc. Dec. 17,2013. doi:10.1136/amiajnl-2013-002381
- [2]. Jianbo Lei, Pengcheng Guan, Kaihua Gao, Xueqin Lu, Yuefeng Li, Qun Meng, Jiajie Zhang, Dean F. Sittig, Kai Zheng. Characteristics of Health IT Outage and Suggested Risk Management Strategies: An Analysis of Historical Incident Reports in China.

International Journal of Medical Informatics. 2014 Feb;83(2):122-30. doi:  
10.1016/j.ijmedinf.2013.10.006. Epub 2013 Oct 22.

- [3]. Yonghui Wu, Jianbo Lei\* (co-first author), Wei-Qi Wei, Buzhou Tang, Joshua C. Denny, S. Trent Rosenbloom, Randolph A. Miller, Dario A. Giuse, Kai Zheng, Hua Xu. Analyzing Differences between Chinese and English Clinical Text: A Cross-Institution Comparison of Discharge Summaries in Two Languages. *Stud Health Technol Inform.* 2013;192:662-6.
- [4]. Jianbo Lei, Paulina Sockolow, Pengcheng Guan, Qun Meng, Jiajie Zhang, A Comparison of Electronic Health Records at Two Major Peking University Hospitals in China to United States Meaningful Use Objectives. *BMC medical informatics and decision making.* 2013, 13:96(28 Aug 2013). DOI: 10.1186/10.1186/1472-6947-13-96, URL: <http://www.biomedcentral.com/1472-6947/13/96> (became as “highly accessed”)
- [5]. Jianbo Lei, Lufei Xu, Qun Meng, Jiajie Zhang, Yang Gong. The Current Status of Usability Studies of Information Technologies in China: a systematic study. *Biomed Research International* vol. 2014, Article ID 568303, 10 pages, 2014.  
doi:10.1155/2014/568303 (<http://www.hindawi.com/journals/bmri/2014/568303/>)

## **Field of Study**

Health Informatics

## Table of Contents

Dedication .....	ii
Acknowledgements .....	iii
Abstract .....	iv
Vita .....	vii
Publications .....	vii
Table of Contents .....	ix
List of Tables .....	xi
List of Figures .....	xii
<b>Chapter 1: Introduction</b> .....	1
1.1. MOTIVATIONS .....	1
1.2. HYPOTHESIS .....	2
1.3. SPECIFIC AIMS .....	2
<b>Chapter 2: Literature Review</b> .....	5
2.1. AN OVERVIEW OF CLINICAL NLP RESEARCH .....	5
2.1.1. Major tasks of clinical NLP .....	5
2.1.2. Clinical NLP systems .....	6
2.1.3. NLP applications in the medical domain .....	9
2.2. NER IN CLINICAL TEXT .....	10
2.2.1. An introduction to NER .....	10
2.2.2. Relevant work of clinical NER in English .....	13
2.2.3. Relevant work of clinical NER in Chinese .....	17
2.3. SUMMARY .....	20
<b>Chapter 3: Create an annotated corpus of Chinese clinical texts</b> .....	21
3.1. INTRODUCTION TO CLINICAL CORPORA CONSTRUCTION .....	21
3.2. METHODS .....	24
3.2.1. Data sets .....	25
3.2.2. Development of annotation guideline for Chinese clinical texts .....	25

3.2.3. <i>Development of annotation tool for Chinese clinical text</i> .....	26
3.2.4. <i>Conducting of annotation</i> .....	27
3.4. RESULTS.....	28
3.4.1. <i>Corpus statistics</i> .....	28
3.4.2. <i>Quality of the corpora</i> .....	29
3.5. DISCUSSION.....	29
<b>Chapter 4: Compare entity distribution between Chinese and English clinical documents</b> .....	32
4.1. INTRODUCTION.....	32
4.2. METHODS .....	35
4.2.1 <i>Data sets</i> .....	35
4.2.2. <i>Analytic Methods</i> .....	35
4.3. RESULTS.....	38
4.4. DISCUSSION.....	42
<b>Chapter 5: Develop and evaluate machine learning based NER approaches for Chinese clinical text</b> .....	47
5.1. INTRODUCTION.....	47
5.2. METHODS .....	48
5.2.1. <i>Datasets and annotation</i> .....	48
5.2.2. <i>ML-based NER</i> .....	48
5.3. EXPERIMENTS AND EVALUATION.....	53
5.4. RESULTS.....	56
5.5 DISCUSSION.....	59
<b>Chapter 6: Key findings, Contribution, Future work and Conclusions</b> .....	63
6.1. OVERVIEW AND SUMMARY OF KEY FINDINGS .....	63
6.2. INNOVATIONS AND CONTRIBUTIONS.....	64
6.3. FUTURE WORK .....	65
6.4. CONCLUSION .....	66
<b>References</b> .....	67

## List of Tables

Table 1. General NLP systems in the medical domain .....	8
Table 2. Methods of the top teams in 2010 i2b2 challenge for concept extraction .....	15
Table 3. Comparison of NER studies in Chinese .....	19
Table 4. A list of available annotation tools .....	23
Table 5. Summary statistics of annotated datasets of Chinese discharge summaries and admission notes .....	29
Table 6. Distribution of different types of entities .....	37
Table 7. Entity density within 9 common sections across four institutions .....	41
Table 8. The performance of the CRF-based NER systems on Chinese admission and discharge notes when different features were used .....	57
Table 9. The detailed results of the best CRF-based NER system on admission and discharge summaries for each entity type .....	58
Table 10. Comparison of four state-of-the-art machine learning algorithms on Chinese admission and discharge summaries when optimized features were used .....	59

## List of Figures

Figure 1. Number of NLP publications in PubMed.....	10
Figure 2. Top 10 teams for 2009 i2b2 challenge .....	15
Figure 3. Screenshot of annotation tool .....	27
Figure 4. Annotation workflow and IAA results .....	28
Figure 5. Workflow of the entity distribution comparison study.....	38
Figure 6. Zipf’s distribution of vocabularies .....	39
Figure 7. Normalized distribution of annotated entities .....	39
Figure 8. Relative frequency of Problems, Tests, and Treatments in three English institutions: UPMC, PARTNERS, and BETH, and one Chinese institution: PUMCH.....	42
Figure 9. Visualization of entity density within 9 common sections across four institutions .....	45
Figure 10. Examples of Chinese medical named entity recognition (NER) representation .....	49
Figure 11. Features used for Chinese medical entity recognition .....	50
Figure 12. An example of word segmentation in Chinese .....	50

## **Chapter 1: Introduction**

### **1.1. Motivations**

Clinical documents are an important type of data in electronic health records (EHRs) and often contain valuable and detailed patient information for many clinical applications. Natural language processing (NLP), a technology that can unlock information embedded in free text, has received much attentions in the medical domain (Meystre, Savova, Kipper-Schuler, & Hurdle, 2008). Clinical NLP has become an active research area in the Biomedical Informatics field and many studies have successfully demonstrated its uses in clinical practice (e.g., facilitating clinical decision support systems) (Demner-Fushman, Chapman, & McDonald, 2009) as well as in biomedical research (Kho et al., 2011).

Named entity recognition (NER) in clinical text, which is to identify the boundaries of clinically relevant entities such as diseases and drugs, is one of the fundamental tasks in clinical NLP research and has been extensively studied, including dictionary-based approaches used in early general clinical NLP systems such as MedLEE (Friedman, 1997; Friedman, 2000) and MetaMap (Aronson & Lang, 2010; Aronson, 2001) as well as more recent machine learning (ML) based approaches in shared clinical NLP tasks. (O. Uzuner & DuVall, 2010; Uzuner, Solti, & Cadag, 2010) However, most previous studies in clinical NER have primarily focused on clinical text written in English. Very few studies have investigated NLP methods such as NER for clinical text in Chinese.



With the rapid growth of EHRs in China, huge amounts of clinical data including narrative text have been generated every day in China. To efficiently utilize these data for computerized clinical applications or for biomedical research, automated methods have to be developed to extract structured information from narrative clinical text in Chinese. Although there are extensive research efforts on NLP in Chinese (Zheng, 2008; Duan, 2003; Shi et al., 2007; Zhao, 2008; Aaron & Lidia, 2013; Zheng, Liu, & Du, 2012; Xiaoshan, 2002; Tiejun, Ting, & Qiang, 2007), few researchers have investigated NLP methods for Chinese clinical text (Wang et al., 2014; Wang et al., 2014; Wang et al., 2010), probably due to the lack of access to Chinese clinical data, as well as the scarcity of medical informatics researchers in China. Therefore, the ultimate goal of my research is to promote clinical NLP research in Chinese, by developing comprehensive resources and novel methods for processing Chinese clinical text. As a first step, this dissertation work focuses on the NER task for Chinese clinical text.

## **1.2. Hypothesis**

After investigating previous clinical NER studies in English and current status of NER methods in general Chinese text, we propose the following hypothesis:

By creating annotated clinical Chinese corpora, optimizing linguistics and domain specific features, and implementing state-of-the-art machine learning algorithms, we will be able to build ML-based NER methods to detect clinical entities in Chinese clinical text with a reasonable performance similar to that obtained by English language systems.

## **1.3. Specific aims**

In this study, our goal is to develop resources and methods for NER in Chinese clinical documents. Our specific aims include:

*Specific Aim 1: create annotated corpora of Chinese clinical text.*

An annotated corpus is required for developing and evaluating NER approaches in Chinese clinical text. We collected 400 discharge summaries and 400 admission notes from PUMCH (Peking Union Medical College hospital) and recruited two MD students to annotate four types of clinical entities: medical problems, tests, medications, and procedures. An annotation guideline and an internet-based annotation tool were also developed by the research team to assist the annotation procedure.

*Specific Aim 2: compare entity distribution between Chinese and English clinical documents.*

There has been an increasing trend in cross-country collaboration on medical research using EHR data, e.g., between the US and China. However, few studies have investigated characteristics of Chinese EHR data and the differences in EHR data from the US and China are unknown. This aim attempts to understand system and cultural differences that may exist between Chinese and English clinical documents and identify those features that may be valuable for the next step in NER method development. We compared the annotated Chinese discharge summary corpus (in Aim 1) with corpora from three US institutions (646 notes) and manually analyzed distributions of clinical entities and potential features for ML-based NER methods.

*Specific Aim 3: develop and evaluate machine learning based NER approaches for Chinese clinical text.*

According to Rosenbloom et al. (Rosenbloom et al., 2011) , NLP is a potential solution to make narrative clinical data re-usable. However, performance of current NLP methods on Chinese clinical text is unknown. Although there have been a number of studies on NER in English clinical notes, very limited NER research has been done on clinical notes written in Chinese. Therefore, we propose to systematically investigate features and machine learning algorithms for NER in Chinese clinical text, in order to develop state-of-the-art methods and systems for clinical NER in Chinese.

## Chapter 2: Literature Review

### 2.1. An overview of clinical NLP research

#### 2.1.1. Major tasks of clinical NLP

NLP is defined as “*a field of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human (natural) languages*” (Wikipedia). In open domains, massive textual data such as web content have guided the development of NLP technologies for different applications such as information retrieval, information extraction, and sentiment analysis (R. Subhashini, 2011; Riloff, 2003; Luo, Michael, & Stephan, 2004; Alberto et al., 2008; Heng-Li, 2014). In the medical domain, clinical documents in EHRs present unique characteristics (e.g., pervasive abbreviations (Pakhomov, Pedersen, & Chute, 2005; Xu, Stetson & Friedman, 2007), telegraphic styles (Fan et al., 2013), and restricted semantic patterns (Friedman, Kra, & Rzhetsky, 2002) and novel methods are required to address these challenges (Friedman, Kra, & Rzhetsky, 2002; Carol & Milton, 2013). Nadkarni et al. (Prakash & Wendy, 2011) systematically summarized current ongoing efforts of clinical NLP research and categorized them into two groups: lower level tasks and high level tasks.

Lower level tasks refer to problems such as sentence boundary detection (Florian et al., 2010), tokenization (Neil & Jens, 2011), part-of-speech tagging (Barrett & Weber-Jahnke, 2014), morphological analysis of medical terms (Grabar, Rizand, Livartowski, &

Hamon, 2009; Deleger & Namer, 2009), shallow parsing (chunking) for identifying phrases from constituent part-of-speech tagged tokens, and problem-specific segmentation such as section identification (Denny et al., 2009).

Higher-level tasks build on low-level tasks and are usually problem-specific. They include: spelling/grammatical error identification and recovery, word sense disambiguation (WSD) (Rindflesch & Aronson, 1994; Pedersen, 2011; Weeber & Aronson, 2001), clinical concept extraction (Aronson, 2001; Zou et al., 2003), identification of negation and uncertainty of concepts (Chapman, 2001; Mutalik & Nadkarni, 2001; Huang, 2007), relationship extraction including determining relationships between entities or events and temporal relationship extraction (Tao et al., 2011; Hripcsak et al., 2009), and co-reference resolution.

### **2.1.2. Clinical NLP systems**

Many successful NLP systems in the clinical domain have been developed since the 1960s. Many early clinical NLP systems used symbolic-based approaches, which often rely on manually extracted explicit representation of facts about language. More recently, statistical (or hybrid) NLP methods such as supervised machine learning algorithms are being increasingly applied to the medical domain and have shown promising results.

Naomi Sager's pioneering work in the 1970s, based on language theories of Zellig Harris

(Harris, 1968; Harris, 1982; Harris, 1991), demonstrated an approach to structure clinical information occurring in text (Sager et al., 1987). In the late 1980s, other early NLP systems also showed that NLP was feasible in the clinical domain (F. W., 1987; AT. M., 1991; Baud & Scherrer, 1992; Friedman et al., 1994; Haug et al., 1994) and improved healthcare (Hripcsak et al., 1995). The 1980s and 1990s witnessed development of substantial resources for NLP. In the early 2000s, open source NLP tools in the biomedical domain also became available and those tools can now be registered and accessed online via the orbit project (<http://orbit.nlm.nih.gov>). Currently, there are a number of general purpose clinical NLP systems available, such as MedLEE (Medical Language Extraction and Encoding) (Friedman, 1997; Friedman et al., 1994), MetaMap (Aronson & Lang, 2010; Aronson, 2001), KMCI (Knowledgemap Concept Identifier) (Denny, Smithers, Miller, & Spickard, 2003), and cTAKES (Clinical Text Analysis and Knowledge Extraction System) (Savova et al., 2010). MedLEE was developed by Dr. Carol Friedman in the 1990s. It is mainly a semantic rule-based system founded on the sublanguage theory and implemented in Quintus Prolog. MedLEE started with radiological reports and was later extended to mammography notes (Jain & Friedman, 1997), discharge summaries (Melton & Hripcsak, 2005; Friedman, Lussier, & Hripcsak, 2004), radiology reports (Jain, Knirsch, Friedman, & Hripcsak, 1996) and pathology notes (Xu, Grann, & Friedman, 2004). MetaMap was initially developed for biomedical literature to map biomedical text to UMLS (Unified Medical Language Systems) concepts (AR A., 2001) and recently used to process clinical notes. cTAKES was initiated from a Mayo-IBM collaboration in 2000 (Savova et al., 2010). It is a modular

system of pipelined components based on the IBM UIMA framework using both rule-based and ML techniques. The KnowledgeMap Concept Identifier (“KnowledgeMap”) was developed by Denny et al (Denny, Miller, & Spickard, 2003; Denny et al., 2005). HiTEX is another clinical NLP system based on the GATE framework (Goryachev, Sordo, & Zeng, 2006). Table 1 shows the description of these general purpose clinical NLP systems.

Table 1 *General NLP systems in the medical domain*

System	Description	Publicly available	Publication
MedLEE	An expert-based system for unlocking clinical information from narratives	No	Friedman and Hripcsak (Friedman et al., 1994; Hripcsak et al., 1994)
MetaMap	An expert based system for mapping text to the Unified Medical Language System	Yes	Aronson and Lang(Aronson, 2001; Aronson & Lang, 2010)
HiTEX	An NLP system distributed through i2b2	Yes	Goryachev, Sordo et al. (Goryachev, Sordo, & Zeng, 2006)
KnowledgeMap Concept Identifier(KMCI)	Rigorous NLP techniques and document-and context-based disambiguation methods to identify UMLS concepts in biomedical documents	No	Denny et al.(Denny, Miller & Spickard, 2003; Denny et al., 2005)

cTAKES	A pipeline built around openNLP, Lucene, and LVG for concept normalization	Yes	Savova, Masanz, et al, (Savova et al., 2010)
--------	--	-----	--

In addition to these general purpose NLP systems, researcher have also developed various tools for specific tasks in clinical NLP. For example, NegEx (Mitchell et al., 2004) and ConText (Harkema, Dowling, Thornblade, & Chapman, 2009) are two widely used tools for detecting negation of and other contextual information about clinical concepts. There are also other systems that focus on extracting specific types of entities, such as medications (Xu et al., 2010; Li et al., 2013) or temporal expressions (Tang et al., 2013; Wu, Juhn, Sohn, & Liu, 2014) from clinical text.

### **2.1.3. NLP applications in the medical domain**

In the medical domain, NLP is crucial for advancing healthcare because it is needed to transform relevant information locked in text into structured data that can be used by computer processes aimed at improving patient care and advancing medicine (Carol & Milton, 2013). Research and applications in biomedical NLP have increased enormously in the last 30 years and have become a prominent activity because of the explosive amount of information in text concerned with biomedical research, clinical care, as well as consumer health information on the web. The broad clinical areas that require NLP techniques and applications include decision support (Imler & Imperiale, 2013; Pai et al., 2014; Byrd, Sun, Ebadollahi, & Stewart, 2013), cohort identification (Zhu, Carterette, &



Liu, 2014), patient management (Gawron, Keswani, Rasmussen, & Kho, 2014), question answering (Terol & Palomar, 2007), knowledge acquisition (Weng, Velez, Johnson, & Bakken, 2014), phenotype characterization (Shivade et al., 2014; Wang & Friedman, 2009; Chen, 2004; C C., 2002), data mining and clinical research (Doddi, Ravi, & Torney, 2001; Hripcsak, Alderson, Friedman, 2002; Wilcox, 2000), biosurveillance (Chapman & Wagner, 2004; Chapman et al., 2004), and adverse drug reaction detection (Wang et al., 2009). Figure 1 illustrates the increasing number of publications on NLP in MEDLINE per year, showing a remarkable increase starting in the 1990s.

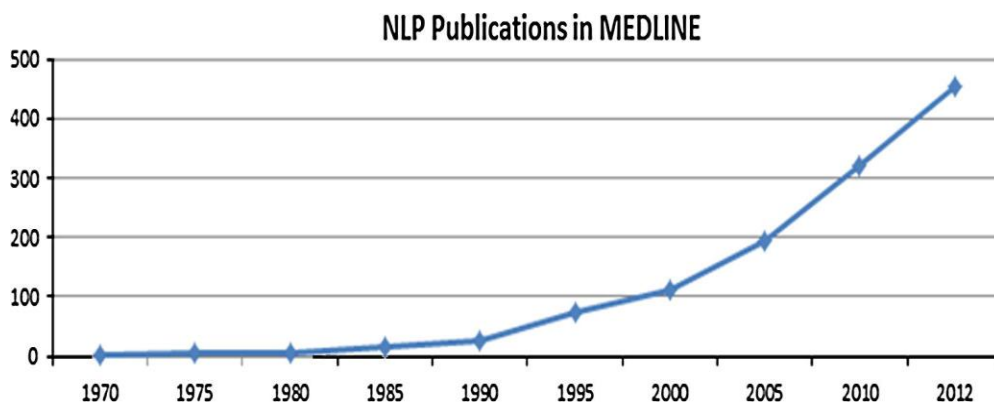


Figure 1. Number of NLP publications in PubMed (from Carol Friedman, et al (Carol & Milton, 2013))

## 2.2. NER in clinical text

### 2.2.1. An introduction to NER

NER (also known as entity identification, entity chunking and entity extraction) is a

subtask of information extraction that seeks to locate and classify elements in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. A common NER task requires detecting the boundary of an entity and determining the semantic category of the entity. The term “Named Entity”, now widely used in NLP, was coined at the Sixth Message Understanding Conference (MUC 6) (R. Grishman & Sundheim 1996). At that time, MUC was focusing on IE tasks where structured information of company and defense related activities was being extracted from unstructured text, such as newspaper articles. When defining the task, people noticed that it was essential to recognize information units like names, including person, organization and location names, and numeric expressions including time, date, money and percent expressions. Identifying references to these entities in text was recognized as one of the important sub-tasks of IE and was called “Named Entity Recognition and Classification (NERC)” .

The current dominant techniques for addressing the NER problem are supervised ML-based approaches. In ML-based NER approaches, annotated data are typically represented in the BIO format, in which each word is assigned to one of the three classes: B, beginning of an entity; I, inside an entity; O, outside of an entity. Therefore, the NER problem now becomes a classification problem to assign one of the three class labels to each word. ML algorithms and features are two most important factors that affect the performance of ML-based NER systems. (E. Tjong Kim Sang & De Meulder 2003). Different supervised machine learning algorithms have been applied to NER, including

Hidden Markov Models (HMM) (D. Bikel et al. 1997), Decision Trees (S. Sekine 1998), Maximum Entropy Models (ME) (A. Borthwick 1998), Support Vector Machines (SVM) (M. Asahara & Matsumoto 2003) , and Conditional Random Fields (CRF) ( A. McCallum & Li 2003). Among them, conditional random fields (CRF) (J. Lafferty & Pereira, 2001) and support vector machines (SVM) (Cortes & Vapnik, 1995) are two of the widely used algorithms. In theory, CRF is a representative sequential labeling algorithm, which is suitable for the NER problem. SVM is a robust classification algorithm that is based on large margin theory. To include information about neighbor tokens in sequences, researchers have developed methods to incorporate neighbor information into features for SVM-based NER systems (Kudoh & Matsumoto, 2000; Kudoh & Matsumoto, 2001).

Features for NER are descriptors or characteristic attributes of words designed for algorithmic consumption. Various types of features have been used in NER. For example, information about the word itself, such as upper/lower case, punctuation, numerical value, prefix and suffix, and special characters, is often useful. Dictionaries containing words with their semantic categories are often used to generate valuable features for NER as well. In addition, contextual information within a document is also helpful. More recently, word representation information generated from unsupervised analysis (B.Tang & Xu, 2012; B.Tang & Xu, 2013), has been investigated and showed beneficial improvement on NER performance.

Despite the high F1 numbers reported on the MUC-7 dataset, the problem of Named Entity Recognition is far from being solved. The main efforts are directed to reducing the annotation labor by employing semi-supervised learning (Lin & Xiaoyun, 2009), robust performance across domains (Lev, 1999; Iii HD) and scaling up to fine-grained entity types (Changki et al., 2006). For example, research indicates that even state-of-the-art NER systems are brittle, meaning that NER systems developed for one domain do not typically perform well in other domains (Poibeau, 2001). Considerable effort is involved in tuning NER systems to perform well in a new domain, which is equally applicable for both rule-based and trainable statistical systems. Abundant literature is available for NER approaches in both English and Chinese text in open domains (Asif, 2013) (Zhao, 2008; Aaron & Lidia, 2013; Rohini & Srihari, 2008).

### **2.2.2. Relevant work on clinical NER in English**

Recognition of medically relevant entities in clinical documents is obviously one of the most important tasks of clinical NLP. Compared with open domains, clinical text has its unique challenges for NER. One of them is the high ambiguity of medical terms, for example, “direct bilirubin” can refer to a substance, laboratory procedure, or result. Moreover, abbreviations are widely used in clinical text and they are often highly ambiguous, e.g., “APC” has 12 expansions, including “activated protein C” and “adenomatous polyposis coli”. Furthermore, some clinical entities can be disjoint, instead

of being continuous. For example, in the sentence “chest wall shows slight tenderness on pressure”, the concept “chest tenderness” spans across multiple disjoint phrases. NER methods that rely on shallow parsing may not be sufficient to resolve this type of disjoint entities.

Early clinical NLP systems often recognize entities using rule-based approaches that rely on dictionary resources (Aronson & Lang, 2010; C. Friedman & Johnson, 1994). More recently, ML-based NER approaches have been studied for clinical text, largely due to the increasing availability of annotated clinical corpora. For example, i2b2 (the Center of Informatics for Integrating Biology and the Bedside) at Partners Health Care System has organized a few clinical NER challenges and created annotated corpora for recognizing various clinical entities including medications and signature (the 2009 challenge) (O.Uzuner & Cadag, 2010; Murphy et al., 2010), medical problems, treatments, and laboratory tests (the 2010 i2b2 challenge) (O. Uzuner & DuVall, 2011; South et al., 2010), and temporal expressions (the 2012 i2b2 challenge) (Xu, Tsujii, & Chang, 2013; Sun & Uzuner, 2013). The top-ranked systems in the i2b2 NLP challenges were primarily based on ML approaches (O. Uzuner & DuVall, 2011; O. Uzuner & Cadag, 2010; B. de Bruijin et al., 2011; M. Jiang, & Xu, 2011). Figure 2 shows the top 10 teams who participated in the 2009 i2b2 challenge on medication information extraction. The systems ranked #1 and #10 were ML-based and the others were rule-based systems.

Rank	Group (external resources, medical experts)	Phrase-level horizontal evaluation								
		Overall			Narrative			List		
		Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
1	USyd (N,Y)	0.896	0.820	0.857	0.685	0.63	0.656	0.914	0.835	0.873
2	Vanderbilt (Y,Y)	0.840	0.803	0.821	0.571	0.606	0.588	0.901	0.814	0.855
3	Manchester (N,N)	0.864	0.766	0.812	0.692	0.542	0.608	0.858	0.805	0.831
4	NLM (N,N)	0.784	0.823	0.803	0.54	0.623	0.579	0.861	0.86	0.861
5	BME-Humboldt (N,N)	0.841	0.758	0.797	0.505	0.576	0.538	0.894	0.701	0.786
6	OpenU (N,N)	0.850	0.748	0.796	0.4	0.585	0.475	0.903	0.536	0.673
7	UParis (N,N)	0.799	0.761	0.780	0.336	0.482	0.396	0.765	0.552	0.641
8	LIMSI (N,N)	0.827	0.725	0.773	0.491	0.535	0.512	0.887	0.685	0.773
9	UofUtah (N,Y)	0.832	0.715	0.769	0.504	0.531	0.517	0.859	0.657	0.744
10	UWiscconsinM (N,N)	0.904	0.661	0.764	0.366	0.405	0.384	0.931	0.51	0.659

N and Y in parentheses indicate whether the system was declared to have used external resources or benefited from medical experts. N, No; Y, Yes.

Figure 2. Top 10 teams for 2009 i2b2 challenge (O. Uzuner & Cadag, 2010)

Table 2 shows the teams and performance in the 2010 i2b2 challenge of clinical NER and most of the top ranked systems were machine learning-based, as the organizers provided an annotated training corpus (O. Uzuner & DuVall, 2011).

Table 2 *Methods used by the top teams in the 2010 i2b2 challenge for concept extraction*

Concept extraction System by	Medical experts	Method	External?	Exact F measure	Inexact F measure
deBruijn <i>et al</i> ( <i>deBruijn et al., 2010</i> )	N	Semi-supervised	N	0.852	0.924

Jiang <i>et al</i> (Jiang <i>et al.</i> , 2010)	Y	Hybrid	Y	0.839	0.913
Kang <i>et al</i> (Kang <i>et al.</i> , 2010)	N	Hybrid	Y	0.821	0.904
Gurulingappa <i>et al</i> (Gurulingappa & Fluck, 2010)	N	Supervised	Y	0.818	0.905
Patrick <i>et al</i> (Patick <i>et al.</i> , 2010)	N	Supervised	Y	0.818	0.898
Torii and Liu (Torii, 2010)	N	Supervised	N	0.813	0.898
Jonnalagadda and Gonzalez (Jonnalagadda 2010)	N	Semi- supervised	N	0.809	0.901
Sasaki <i>et al</i> (Sasaki <i>et al.</i> , 2010)	N	Supervised	N	0.802	0.887
Roberts <i>et al</i> (Roberts & Harabagiu, 2010)	N	Supervised	N	0.796	0.893
Pai <i>et al</i> (Pai <i>et al.</i> , 2010)	Y	Hybrid	N	0.788	0.884

Note: Credit to Uzuner et al, (O. Uzuner & DuVall, 2011)

In these clinical NER studies, different types of features, including morphological (e.g., prefix and suffix), syntactic (e.g., part-of-speech tags), and semantic (e.g., semantic classes in UMLS) information of context words have been used to improve NER performance (Jiang et al., 2011). In addition, word representation information generated from unsupervised analysis such as brown clustering (B. Tang & Xu, 2012; B. Tang & Xu, 2013), has also been investigated, and showed beneficial improvement on the clinical NER task (de Bruijn et al., 2010). Among different ML algorithms, CRF (J. Lafferty & Pereira, 2001) and SVM (Cortes & Vapnik, 1995) have been widely applied to clinical NER (Kudoh & Matsumoto, 2000; Kudo & Matsumoto, 2001). Another emerging algorithm for NER is the structural SVM (SSVM) (B. Taskar & Koller, 2003; I. Tsochantaridis & Altun, 2005), which is an SVM-based discriminative algorithm for structural prediction. Therefore, SSVM combines the advantages of both CRF and SVM and is also suitable for sequence-labeling problems. Recent studies demonstrated that SSVM achieved a slightly better performance on recognizing clinical entities in discharge summaries from US hospitals (B. Tang & Xu, 2012; B. Tang & Xu, 2013; M. Jiang & Xu, 2011).

### **2.2.3. Relevant work on clinical NER in Chinese**

Although NER has been extensively studied in open domains for both English and



Chinese text (Sang & Meulder, 2003; Sang, 2002; Nadeau & Sekine, 2007), few studies have investigated NER in clinical text written in Chinese. With the rapid growth of EHRs in China, there is a huge need to extract information from clinical notes written in Chinese. Prior work of NER in Chinese clinical text has focused on traditional Chinese medicine documents. Wang et al (S.Wang & Chen, 2009) applied CRF, SVM, and ME to recognize symptoms and pathogenesis in ancient Chinese medical records and showed that CRF achieved a better performance. Wang et al (Y. Wang & Jiang, 2012) conducted a preliminary study on symptom name recognition in clinical notes of traditional Chinese medicine. Despite the valuable insights, these studies studied traditional Chinese medical records which comprise only a small proportion of Chinese EHRs and only symptom was recognized. A more recent and related study by Xu et al (Y. Xu & Chang, 2013) proposed a joint model that integrates segmentation and NER simultaneously to improve the performance of both tasks in Chinese discharge summaries. However they did not investigate the performance of other common algorithms on Chinese clinical notes and the effects of different features on NER performance. Table 3 compares different NER studies on Chinese clinical text.

During the period of our study, we found that several studies on Chinese clinical NER were being conducted simultaneously, indicating the needs of clinical NLP in China. Wang H et al. conducted a study to extract tumor-related information from operation notes of hepatic carcinomas which were written in Chinese using both rule-based and supervised ML approaches with the best approach yielding a precision of 69.6% and

recall of 58.3% (Wang et al. 2014). Wang Y et al. performed experiments to adapt general label sequencing classifiers to recognize symptoms in the chief complaints of free-text traditional Chinese medicine record and achieved very good performance with the highest *FMrec* of 95.12% (Wang, et al., 2014). Table 3 summarizes and compares the characteristics of above relevant studies of clinical NER in Chinese.

Table 3. *Comparison of NER studies in Chinese*

Authors	Method	Entities	Note type	Conclusion	Drawback
Wang S, et al (S. Wang & Chen, 2009)	CRF, ME, SVM	Symptoms and pathogenes is	Traditional Chinese Record	CRF achieved better performance	Only on traditional Chinese record which is unique and a small part; only two entities
Wang Y, et al (Y. Wang & Jiang, 2012)		Symptom	Traditional Chinese Record		Only on traditional Chinese record which is unique and small portion; only one entity
Xu Y, et al (Y. Xu & Chang, 2013)	Joint model(segmentation and NER)		Discharge summaries	Improve performance of both tasks	Focus on one fixed method; No comparison of algorithms and features
Wang H, et al (Wang et al., 2014)	Rule based and CRF	Tumor information	Operation notes	Best approach yielded 69.6% in precision, 58.3% in recall	Performance is poor

Wang Y, et al (Wang et al., 2014)	Supervised methods(CRF, MEMM, HMM)	SNR(symptom name recognition)	Chief complaints of TCM notes	highest <i>F</i> <sub>max</sub> (95.12%)	Chief complaints are short and less free, a less difficult task
-----------------------------------	------------------------------------	-------------------------------	-------------------------------	--	---

### 2.3. Summary

As shown in the previous studies, clinical NLP has been successfully used in various applications in healthcare operation and clinical research, becoming an important research area of Biomedical Informatics. Compared with NLP research in English clinical text, very limited resources and methods have been developed for clinical NLP in Chinese. NER is one of the fundamental tasks of NLP and it would be a good starting project for promoting clinical NLP research in Chinese.

## **Chapter 3: Create an annotated corpus of Chinese clinical texts**

### **3.1. Introduction to clinical corpora construction**

An annotated corpus is a collection of texts that have been enhanced with mark-ups specifying linguistic and domain information such as syntactic structure, named entity identification, and entity relationships. The impact of a common shared corpus resource has been shown by a series of important NLP related conferences such as MUC (MUC-7), TREC(conference Tr.) and SENSEVAL (SENSEVAL), to stimulate increased focus. Large annotated corpora in open domains such as newswire have been created to advance general NLP research (see the Linguistic Data Consortium - <https://www ldc.upenn.edu/>). In the medical domain, the need for annotated clinical corpora was also highlighted in a recent clinical NLP workshop hosted by the National Library of Medicine (Friedman, Rindflesch, & Corn, 2013). In the past decade, a number of clinical corpora have been developed for various NLP tasks, including POS tagging (Fan et al., 2011; Pkahomov, Coden, & Chute, 2006), named entity recognition (Uzuner, Solti, Xia, & Cadag, 2010; Uzuner, South, Shen, & DuVall, 2010), word sense disambiguation (Moom et al., 2014), syntactic parsing (Fan et al., 2013; Cairns et al., 2011), temporal reasoning (Sun, Rumshisky, & Uzuner, 2013), co-reference resolution (Savova, Chapman, Zheng, & Crowley, 2011; Uzuner et al., 2012) and concept encoding (Suominen et al., 2013). Among them, the MiPACQ corpus is a relatively large clinical text collection with multiple layers of annotations, including POS, parse tree, semantic role labeling, named

entity recognition, and concept encoding (Cairns et al., 2011). Another significant annotation effort on clinical text is from the i2b2 clinical NLP challenges (Uzuner, Solti, & Cadag, 2010; Uzuner, Solti, Xia, & Cadag, 2010; Uzuner, South, Shen, & DuVall, 2011; Sun, Rumshisky, & Uzuner, 2013; Uzuner et al., 2012; Uzuner, Luo, & Szolovits, 2007; Uzuner, 2008; Uzuner, 2009; Sun, Rumshisky, & Uzuner, 2012). More specifically, a number of i2b2 clinical NLP challenges have focused on various NER tasks in clinical text, including recognizing identifiers (Uzuner, Luo, & Szolovits, 2007), medications (Uzuner, Solti, & Cadag, 2010; Uzuner, Solti, Xia, & Cadag, 2010), clinical problems, treatments, and tests (Uzuner, South, Shen, & DuVall, 2011), and temporal expressions (Sun, Rumshisky, & Uzuner, 2013; Sun, Rumshisky, & Uzuner, 2012; Sun, Rumshisky, & Uzuner, 2013).

Creating high-quality annotated corpora is not a trivial task. A good annotation schema (or guideline) is critical in the annotation process. One challenge for annotating clinical text is to balance the linguistic and medical domain knowledge (Chapman & Dowling, 2006). Chapman and Dowling proposed an iterative process to induce an annotation schema for emergency department reports and showed promising results ((Chapman & Dowling, 2006). Researchers also demonstrated the effectiveness of this approach on other clinical NLP tasks such as parse tree annotation. Other factors such as background of annotators and training of annotators also significantly affects the annotation quality ((Chapman, Dowling, & Hripcsak, 2008).

In addition, graphic interfaces are often used in the annotation process and are very important for improving the efficiency of annotation. Table 4 shows a list of common annotation tools that have been used in clinical text annotation. It also compares different features of these tools. Brat (available at <http://brat.nlplab.org/index.html>) is a web-based tool for text annotation and is designed in particular for structured annotation. GATE (Cunningham, Maynard, & Bontcheva, 2011) is in active use for all types of computational tasks involving human language. GATE includes a desktop client for developers, a workflow-based web application, a Java library, an architecture and a process. Knowtator (Ogren, 2006) is a general-purpose text annotation tool that is integrated with the Protégé knowledge representation system. It facilitates the manual creation of training and evaluation corpora for a variety of biomedical language processing tasks. MAE (Multi-purpose Annotation Environment) (Stubbs, 2011) is an annotation tool created by Amber Stubbs (<http://amberstubbs.net>) at Brandeis University and was used in the i2b2 challenges. Teamware (Bontcheva et al., 2013) is another web-based management platform for collaborative annotation & curation. It is a cost-effective environment for annotation and curation projects, enabling a broadly distributed workforce with the capability to monitor progress and results remotely in real time.

Table 4. *A list of available annotation tools*

	<b>BRAT</b>	<b>GATE</b>	<b>KNOWTATOR</b>	<b>MAE</b>	<b>Teamware</b>
Web based?	Yes	No	No	No	Yes

Multiple users?	No	No	No	No	Yes
Team Support?	No	No	Yes	No	Yes
Prerequisite	Python	Java	Java, Protege	Java	Java/Tomcat/Mysql
Easy to install?	Yes	Yes	No	Yes	No
Calculate Inter annotator agreement?	No	Has plugins for IAA	Yes	No	Yes
Allow fast-annotation?	No	Yes	Yes	No	Same as GATE

As mentioned in the previous chapters, few studies have investigated the NER tasks in Chinese clinical text and very limited annotated corpora of Chinese clinical text exist. Therefore, in this study, we propose to develop an annotation scheme for NER in Chinese clinical text and create annotated corpora of admission notes and discharge summaries in Chinese, thus to advance the NER study in Chinese clinical text. To the best of our knowledge, this is one of the first efforts at developing an annotation guideline and large scale annotated clinical corpora in Chinese.

### 3.2 Methods

### **3.2.1. Data sets**

We collected one-month of admission + discharge summaries from the EMR database of Peking Union Medical College Hospital (PUMCH) in China. PUMCH is one of the most prestigious top tier hospitals in China. Founded by the Rockefeller Foundation and staffed by U.S physicians in its early year in the 1910s, this hospital has a long tradition of following rigorous clinical training. Therefore, it has maintained the best quality of practice and Health Record is one of its three best assets in the nation. A typical health record in PUMCH comprises the admission note, first progress note, progress note and the discharge note. We chose only the admission and discharge summary from its electronic medical records. The rationale for extracting and annotating only admission + discharge summaries are the following: first, admission and discharge summaries are two important components of medical record for an inpatient and second, these two parts contain a wealth of clinical concepts. After excluding very short notes (incomplete notes), we randomly selected 400 admission +discharge summaries from the PUMCH pool for annotation. All patient identifiers in the notes were manually removed by PUMCH physicians before the notes were sent to researchers for annotation.

### **3.2.2. Development of annotation guideline for Chinese clinical texts**

The annotation guidelines were similar to those used in the 2010 i2b2 NLP challenge (<https://www.i2b2.org/NLP/Relations/Documentation.php>), but were translated into Chinese. One main difference is that we broke the “treatment” category in the i2b2 challenge into two categories: “procedures” and “medications”. Thus, we had four types



of entities in this study (medical problems, tests, medications, and procedures), instead of three as in the i2b2 challenge. In addition, we also specified some rules for determining entity boundary in Chinese text. The Chinese concept annotation guideline is attached as Appendix I.

The annotation schema was developed using inductive approaches similar to that utilized in Chapman and Dowling (Chapman & Dowling, 2006). We started with the i2b2 guidelines and iterative modifications were made by testing more reports. Two Chinese MD students iteratively annotated a batch of clinical notes (i.e., 5) to identify potential issues and made changes to the Chinese concept annotation guideline as necessary. Many examples were included in the guidelines to train the annotators. One difficulty of Chinese clinical concept annotation is the determination of the boundary of an expression.

### **3.2.3. Development of the annotation tool for Chinese clinical text**

To effectively perform annotation and create annotated corpus, we developed an annotation tool for Chinese clinical texts. The tool uses MySQL database, Java programming language based MyEclipse 6.5 and runs under webserver Tomcat 5.5. The tool manages the whole process of annotation including user management, uploading annotation files, assignment of annotation jobs, annotation, and output of the results. The tool took the B/S architecture so that multiple annotators can work simultaneously online in different places. The design and major functions of the tool are attached as appendix II. Following is one of the screenshots of the annotation tool.



Figure 3. Screenshot of the annotation tool

### 3.2.4. Conducting of annotation

Using the developed Chinese annotation guidelines, two native Chinese-speaking domain experts (MD students) manually annotated the problems, tests, procedures and medications in each note. To calculate the inter-rater agreement for annotation, 40 identical notes were presented to the two annotators. Therefore, each annotator completed 180 different notes and 40 identical notes.

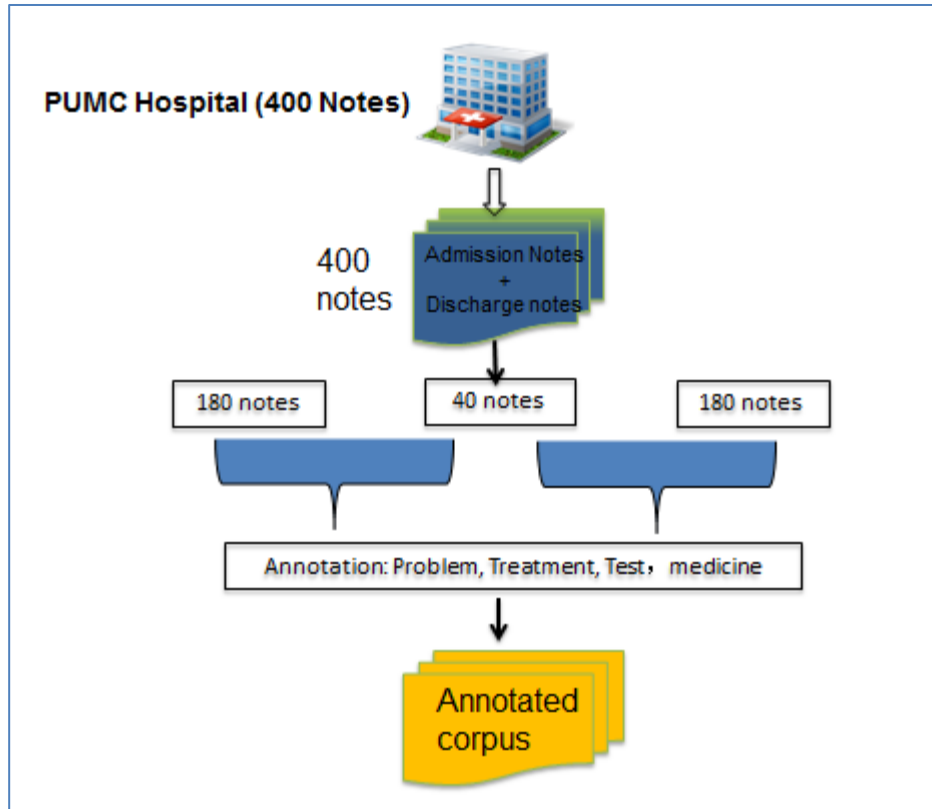


Figure 4. Annotation workflow

### 3.4. Results

#### 3.4.1. Corpus statistics

One of the largest known annotated Chinese clinical text corpus has been created in this study. Table 5 shows the statistics of the corpora of Chinese discharge summaries and admission notes we obtained from the annotation process that will be used in the NER study. There are 30,793 sentences and 38,973 entities in the 400 admission notes, and 22,838 sentences and 39,334 entities in the 400 discharge summaries. The proportion of

each type of concept in the 800 notes (both admission and discharge summaries) is 56.95% for problem, 29.73% for lab test, 8.54% for procedure and 4.78% for medicine, respectively. The problems and lab tests are almost equally distributed in admission and discharge summaries, while procedures and medicines are mainly found in discharge summaries.

Table 5. *Summary statistics of annotated datasets of Chinese discharge summaries and admission notes.*

Type	# note	# sen	# char	# NER				
				# problem	# procedure	# labtest	# medicine	# total
Admission	400	30,793	417,586	24,433	2,171	11,168	1,201	38,973
Discharge	400	22,838	368,404	20,159	4,517	12,114	2,544	39,334
All	800	53,631	785,990	44,592	6,688	23,282	3,745	78,307

### 3.4.2. Quality of the corpora

Based on the annotations on the 40 notes, the inter-annotation agreements using kappa statistics (Hripcsak & Rothschild, 2005) on admission and discharge summaries were 0.957 and 0.924, respectively, which indicates that the annotation was reliable.

## 3.5. Discussion

In this study, we developed an annotation schema, a web-based annotation tool, and two annotated clinical corpora (admission notes and discharge summaries) for clinical NER in Chinese. To the best of our knowledge, this is one of the first efforts at generating annotated clinical corpora in Chinese. During the development of this project, parallel efforts on clinical NER in Chinese were also reported (Y. Xu & Chang, 2013).

Developing an annotation schema that different annotators can agree on is not trivial. One issue of annotation entities in Chinese text is about boundary determination. For example, in the case of problem boundary identification, if a problem is followed by a change in the problem, the “change” is not included in the problem, as shown in the this text: “左下胸腔积液+增加” (“lower left pleural effusion”+ “increased”), “increased” is not included in the problem. However if it is a body location/function followed by a modifier, the modifier is included in the problem: “心脏+扩大” (“heart”+“enlarged”) , “enlarged” is included in the problem. Although we reported a high IAA between the two annotators, the initial agreement between the annotators while developing the guidelines was low. The iterative process for guideline development and the extensive training of the annotators were two important factors that contributed maximally to the final high IAA.

To advance clinical NLP in Chinese, more large annotated corpora should be developed. However, annotation is a labor-intensive and time-consuming process. Thus, methods that can efficiently reduce annotation cost would be highly desirable. We plan to

investigate such methods in our future work. One possibility is to use machine assisted pre-annotation, which has been reported by a few studies and has shown promising results (Lingren et al., 2014; South et al., 2014; Gobbel et al., 2014). Other methods, such as active learning (Chen, Mani, & Xu, 2012; Figueroa et al., 2012; Chen et al., 2013) would be interesting to investigate as well.

## **Chapter 4: Compare entity distribution between Chinese and English clinical documents**

### **4.1. Introduction**

Recently, the Chinese government announced ambitious national health reform plans. It has allocated enormous funds to improve the health care system in China. For example, a recent report indicated that health insurance now covers 95.6% of the population in China (Lim, 2004). The latter may be one of the greatest healthcare accomplishments worldwide. Health information technology (HIT) stands as one of the eight supporting pillars necessary to achieve Chinese healthcare reform goals. The Chinese government views Electronic Medical Record systems (EHRs) as an essential component for modern hospital management, with the potential to improve the efficiency, quality, and safety of health care. The Chinese Ministry of Health (MOH) has established a standards bureau that in 2009 proposed a series of HIT standards such as those covering EHR basic architecture and data standards (EMR basic architecture and data standards, 2009). Up to now, many urban hospitals in China have adopted and used EHR systems to a variable extent (Report of 2011-2015 Market Survey and prediction of development of China's EMR , 2011). To accelerate EHR adoption in rural hospitals, the Chinese government has allocated 3.9 billion RMB (approximately \$600 million US) in 2011 to a pilot program for implementing EHRs in about 200 hospitals (Chinese MOH Notice on First 97 trial hospitals for EMR., 2011; Chinese MOH Notice on Second 92 trial hospitals for EMR ,

2011). Given the large population of China, the rapid growth in standardized EHR databases is anticipated to accumulate unprecedented amounts of electronically available clinical data that can support clinical and translational research.

In the US, large academic medical centers have implemented EHR systems for more than three decades and have established large practice-based longitudinal datasets (Roden et al., 2008). Recently, the growth of EHRs in the US is being fueled by federal legislation that provides generous financial incentives to institutions demonstrating aggressive application and “meaningful use” of comprehensive EHRs (Shea & Hripcsak, 2010; Health USD-po, 2009). Major efforts are already underway to link these EHRs across the US institutions for clinical and translational research. The US EHR databases have been successfully used for various types of studies such as observational comparative effectiveness research (Pace et al., 2009), genomic (McCarty, et al., 2011), and pharmacogenomic studies (Xu et al., 2011).

Of late, there has been an increasing trend in the US and western institutions towards collaborating with China on public health, clinical, and translational research based on EHRs (Wang, 2011; Jiebai et al., 2012). It is very likely that patient records stored in the EHR systems in China will also become an invaluable asset supporting international collaborative research endeavors. Due to the differences in culture and practice patterns between China and US, EHR data in Chinese hospitals is likely to have different



characteristics than data from US institutions. It is important for international collaborators to understand any differences that might exist. Nevertheless, few published studies have compared available EHR data in China versus those in the US. Various EHR systems can contain data in numerous formats, including both structured and unstructured information. For example, EHR systems typically store narrative clinical reports, containing detailed treatment and outcome information for individual patients. Such reports comprise a highly valuable resources for clinical research.

As an initial step towards differentiating EHR data in Chinese and US systems, this study attempts to understand differences that exist between Chinese and English clinical documents. More specifically, the study collected 1046 inpatient discharge summaries from one Chinese (400 notes) and three US institutions (646 notes). The investigators manually analyzed the three major types of clinical entities: medical problems, tests, and treatments in all discharge summaries and reported comparison results at both the document and section levels. Documenting and understanding differences in clinical reports from the US and Chinese EHRs are important for cross-country collaborations and our study also provides valuable insights for developing natural language processing tools for Chinese clinical text.

## **4.2. Methods**

### **4.2.1 Data sets**

Organizers of the 2010 i2b2 clinical NLP challenge (I2b2 2012 annotation guidelines, 2012), collected 826 clinical notes, of which 646 are inpatient discharge summaries, from three US hospitals: University of Pittsburgh Medical Center (UPMC), Partners Healthcare (PARTNERS), and Beth Israel Deaconess Medical Center (BETH). For each clinical note in the collection, domain experts manually annotated three clinically important components: medical problems (e.g., diseases and symptoms), tests (e.g., lab tests), and treatments (e.g., medications and procedures), by following annotation guidelines developed by the i2b2 challenge organizers (O. Uzunur & DuVall, 2010). This study included and analyzed all 646 English discharge summaries from the 2010 i2b2 challenge. We chose 400 annotated discharge summaries in Chinese (as described in Chapter 3) and compared it with the 646 discharge summaries in English in this study.

### **4.2.2. Analytical Methods**

We conducted content analysis on the 646 English and 400 Chinese discharge summaries using Charmaz's grounded theory approach (Charmaz, 2006). We approached the data with no prior assumptions and generated descriptive statistics based on the content of the notes. We analyzed the data with a focus on understanding the distributions of three types of important clinical entities (Problems, Tests, and Treatments) at both document and

section levels, as well as the differences of such distributions between Chinese and English clinical text.

#### **4.2.2.1. Document level analysis**

At the document level, we conducted two experiments: (1) compared the vocabulary distribution and the density of clinical entities (defined as the average number of clinical entities in each document) in the Chinese and English corpora; and (2) reported relative frequency of the three types of entities for each institution. Zipf's distribution is widely used to describe the vocabulary frequency by plotting a log-scale graph between frequency and rank. We collected all the words from the two corpora and then ranked the words according to their frequencies to present the curve in log scale. As there are no spaces between words in the Chinese corpus, the Stanford Word Segmenter (Pi-Chuan & Christopher, 2008) trained on Penn Chinese Treebank corpus (Naiwen & Marta, 2005) was used to identify individual Chinese words. In the second experiment, the relative frequency for a specific entity type was defined as the number of entities that belong to this type divided by the total number of all three types of entities. We calculated the relative frequencies of three different entity types: Problem, Test, and Treatment for all four institutions.

#### **4.2.2.2. Section level analysis**

At the section level, we focused on measuring the density of clinical entities in each

section across the four institutions. Section identification in clinical text is not a trivial task (Denny et al., 2009) . In this study, we developed an ad-hoc approach to identify sections in Chinese and English notes.

- Detect candidate section headers -- a regular expression based program was developed to detect all candidate section headers using the colon, upper case letter and other features.
- Group section headers -- we manually reviewed all candidate section headers, removed false positives, and grouped all the variations according to the contents under section header.
- Match section headers -- two domain experts (author WW -- who is familiar with both Chinese clinical notes and English clinical notes, JD -- a domain expert in English clinical notes) worked together to match the corresponding section headers between the English corpus and the Chinese corpus according to the content in each section.

Once sections were identified and mapped, we reported the average number of clinical entities for each section. To further understand the differences in section content, we also compared the average number of entities within each section in both the English and Chinese corpora.

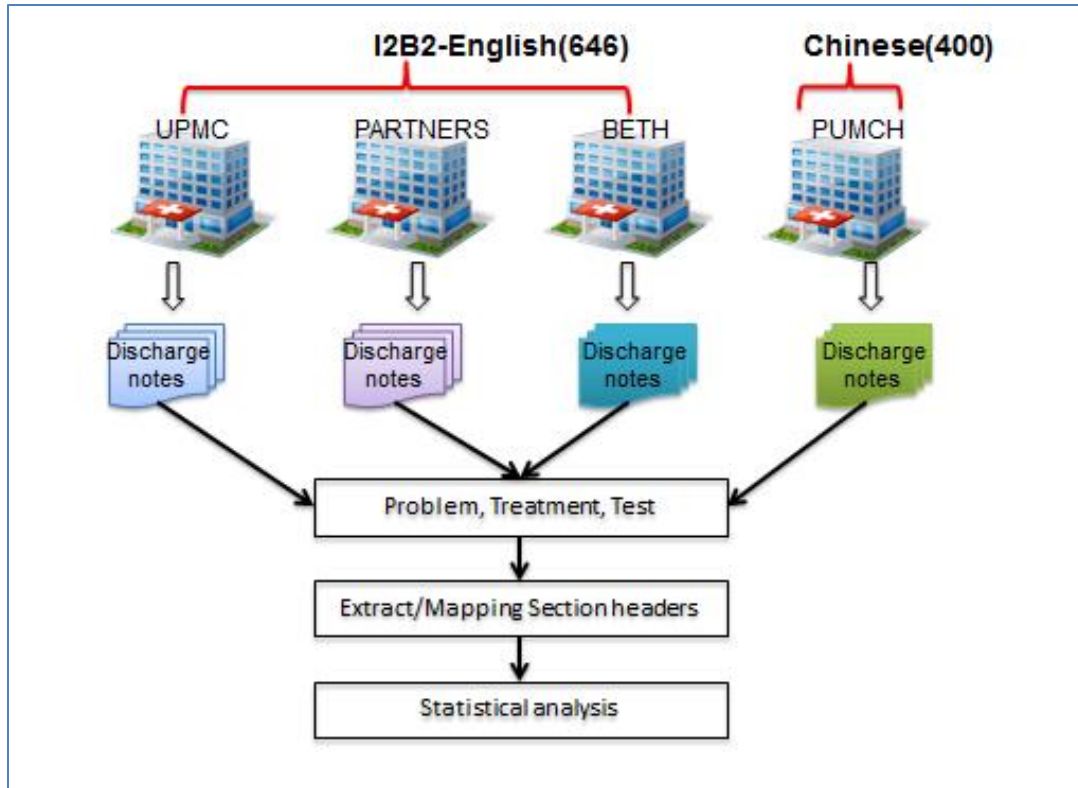


Figure 5. Workflow of the entity distribution comparison study

### 4.3. Results

Figure 6 draws the word frequency distribution for the English corpus and Chinese corpus, the curve shows a typical distribution of Zipf's law. As the English corpus contains more notes than Chinese corpus, the curve for English is above the Chinese corpus (labelled as PUMCH). Figure 7 shows the normalized distribution of entities in the English corpus and the Chinese corpus. The curve for the English corpus descended smoothly, whereas, the curve for PUMCH ended with a sharp decrease, indicating that the English corpus appeared to use a more diverse vocabulary; however, such analysis is complicated by the differences in word form variation between the two languages.

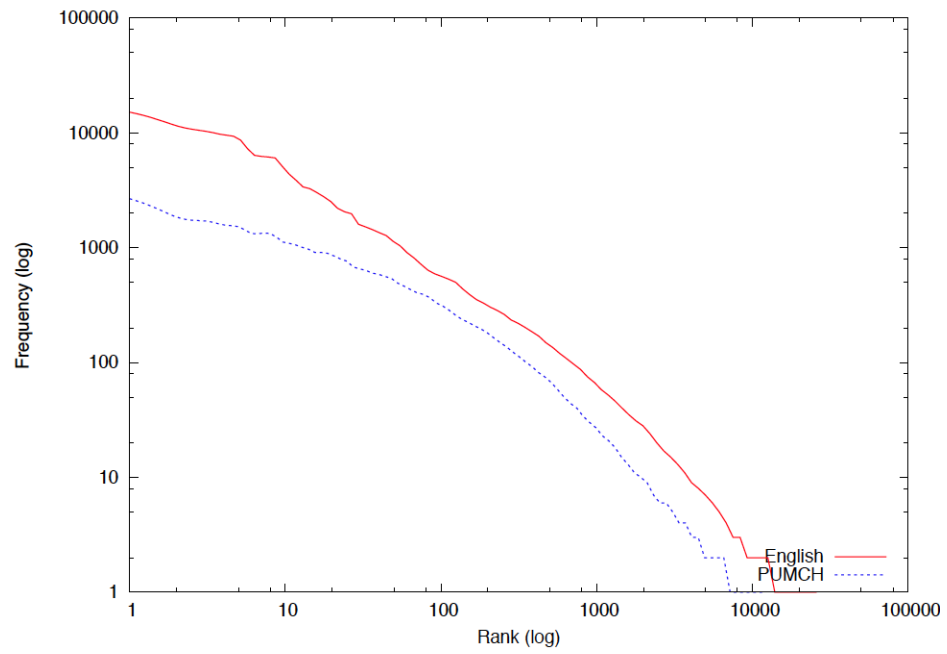


Figure 6. Zipf's distribution of vocabularies

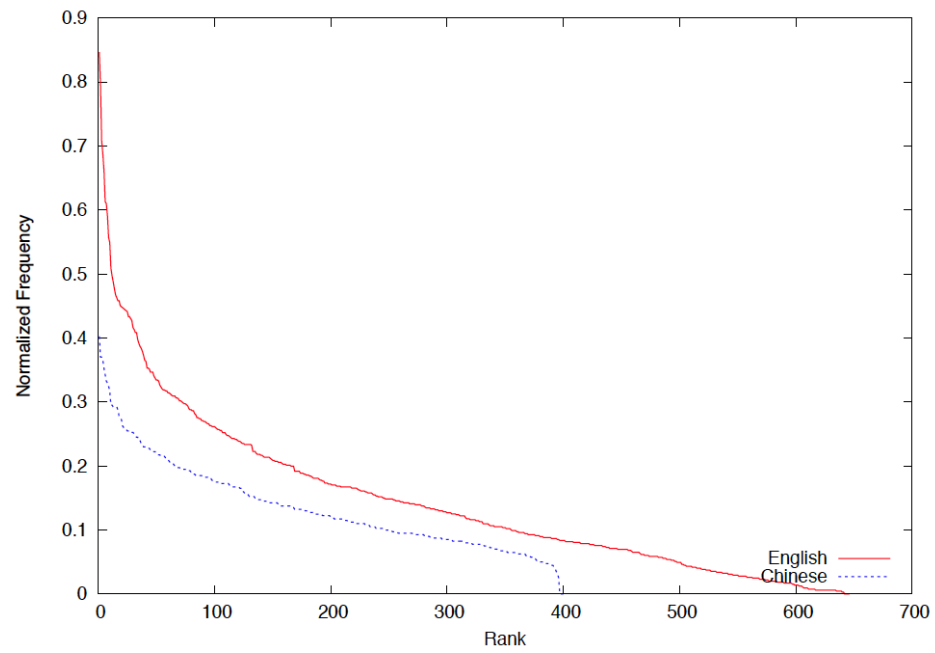


Figure 7. Normalized distribution of annotated entities

Table 6 shows the average numbers of different types of entities in one note, across four different institutions. Compared with the three US institutions, the PUMCH corpus had fewer entities than the English corpora, especially for ‘Test’ and ‘Treatment’ entities. The relative frequencies of the three types of entities within each individual institute are shown in Figure 10. The relative frequencies are different among the four institutions, with the unique traits of PUMCH compared to the three English institutions more obvious. PUMCH had a higher proportion of ‘Problem’ entities and fewer ‘Test’ and ‘Treatment’ entities than in English institutions.

Table 6. *Distribution of different types of entities*

Corpus	# of Doc	Type	# of Entity	Average # of entity per note	Relative Frequency
UPMC (English)	220	Prob	5805	26.39	43.76%
		Test	2762	12.55	20.82%
		Treat	4700	21.36	35.43%
		All	13267	<b>60.30</b>	--
PARTNERS (English)	235	Prob	8542	36.35	44.69%
		Test	4884	20.78	25.55%
		Treat	5686	24.20	29.75%
		All	19112	<b>81.33</b>	--
BETH (English)	191	Prob	11122	58.23	38.93%
		Test	8947	46.84	31.32%
		Treat	8499	44.50	29.75%
		All	28568	<b>149.57</b>	--
PUMCH (Chinese)	400	Prob	12550	31.38	57.77%
		Test	3890	9.95	17.91%
		Treat	5284	13.12	24.32%
		All	21724	54.45	--

\*Prob -- Problems, Test – Tests, Treat -- Treatments

After grouping the variations and matching the section headers between the Chinese corpus and the English corpus, two domain experts detected that 9 common, high-level sections appeared in both English and Chinese corpora. These 9 common sections appeared in at least 10 notes in every institution. Table 7 shows the counts of documents, the counts of clinical entities, and the density of entities (the average number of entities in a given section), for the 9 sections across four institutions. Figure 9 gives a visualized view of the same data in Table 7. The results show that the density of entities is markedly different between PUMCH corpus and the English corpora, where the minimum density in the English corpora is at least twice of PUMCH corpus in the following five sections: PS, DM, DI, PE, and PMH.

Table 7. Entity density within 9 common sections across four institutions

	UPMCD (English)			PARTNERS (English)			BETH (English)			PUMCH (Chinese)		
Section	Doc	Entity	Density	Doc	Entity	Ave	Doc	Entity	Ave	Doc	Entity	Ave
<b>PS</b>	131	4453	<b>33.99</b>	174	5259	<b>30.22</b>	151	6211	<b>41.13</b>	389	383	<b><u>0.98</u></b>
<b>DM</b>	95	1224	<b>12.88</b>	138	1113	<b>8.07</b>	123	1418	<b>11.53</b>	197	494	<b><u>2.51</u></b>
<b>DI</b>	47	314	<b>6.68</b>	54	271	<b>5.02</b>	100	713	<b>7.13</b>	167	271	<b><u>1.62</u></b>
CC	33	377	<b>11.42</b>	34	67	<b>1.97</b>	77	127	<b>1.65</b>	399	770	<b>1.93</b>
DD	105	1005	<b>9.57</b>	35	126	<b>3.60</b>	136	793	<b>5.83</b>	389	2259	<b>5.81</b>
HOPI	30	486	<b>16.20</b>	151	3481	<b>23.05</b>	159	4612	<b>29.01</b>	400	7649	<b>19.12</b>
<b>PE</b>	25	479	<b>19.16</b>	142	2489	<b>17.53</b>	157	3039	<b>19.36</b>	266	1176	<b><u>4.42</u></b>
<b>PMH</b>	59	659	<b>11.17</b>	140	2209	<b>15.78</b>	166	3812	<b>22.96</b>	222	1153	<b><u>5.19</u></b>
PL	48	187	<b>3.90</b>	41	237	<b>5.78</b>	35	699	<b>19.9</b>	400	1843	<b>4.61</b>



									7			
--	--	--	--	--	--	--	--	--	---	--	--	--

*Note.* PS-patient summary, DM-discharge medications, DI- discharge instructions, CC- chief complaint, DD-discharge diagnosis, HOPI- history of present illness, PE- physical examination, PMH-past medical history, PL- problem list

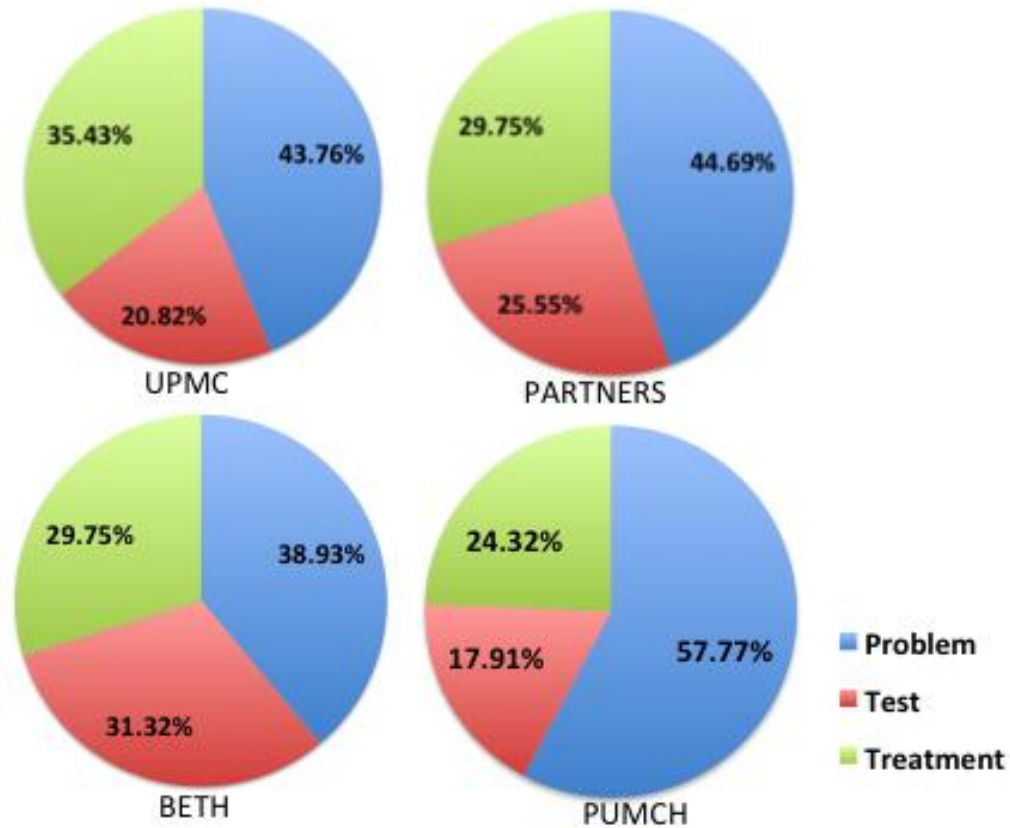


Figure 8. Relative frequency of Problems, Tests, and Treatments in three English institutions: UPMC, PARTNERS, and BETH, and one Chinese institution: PUMCH

#### 4.4. Discussion

This study compared the distribution of three types of important clinical entities (problems, tests, and treatments) in inpatient discharge summaries among three US institutions and one Chinese institution. Understanding such structural differences may help to maximize the value of EHR data acquired in Chinese hospitals when the data are

utilized for secondary purposes such as international collaborations on clinical, translational, and global health research. These structural differences in clinical documentation may also reflect more fundamental system and cultural differences in patient care delivery in China vs. that in US. This knowledge can be critical to the success of collaborative research efforts between the two countries, and between China and other western countries more broadly.

The study revealed some interesting data and discrepancies. First, the average number of clinical entities per document varied widely among different institutions, even for the three US institutions (e.g., 60.30 for UPMC vs. 149.57 for BETH). Further investigation should examine potential explanations for this variability – for example, the effects of clinical documentation methods at different institutions (e.g., directly typed vs. dictated and transcribed notes). Of note, the Chinese discharge summaries contained fewer clinical entities than any of the US institution's discharge summaries. Again, further investigation should determine why this difference exists, e.g., whether physician workloads varied between settings. Similarly, more research should evaluate why Chinese discharge summaries had a much lower relative frequency for Test entities than that of US discharge summaries. Whether it indicates that less lab tests are ordered in clinical practices in China is not certain; but it is interesting and worth conducting further investigation. Other potential causes for the greater content in US include 1) billing requirements and 2) a more complex US medico-legal environment in which more

thorough testing and discussion of problems may be performed in order to provide defence against a perceived higher risk of litigation.

When analysing clinical term distributions within different document sections, we noticed that some frequent sections in English discharge summaries, such as “Current Medications” and “Social History”, were not found in Chinese notes. Manual review by a Chinese physician showed that this information could be scattered among different sections. For example, medication information could be recorded in a patient’s Past Medical History section, e.g., "the patient was diagnosed with HTN in 1995. She is taking a beta blocker (Metoprolol) and her BP is normal”. This may also explain the differences between US and Chinese notes in entity frequency distribution for a given section (Figure 9). Chinese physicians in the team thought this was an important finding, as it provides valuable information about how to re-organize the structure of Chinese clinical notes for better representation and communication of patient information.

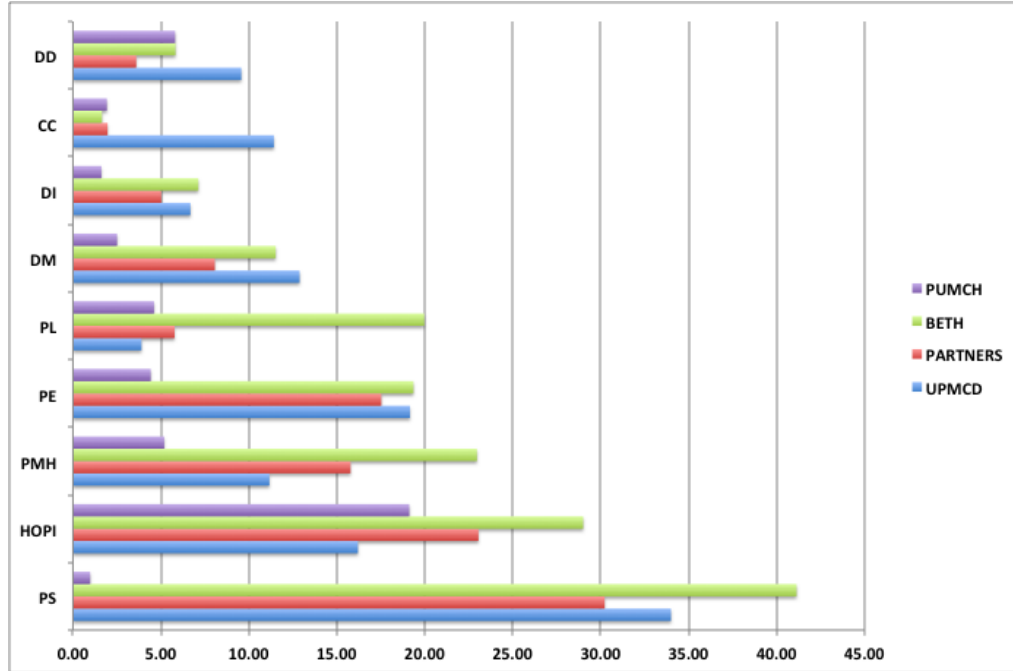


Figure 9. Visualization of entity density within 9 common sections across four institutions

One of the challenges of using EHR data for medical research, which exists for both US and Chinese EHRs, is that much of the detailed clinical information is embedded in the narrative clinical reports, which are not directly usable for analysis. Much effort has been devoted to develop natural language processing (NLP) technologies for English clinical text (Aronson, 2001; Friedman et al., 1994; Savova et al., 2010) and some approaches have shown promising results (Meystre, Savova, Kipper-Schuler, & Hurdle, 2008). However, little work has been done on NLP regarding Chinese clinical text in EHRs. This study also provides potential insights relevant to the development of NLP tools for Chinese clinical text. During the vocabulary distribution analysis (Figure 6), we explored the word segmentation methods for Chinese clinical corpus. Different from English,

Chinese texts do not have spaces between words, which makes it more difficult for identifying word boundaries. Our initial analysis showed that clinical dictionary resources helped in word segmentation of Chinese clinical text. In addition, the section analysis of Chinese clinical text is also helpful for NLP research. Further studies on Chinese clinical text processing are one of the areas earmarked for future work.

This study has limitations. One of the major limitations was that the analysis of Chinese clinical text was conducted on notes from one institution in China only. Therefore the results regarding Chinese notes might not be representative. Future studies should include Chinese clinical notes from multiple institutions in China. Another limitation was that we were unable to compare the clinical settings among the i2b2notes due to the lack of intimate knowledge of these healthcare systems.

Documenting and understanding system/cultural differences in EHR documents from the US and China are important. These differences may reflect fundamental discrepancies in patient care delivery, and the different structures of healthcare systems. Mastering the differences will be critical in helping US/western researchers understand how to properly interpret and computationally reuse clinical documents produced in either healthcare system relative to the other. In addition, such learning may also inform opportunities to develop novel NLP tools for processing narrative documents in Chinese, or fine-tune tools that were originally developed in the English context.

## **Chapter 5: Develop and evaluate machine learning based NER approaches for Chinese clinical text**

### **5.1. Introduction**

Despite the important contributions of previous studies on Chinese clinical NER, none has systematically evaluated the effects of different features and different ML algorithms on NER in Chinese clinical text. It is important to investigate whether the NER methods that we have developed for English clinical text are also effective with Chinese clinical text. For example, one major difference between English and Chinese is that segmentation of Chinese text is more difficult because you cannot rely on white spaces to separate words.

The goal of this study is to assess the performance of ML-based NER approaches that have been developed for English clinical text on Chinese clinical documents. Using manually annotated datasets of admission notes and discharge summaries in Chinese, we systematically evaluated different types of feature (e.g., syntactic, semantic, and segmentation information) and four ML algorithms, CRF, SVM, SSVM, and ME. To the best of our knowledge, this is one of the earliest comprehensive studies in Chinese

clinical NER research and we believe it will provide valuable insights into NLP research in Chinese clinical text.

## **5.2. Methods**

We randomly selected 400 admission notes and 400 discharge summaries from Peking Union Medical College Hospital (PUMCH) in China. For each note, four types of entities including clinical problems, procedures, labs, and medications were annotated according to a predefined guideline. Two-thirds of the 400 notes were used for training the NER systems and one-third were used for testing. We investigated the effects of different types of features including bag-of-characters, word segmentation, part-of-speech, and section information, and different machine learning algorithms including Conditional Random Fields (CRF), Support Vector Machines (SVM), Maximum Entropy (ME), and Structural Support Vector Machines (SSVM) on the Chinese clinical NER task. All classifiers were trained on the training dataset, evaluated on the test set, and micro-averaged precision, recall, and F-measure were reported.

### **5.2.1. Datasets and annotation**

The process of corpora development was described in Chapter 3. Both annotated 400 admission notes and 400 discharge notes were used in this study.

### **5.2.2. ML-based NER**

To convert the NER task into an ML-based classification problem, we used the ‘BIO’ tags to represent the boundaries of entities. As we have four types of entity in this study,

we generated nine different tags in total: B-problem, B-procedure, B-test, B-medication, I-problem, I-procedure, I-test, I-medication, and O. Figure 10 shows examples of annotated entities labelled with BIO tags.

Sentence	BIO Tags
约1周前因受凉“感冒”后出现咳嗽，咳少量白痰。 (About one week ago, developed cough with small amount of white sputum, after catching a cold.)	约/O 1/O 周/O 前/O 因/O 受/O 凉/O “/O 感/B-problem 冒/I-problem ”/O 后/O 出/O 现/O 咳/B-problem 嗽/I-problem , 咳/B-problem 少/I-problem 量/I-problem 白/I-problem 痰/I-problem ./O

Figure 10. Examples of Chinese medical named entity recognition (NER) representation

### 5.2.2.1. Features

As shown in Figure 11, we used four types of feature: (1) bag-of-characters; (2) bag-of-words (based on two segmentation approaches); (3) part-of-speech (POS) tags (only available for one segmentation approach); and (4) section information. One major difference between Chinese and English text is that words in Chinese are formed by continuous Chinese characters without any space. For example, the word ‘cough’ (咳嗽) is formed by two Chinese characters: “咳 ” and “嗽 ”. Figure 12 shows the output of a sentence after segmentation. The bag-of-characters approach simply used individual Chinese characters as features. If word segmentation is applied to Chinese text, we can use identified word segments as features (called ‘bag-of-words’ here). We implemented two word segmentation methods in this study: (1) the Stanford Word Segmenter (<http://nlp.stanford.edu/software/segmenter.shtml>), which supports general Chinese text, but not clinical Chinese text; and (2) a simple dictionary lookup approach, which uses the



forward maximum match algorithm to search the New dictionary of medicine and drugs, a clinical dictionary containing 704,896 medical concepts in Chinese. When the Stanford Word Segmenter was used, POS tags were generated by the system as well, which were also used as features in this study. In addition, we manually reviewed some notes and defined 35 different section headers (e.g., “history of illness”) as additional features.

Feature type	Explanation
Bag-of-characters	Individual Chinese characters in a window
Bag-of-words	Individual Chinese words in a window. Two methods were used for word segmentation: the Stanford Word Segmenter and a dictionary lookup program.
Part-of-speech (POS)	POS tags, only available from the Stanford Word Segmenter
Section information	Section headers from a predefined list

Figure 11. Features used for Chinese medical entity recognition

Original Text: 约1周前因受凉“感冒”后出现咳嗽，咳少量白痰。  
Word Segmentation: 约/1/周/前/因/受凉/“/感冒/”/后/出现/咳嗽/，/咳/少量/白/痰/。

Figure 12. An example of word segmentation in Chinese

#### 5.2.2.2. Machine learning algorithms

NER problems can be considered as either a pure classification problem or a sequence labelling problem. In this study, we compared four state-of-the-art ML algorithms: two for classification problems (SVM (Cortes & Vapnik, 1995) and ME (Miller, 1998) and two for sequence labelling problems (CRF (J. Lafferty & Pereira, 2001) and SSVM (B. Taskar & Koller, 2003; I. Tsochantaridis & Altun, 2005)). SVM and SSVM are

discriminative statistical algorithms based on large margin theory, while ME and CRF are discriminative statistical algorithms based on probability theory. All of them have been widely used in NLP.

Assume there is a sequence labeling problem of independent and identically distributed training samples  $S^L = \{(x^k, y^k) | k = 1, \dots, N\}$ . We use  $l(x)$  to denote the sequence length of input  $x$ ,  $x_i^k$  denotes the  $i$ -th subinput of  $x$ ,  $y_i^k$  denotes the  $i$ -th sublabel of  $y^k$ , and  $L$  denotes the sub-label set respectively. This problem can be treated as a classification problem of training samples  $S^C = \{(x_i^k, y_i^k) | (x^k, y^k) \in S^L \text{ and } i = 0, \dots, l(x^k)\}$  if we assume all sub-labels are independent from each other.

SVM uses a linear discriminative function to model the score of an observation  $x_i^k$  and a random variable  $y_i^k$ :  $s(x_i^k, y_i^k) = wf(x_i^k, y_i^k)$ , where  $f(x_i^k, y_i^k)$  are features. The total loss function on the training samples  $S^C$  can be written as:

$$\text{loss}(S^C) = \sum_{k=1}^N \sum_{i=1}^{l(x^k)} \text{loss}(x_i^k, y_i^k, \hat{y}_i^k) \quad (1)$$

where  $\hat{y}_i^k = \arg\max_y s(x_i^k, y)$ ,  $y \in L$  and  $\text{loss}(x_i^k, y_i^k, \hat{y}_i^k) = \max\{s(x_i^k, \hat{y}_i^k) - s(x_i^k, y_i^k), 0\}$ . This problem can be transformed into a quadratic programming problem as follows:

$$\begin{aligned} \arg\min_w \quad & \frac{1}{2} \|w\|^2 \\ \text{s. t.} \quad & \text{loss}(x_i^k, \bar{y}_i^k, y_i^k) \geq 1 \text{ for } (x_i^k, y_i^k) \in S^C \\ \text{where,} \quad & \bar{y}_i^k = \arg\max_y s(x_i^k, y), y \in L - \{y_i^k\} \end{aligned} \quad (2)$$

Many algorithms have been proposed to optimize (2), such as Sequential Minimal Optimization (SMO) (J. P., 1998) and Cutting Plane (CP) (J. P. 1998; Franc, 2008; Keerthi et al., 2008). In our experiments, we used liblinear

(<http://www.csie.ntu.edu.tw/~cjlin/liblinear/>) as an implementation of SVM, which optimizes (2) by CP.

SSVM uses a similar method to model the sequence labelling problems. The discriminative function for a sequence labelling sample  $(x^k, y^k)$  can be represented by 1st-order Markov chain in the following form:

$$s(x^k, y^k) = \sum_{i=1}^l (w_e f_e(y_i, x_i) + w_s f_s(y_i, y_{i-1}, x_i)) \quad (3)$$

where  $f_e(y_i, x_i)$  are emission features, and  $f_s(y_i, y_{i-1}, x_i)$  are transmission features. The sequence labeling problem can be formatted as a quadratic programming problem as follows:

$$\text{argmin}_w \frac{1}{2} \|w\|^2 \quad (4)$$

$$\text{s. t. } \text{loss}(x^k, \bar{y}^k, y^k) \geq \text{loss}(\bar{y}^k, y^k) \text{ for } (x^k, y^k) \in S^L$$

$$\text{where } \bar{y}^k = \text{argmax}_y s(x^k, y), y \in L^{(x^k)} - \{\bar{y}^k\}$$

$$\text{loss}(\bar{y}^k, y^k) \text{ is a loss function of } \bar{y}^k \text{ and } y^k.$$

↵

There are several types of loss function and the Hamming window distance is usually used for sequence labeling problems. It can be written as  $\text{loss}(\bar{y}^k, y^k) = \sum_{i=1}^{l(y^k)} I(y_i^k \neq \bar{y}_i^k)$ , where  $I(\cdot)$  means whether the condition in the parenthesis is satisfied. The equal (4) can also be solved by SMO and CP. In our experiments, we used SVMhmm ([http://www.cs.cornell.edu/people/tj/svm\\_light/svm\\_hmm.html](http://www.cs.cornell.edu/people/tj/svm_light/svm_hmm.html)) as an implement of SSVM, which solved (4) by CP.

↵

ME uses an exponential distribution to model the conditional distribution of a random variable  $y_i^k$  on an observation  $x_i^k$ :  $p(y_i^k | x_i^k) = \frac{1}{Z(x_i^k, w)} \exp(wf(x_i^k, y_i^k))$ , where  $Z(x_i^k, w) = \sum_y \exp(wf(x_i^k, y_i^k))$ . The maximum log-likelihood estimation function on the training samples  $S^C$  can be written as:

$$L(S^C) = -\log(\prod_{k=1}^N \prod_{i=1}^{l(x^k)} p(y_i^k | x_i^k, w)) \quad (5)$$

which can be solved by Generalized Iterative Scaling (GIS) (Darroch, 1972), Broyden–Fletcher–Goldfarb–Shanno (BFGS) (Head & Zener, 1985), limited-memory BFGS (L-BFGS) (Liu, 1989), stochastic gradient (SG) (Vishwanathan et al., 2006), and so on. In our experiments, we used maxent ([http://homepages.inf.ed.ac.uk/lzhang10/maxent\\_toolkit.html](http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html)) as an implement of ME, and set L-BFGS as its training algorithm.

CRF uses an undirected graph to model the conditional distribution of random variables  $Y$  conditioned on observations  $X$ :  $p(Y|X)$ . For example, given a sample of length  $l$ ,  $(x, y)$ , the conditional probability  $p(y|x)$  can be represented by 1st-order Markov chain in the following form:

$$p(y|x) = \frac{1}{Z(x, w)} \exp \sum_{i=1}^l (w_e f_e(y_i, x_i) + w_s f_s(y_i, y_{i-1}, x_i)) \quad (6)$$

where  $Z(x, w) = \sum_y \exp \sum_{i=1}^l (w_e f_e(y_i, x_i) + w_s f_s(y_i, y_{i-1}, x_i))$  is a normalization factor.

The maximum log-likelihood estimation function on the training samples  $S^L$  can be written as:

$$L(S^L) = -\log(\prod_{k=1}^N p(y^k|x^k, w)) \quad (7)$$

The equal (7) can be solved by the similar algorithms used for ME, such as GIS, BFGS, L-BFGS, SG and so on. In our experiments, we used CRF++ (<http://crfpp.googlecode.com/svn/trunk/doc/index.html>) as an implementation of CRF, which optimizes (7) by L-BFGS.

### 5.3. Experiments and evaluation

For both discharge summaries and admission notes, we divided the 400 notes into two subsets: two-third (266 notes) for training and one-third (134 notes) for testing. The

parameters of classifiers were optimized using the training set via a 10-fold cross validation (CV) method. Then we evaluated and reported the performance using the independent test set. As CRF is the most widely used algorithm for NER, we first investigated the effects of different types of features based on the CRF classifier. We started with the baseline system that used bag-of-character features only, and then progressively added bag-of-word features based on different segmentation methods, POS tags, and section information. Once the optimized feature combination was identified based on the CRF classifier, we evaluated the performance of other machine learning algorithms (SVM, ME and SSVM) using the same sets of features.

The performance of NER systems was measured by standard micro-averaged precision, recall, and F-measure for all entities (O. Uzuner & DuVall, 2010). We developed an evaluation tool to calculate their values based on the official evaluation program developed in the 2010 i2b2 NLP challenge. The evaluation program provides two sets of measures: exact-match and inexact-match, where exact-match means that an entity is correctly predicted, if and only if, the starting and ending offsets are exactly the same as those in the gold standard. The inexact-match means that an entity is correctly predicted if it overlaps with any entity in the gold standard.

In this study, we are attempting to systematically investigate features and ML algorithms for the NER task in Chinese clinical text, using a manually annotated corpus of 400 admission notes and 400 discharge summaries.



## 5.4. Results

A comprehensive study of features and NER ML-based algorithms has been investigated. Table 8 shows the performance of the CRF-based systems on test sets when different features were used for admission and discharge summaries, respectively. The numbers in column 2-5 are F-measures followed by corresponding recall and precision values in a parenthesis for all entities using the exact-matching or inexact-matching criteria. Both word segmentation approaches slightly improved the NER performance and the dictionary lookup method seemed to have better performance. For example, on discharge summaries, the Stanford Word Segmenter improved F-measure from 88.89% to 89.01%; while the dictionary lookup approach improved F-measure from 88.89% to 89.19%. The POS tag information following Stanford segmentation did not further improve the NER performance. The section information also helped the NER system slightly (F-measure 88.95% vs. 88.89% at baseline on discharge summaries). However, such an improvement is minimal, as the 95% confidence intervals for the F-measure of “BOC + SECTION” were [88.46, 89.42] ( $88.94 \pm 0.48$ ) using two-tailed t-test based on bootstrapping sampling that randomly selected 5000 sentences with replacement for 200 times.

The best performance, F-measures of 89.23% and 93.52% for discharge and admission notes, respectively, was achieved when bag-of-character, bag-of-word from the dictionary lookup, and section information were combined. In addition, we noticed that the NER systems always showed better performance on admission notes than discharge

summaries, when same features were used. For example, when only the bag-of-character features were used, the F-measure of the CRF-based NER system was 93.18% on admission notes vs. 88.89% on discharge summaries.

Table 8. *The performance of the CRF-based NER systems on Chinese admission and discharge notes when different features were used*

Features <sup>*</sup>	Admission notes		Discharge summaries	
	exact-match F(R/P)	inexact-match F(R/P)	exact-match F(R/P)	inexact-match F(R/P)
BOC	93.18 (93.70/92.66)	94.32 (94.85/93.80)	88.89 (89.80/87.99)	90.75 (91.68/89.83)
BOC+BOW-STAN	93.19 (93.59/92.79)	94.40 (94.81/94.00)	89.01 (89.87/88.16)	90.95 (91.83/90.08)
BOC+BOW-STAN +POS	93.14 (93.46/92.81)	94.37 (94.70/94.04)	88.89 (89.59/88.21)	90.86 (91.57/90.16)
BOC+BOW-DICT	93.30 (93.66/92.94)	94.50 (94.87/94.13)	89.19 (90.16/88.24)	90.97 (91.96/90.00)
BOC+SECTION	93.28 (93.63/92.93)	94.40 (94.76/94.05)	88.95 (89.96/87.96)	90.71 (91.74/89.70)
BOC+BOW-STAN +SECTION	93.22 (93.61/92.83)	94.45 (94.85/94.06)	89.02 (89.95/88.12)	90.89 (91.83/89.96)
BOC+BOW-DICT +SECTION	<b>93.52</b> (93.77/93.26)	<b>94.69</b> (94.95/94.43)	<b>89.23</b> (90.29/88.20)	<b>91.00</b> (92.08/89.94)

*Note:* “BOC”, “BOW-STAN”, “BOW-DICT”, “POS”, and “SECTION” denote “bag-of-character”, “bag-of-word from Stanford segmenter”, “bag-of-word from dictionary lookup”, “part-of-speech information from Stanford segmenter”, and “section information” respectively.

The detailed results of the best CRF-based NER system for each entity type are shown in Table 9. F-measures ranged from 82.89% to 95.06% for admission notes and 78.91% to 91.82% for discharge summaries among the four types of entities. The performance for



labtests was the best, and the worst for procedures. For each type of entity, the precision was always higher than recall.

Table 9. *The detailed results of the best CRF-based NER system on admission and discharge summaries for each entity type*

Entity	Admission notes		Discharge summaries	
	exact-match	inexact-match	exact-match	inexact-match
Overall	93.52(93.77/93.26)	94.69(94.95/94.43)	89.23(90.29/88.20)	91.00(92.08/89.94)
Problem	93.96(93.99/93.92)	95.35(95.39/95.32)	90.19(90.61/89.77)	92.20(92.63/91.77)
Procedure	82.89(85.44/80.48)	85.34(87.97/82.86)	78.51(82.80/74.64)	81.48(85.93/77.46)
Labtest	95.06(95.22/94.91)	95.41(95.56/95.26)	91.82(92.22/91.42)	92.89(93.30/92.49)
Medicine	86.44(88.18/84.76)	88.98(90.78/87.26)	87.41(90.82/84.24)	88.33(91.78/85.13)

Using the optimized features sets (bag-of-character, bag-of-word from the dictionary lookup, and section information), we compared four ML algorithms on admission and discharge notes. Results are reported in Table 10. The sequence labeling algorithms (CRF and SSVM) were superior to the classification algorithms (ME and SVM). For example, SSVM outperformed SVM by 2.99% and 4.45% in F-measure for admission notes and discharge summaries, respectively. The best performance was achieved by SSVM, which was similar to CRF on admission notes (93.53% vs. 93.52%), but was better than CRF on discharge summaries (90.01% vs. 89.23%).

Table 10. *Comparison of four state-of-the-art machine learning algorithms on Chinese admission and discharge summaries when optimized features were used.*

Algorithm	Admission notes		Discharge summaries	
	exact-match	inexact-match	exact-match	inexact-match
SVM	90.54(90.81/90.27)	93.70(93.99/93.42)	85.56(85.89/85.21)	89.87(90.23/89.52)
ME	90.43(91.07/89.80)	93.49(94.15/92.84)	85.15(86.01/84.30)	89.70(90.61/88.80)
CRF	93.52(93.77/93.26)	94.69(94.95/94.43)	89.23(90.29/88.20)	91.00(92.08/89.94)
SSVM	<b>93.53</b> (92.93/94.15)	<b>95.35</b> (94.72/95.97)	<b>90.01</b> (89.19/90.84)	<b>92.65</b> (91.91/93.51)

## 5.5 Discussion

In this study, we investigated ML-based approaches for NER in Chinese clinical text. Using the manually created annotated datasets of 400 admission notes and 400 discharge summaries in Chinese, we systematically evaluated the contributions of different types of features and ML algorithms for NER in Chinese clinical text. The results showed that word segmentation information based on a Chinese medical dictionary and section information was beneficial to NER tasks in Chinese clinical text. When the same features were used, we also demonstrated that SSVM achieved the best performance among the four different ML algorithms. This was consistent with a previous study on NER in English clinical text (B. Tang & Xu, 2012; M.Jiang & Xu, 2011). All of the above findings shed light for future work in Chinese clinical NLP research.

The best performance of our NER system for Chinese discharge summaries was 90.01% in F-measure, which is similar to the best F-measure (90.24%) reported in another recent NER study on Chinese discharge summaries (Y. Wang & Jiang, 2012). These results were much better than the best result in the 2010 i2b2/VA NLP challenge on clinical entity recognition from English discharge summaries (F-measure 85.23%) (O. Uzuner & DuVall, 2010; O. Uzuner & Cadag, 2010). It would be difficult to determine the exact reasons why English clinical text is more difficult for NER tasks. We conducted an analysis of entity frequency in both English (the i2b2 corpus) and Chinese discharge summaries. It seemed that entities in English clinical text were sparser than those in Chinese clinical text. In Chinese discharge summaries, 53.18% of entities occurred once; however, 76.02% of entities occurred once in English clinical text. Therefore, we estimate that the higher percentage of low frequency entities could be one reason for the performance difference between English and Chinese clinical text. Moreover, the difference between the exact-matching and inexact-matching F-measures of our best NER system on Chinese discharge summaries (2.64%) is much smaller than the best result on the i2b2/VA NLP challenge on clinical entity recognition in English discharge summaries (8.39%) (O. Uzuner & DuVall, 2010; O. Uzuner & Cadag, 2010), indicating that boundaries of entities in Chinese clinical text are easier to be determined than entities in English clinical text. It may be another reason for the performance difference between English and Chinese clinical entity recognition.

Word segmentation is one of the major differences between English and Chinese text processing. However, when using the Stanford Word Segmenter, a state-of-art Chinese segmenter in the general domain, the performance of the NER system did not improve much. More improvement was observed when a Chinese medical dictionary was used for word segmentation. This finding suggests that domain knowledge is important for word segmentation in Chinese clinical text. In the future, we plan to work on domain-specific word segmentation approaches for Chinese clinical text by combining medical knowledge bases with statistical word segmentation methods, to further improve the NER performance. It is not surprising that the sequence labeling algorithms were superior to the classification algorithms for NER in Chinese clinic notes, as sequence labeling algorithms take the relationships between neighbor labels into consideration. However, it is important to verify that SSVM, a relatively new sequential labeling algorithm, could achieve slightly better performance than CRF on NER in Chinese clinical text. This finding, together with our previous reported results (B. Tang & Xn, 2012; M. Jiang & Xu, 2011), demonstrated that SSVM could be a competitive alternative to CRF, on NER tasks in both English and Chinese clinical texts.

Furthermore, we conducted an analysis on errors in our best system. We found that most errors occurred in long entities with combined structures. For example, in a long problem entity “肝(liver)功能(function)异常(abnormal)急性(acute)加重(exacerbation)” (acute exacerbation of abnormal liver function), only part of it -- “肝(liver)功能(function)异常”

(abnormal) was predicted as a problem. Information about the syntactic structures of Chinese sentences could potentially help this scenario. However, there is very limited work on syntactic parsing of clinical text in Chinese, which requires extensive resources and effort (e.g., building a Treebank), but is probably worth investigating. Another type of error was caused by unseen samples in the training set. For example, a procedure “停 (discontinue)呼吸机(ventilator)” (discontinue ventilator) was not detected because there were no similar medical concepts in the training dataset. Increasing sample size could potentially solve this problem.

## **Chapter 6: Key findings, Contribution, Future work and Conclusions**

### **6.1. Overview and summary of key findings**

In this dissertation, we conducted three sub-studies to advance NER research in clinical text in Chinese:

- (1) We first developed an annotation schema for NER in Chinese clinical text and created two manually annotated Chinese corpora containing 400 admission notes and 400 discharge summaries respectively. Using an inductive approach for schema development and extensive training for annotators, we demonstrated high IAA on both Chinese clinical corpora, indicating the success of the annotation process.
- (2) We then compared distributions of clinical entities in discharge summaries in US (from the 2010 i2b2 clinical NLP challenge) with those in Chinese (annotated in project #1), which not only reveals potential differences between US and China in clinical practice, but also identifies useful features for NER in Chinese clinical text.
- (3) Finally, we developed a ML-based NER system for Chinese clinical text. We systematically investigated features and ML algorithms for the NER task in Chinese clinical text using annotated corpora of admission notes and discharge summaries. Our results suggests that both word segmentation and section information improved NER in Chinese clinical text, and SSVM, a recent sequential labeling algorithm, outperformed CRF and other classification algorithms. Our best system achieved F-measures of 90.01% and 93.52% on Chinese discharge

summaries and admission notes, respectively, indicating a promising start on Chinese NLP research.

In summary, we demonstrated that by creating annotated clinical Chinese corpora, optimizing linguistics and domain specific features, and implementing state-of-the-art ML algorithms, we could build high-performance ML-based NER methods to detect clinical entities in Chinese clinical text.

## **6.2. Innovations and contributions**

This work is innovative as it is the first comprehensive study of NER in Chinese clinical text. It has created one of the earliest large annotated Chinese corpora in the medical domain. Using the annotated Chinese corpus of discharge summaries, we conducted the first comparative study of content between English and Chinese clinical documents. Moreover, it is also the first time that SSVM was used for NER in Chinese clinical corpora and demonstrated superior performance.

The major contribution of this dissertation work to biomedical informatics is that it develops a framework for NER in Chinese clinical text, by 1) creating annotated corpora; 2) optimizing features for NER; and 3) implementing state-of-the-art ML algorithms. The two Chinese clinical corpora created in this study would be valuable resources for method development in clinical NLP in Chinese. Findings learned from the systematic evaluation of features and ML algorithms in Chinese clinical NER would provide insights

for other researchers in the field. This study also contributes to computer and information science. It demonstrates that ML-based NER approaches are effective in Chinese text in closed domains such as medicine. Furthermore, this study also benefits healthcare. It develops automated methods to extract information from clinical narratives in Chinese, thus making it possible to utilize free text in Chinese EHRs for healthcare operation and clinical research.

### **6.3. Future work**

This dissertation work is just a beginning of research in clinical NLP in Chinese. In the future, we plan to conduct more extensive studies on various aspects of NLP in Chinese clinical text. First, we would like to include more types of clinical documents and clinical data from multiple institutions into our study, to investigate the generalizability and portability of current NER methods. In addition, we plan to develop more clinical corpora in Chinese, by investigating more efficient methods for corpus development, such as machine assisted pre-annotation (Lingren et al., 2014; South et al., 2014; Gobbel et al., 2014), active learning (Chen, Mani, & Xu, 2012; Figueroa et al., 2012; Chen et al., 2013) or domain adaptation (Chiricariu et al., 2010; Guo et al., 2009) technologies. Another interesting research direction is to study unsupervised representation learning technologies (e.g., "deep learning") in NER. Specifically, we will investigate the use of deep learning in two ways: (1) We will feed word representations (Turian, Ratnov, & Bengio, 2010) (a.k.a. word embedding) learned by existing neural language models



(Bengio et al., 2006; Mnih & Hinton, 2009) as features to traditional NER algorithms such as CRF (2). We will investigate deep neural networks that combine feature abstraction and classification in one unified architecture, such as the earlier work by Collobert and Weston using convolutional networks (Collobert & Weston, 2008; Collobert et al., 2011) or more recent approaches, e.g., using recurrent neural networks that incorporate elements of the CRF model (Yao et al., 2014).

#### **6.4. Conclusion**

By creating annotated clinical corpora, optimizing linguistics and domain specific features, and implementing state-of-the-art machine learning algorithms, we have demonstrated that it is feasible to develop high-performance ML-based NER systems to detect entities in Chinese clinical text. We envision that such advanced clinical NLP methods would accelerate the use of clinical documents in ERHs to improve healthcare quality/safety and to facilitate clinical and translational research in China.

## References

- Aaron L. -F. Han DFW, Lidia S. Chao. Chinese Named Entity Recognition with Conditional Random Fields in the Light of Chinese Characteristics. *Language Processing and Intelligent Information Systems. Lecture Notes in Computer Science* 2013 Volume 7912, pp 57-68.
- Alberto Lavelli MEC, Fabio Ciravegna, Dayne Freitag, Claudio Giuliano, Nicholas Kushmerick, Lorenza Romano, Neil Ireson. Evaluation of machine learning-based information extraction algorithms: criticisms and recommendations. *Language Resources and Evaluation*. 2008 Volume 42, Issue 4, pp 361-393.
- AR A. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. *Proc AMIA Symp*. 2001 17–21.
- Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc*. 2010 May-Jun;17(3):229-36.
- Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc*. 2010 vol. 17, no. 3, pp. 229–236.
- Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp*. 2001:17-21.
- Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp*. 2001:17-21.
- Asif Ekbal SS. Combining feature selection and classifier ensemble using a multiobjective simulated annealing approach: application to named entity recognition. *Soft Computing*. 2013 Volume 17, Issue 1, pp 1-16.
- AT. M. Extending a natural language parser with UMLS knowledge. *Proc Annu Symp Comput Appl Med Care*. 1991 p. 194-198.
- B. de Bruijn, C. Cherry S. Kiritchenko J. Martin, X. Zhu. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *J Am Med Inform Assoc*. 2011 vol. 18, no. 5, pp. 557–562.

- B. Tang HCYWMJ, Xu H. Clinical entity recognition using structural support vector machines with rich features. in Proceedings of the ACM sixth international workshop on Data and text mining in biomedical informatics. 2012 New York, NY, USA, pp. 13–20.
- B. Tang HCYWMJ, Xu H. Recognizing clinical entities in hospital discharge summaries using Structural Support Vector Machines with word representation features. BMC Med Inform Decis Mak. 2013 vol. 13, no. 1, pp. 1–10.
- B. Taskar CG, Koller D. Max-margin Markov networks. 2003.
- Barrett N, Weber-Jahnke J.. A token centric part-of-speech tagger for biomedical text. Artif Intell Med. 2014 61(1):11-20. doi: 10.1016/j.artmed.2014.03.005. Epub 2014 Mar 26.
- Baud RH RA, Scherrer JR. Natural language processing and semantical representation of medical texts. Method Inform Med. 1992 31(2):117-25.
- Bengio Y, Schwenk H, Sen écal J-S, Morin F, Gauvain J-L. Neural probabilistic language models. Innovations in Machine Learning: Springer; 2006. p. 137-86.
- Bontcheva K, Cunningham H, Roberts I, et al. GATE Teamware: a web-based, collaborative text annotation framework. Language Resources and Evaluation. 2013;47(4):1007-29.
- Byrd R1 SS, Sun J, Ebadollahi S, Stewart WF. Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records. Int J Med Inform 2013 pii: S1386-5056(12)00246-8. doi: 10.1016/j.ijmedinf.2012.12.005. [Epub ahead of print].
- C C. The horizontal and vertical nature of patient phenotype retrieval: new directions for clinical text processing. Proc AMIA Symposium. 2002 Washington DC: 165-9.
- C. Friedman POAJHAJJC, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc. 1994 vol. 1, no. 2, pp. 161–174.
- Cairns BL, Nielsen RD, Masanz JJ, et al. The MiPACQ clinical question answering system. AMIA Annu Symp Proc. 2011;2011:171-80.

- Carol Friedman TCR, Milton Corn. Natural Language processing: State of the art and prospects for significant process, a workshop sponsored by the National Library of Medicine. *Journal of biomedical Informatics*. 2013 26: 765-773.
- Changki Lee Y-GH, Hyo-Jung Oh, Soojong Lim, Jeong Heo, Chung-Hee Lee, Hyeon-Jin Kim, Ji-Hyun Wang, Myung-Gil Jang. Fine-Grained Named Entity Recognition Using Conditional Random Fields for Question Answering. *Information Retrieval Technology, Lecture Notes in Computer Science*. Volume 4182, 2006, pp 581-587.
- Chapman W BWHPea. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform*. 2001 34:301-10.
- Chapman WW DJ, Wagner MM. Fever detection from free-text clinical records for biosurveillance. *J Biomed Inform*. 2004 37:120-7. .
- Chapman WW FM, Dowling JN, et al. Identifying respiratory findings in emergency department reports for biosurveillance using MetaMap. *Stud Health Technol Inform*. 2004 07:487-91.
- Chapman WW, Dowling JN, Hripcsak G. Evaluation of training with an annotation schema for manual annotation of clinical conditions from emergency department reports. *Int J Med Inform*. 2008 Feb;77(2):107-13.
- Chapman WW, Dowling JN. Inductive creation of an annotation schema for manually indexing clinical conditions from emergency department reports. *J Biomed Inform*. 2006 Apr;39(2):196-208.
- Charmaz K. *Constructing grounded theory*. Sage. 2006.
- Chen L FC. Extracting phenotypic information from the literature via natural language processing. *Stud Health Technol Inform*. 2004 107:758-62.
- Chen Y, Carroll RJ, Hinz ER, et al. Applying active learning to high-throughput phenotyping algorithms for electronic health records data. *J Am Med Inform Assoc*. 2013 Dec;20(e2):e253-9.
- Chen Y, Mani S, Xu H. Applying active learning to assertion classification of concepts in clinical text. *J Biomed Inform*. 2012 Apr;45(2):265-72.

Chinese MOH Notice on First 97 trial hospitals for EMR.

<http://www.moh.gov.cn/publicfiles/business/htmlfiles/mohyzs/s3586/201105/51779.htm>. 2011.

Chinese MOH Notice on Second 92 trial hospitals for EMR.

<http://www.moh.gov.cn/publicfiles/business/htmlfiles/mohyzs/s3586/201111/53273.htm>. 2011.

Chiticariu L, Krishnamurthy R, Li Y, Reiss F, Vaithyanathan S. Domain adaptation of rule-based annotators for named-entity recognition tasks. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing; 2010: Association for Computational Linguistics; 2010. p. 1002-12.

Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural language processing (almost) from scratch. The Journal of Machine Learning Research. 2011;12:2493-537.

Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning. Proceedings of the 25th international conference on Machine learning; 2008: ACM; 2008. p. 160-7.

conference Tr. <http://trechnistgov/>.

Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995 vol. 20, no. 3, pp. 273–297.

Cunningham H, Maynard D, Bontcheva K. Text processing with gate: Gateway Press CA; 2011.

Darroch J RD. Generalized Iterative Scaling for Log-Linear Models. . Ann Math Stat. 1972 43:1470–80.

de Bruijn B, Cherry C, Kiritchenko S, Martin J, Zhu X. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. J Am Med Inform Assoc. 2011 Sep-Oct;18(5):557-62.

deBruijn B CC, Kiritchenko S, et al . . NRC at i2b2: one challenge, three practical tasks, nine statistical systems, hundreds of clinical records, millions of useful features.

- Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data. 2010 Boston, MA, USA: i2b2.
- Deleger L, Namer F, Zweigenbaum P. Morphosemantic parsing of medical compound words: transferring a French analyzer to English. *Int J Med Inform.* 2009 78(suppl 1):S48-55.
- Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? *J Biomed Inform.* 2009 Oct;42(5):760-72.
- Denny JC SA, Miller RA, et al. Identifying UMLS concepts from ECG impressions using KnowledgeMap. *AMIA Annu Symp Proc.* 2005 196–200.
- Denny JC SJ, Miller RA, Spickard III A. Understanding medical school curriculum content using KnowledgeMap. *J Am Med Inform Assoc.* 2003 10(4):351–362.
- Denny JC, Smithers JD, Miller RA, Spickard A. 3rd. "Understanding" medical school curriculum content using KnowledgeMap. *J Am Med Inform Assoc.* 2003 Jul-Aug;10(4):351-62.
- Denny Jc, Spickard A. rd, Johnson K. B., Peterson N. B., Peterson J. F., Miller R. A. Evaluation of a method to identify and categorize section headers in clinical documents. *J Am Med Inform Assoc.* 2009 Nov-Dec;16(6):806-15.
- Doddi S MA, Ravi SS, Torney DC. Discovery of association rules in medical data. *Med Inform Internet Med* 2001 26(1):25–33.
- Duan H BXea. Chinese Word Segmentation at Peking University. *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing.* 2003 pp. 152-155.
- EMR basic architecture and data standards.  
<http://www.moh.gov.cn/publicfiles/business/htmlfiles/mohbgt/s6694/200908/42155.htm>. 2009.
- F. W. Automated indexing of SNOMED statements into ICD. *Method Inf Med.* 1987 26(3):93-8.
- Fan JW, Prasad R, Yabut RM, et al. Part-of-speech tagging for clinical text: wall or bridge between institutions? *AMIA Annu Symp Proc.* 2011;2011:382-91.

- Fan JW, Yang EW, Jiang M, et al. Syntactic parsing of clinical text: guideline and corpus development with handling ill-formed sentences. *J Am Med Inform Assoc.* 2013 Nov-Dec;20(6):1168-77.
- Figuerola RL, Zeng-Treitler Q, Ngo LH, Goryachev S, Wiechmann EP. Active learning for clinical text classification: is it better than random sampling? *J Am Med Inform Assoc.* 2012 Sep-Oct;19(5):809-16.
- Florian Boudin, Jian-Yun Nie, Joan C Bartlett, Roland Grad, Pierre Pluye, Martin Dawes. Combining classifiers for robust PICO element detection. *BMC Medical Informatics and Decision Making.* 2010 May 2010, 10:29,.
- Franco V SS. Optimized cutting plane algorithm for support vector machines. . *Proceedings of the 25th international conference on Machine learning.* 2008 New York, NY, USA, 320-7.
- Friedman C, Alderson P. O., Austin J. H., Cimino J. J., Johnson S. B. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association.* 1994 Mar-Apr;1(2):161-174.
- Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. *J Am Med Inform Assoc.* 1994 Mar-Apr;1(2):161-74.
- Friedman C, Kra P, Rzhetsky A. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *J Biomed Inform.* 2002 Aug;35(4):222-35.
- Friedman C, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. *J Am Med Inform Assoc.* 2004 11(5):392–402.
- Friedman C, Rindflesch TC, Corn M. Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine. *J Biomed Inform.* 2013 Oct;46(5):765-73.
- Friedman C. A broad-coverage natural language processing system. *Proc AMIA Symp.* 2000:270-4.
- Friedman C. Towards a comprehensive medical language processing system: methods and issues. *Proc AMIA Annu Fall Symp.* 1997:595-9.

- Gawron AJ TW, Keswani RN, Rasmussen LV, Kho AN. Anatomic and Advanced Adenoma Detection Rates as Quality Metrics Determined via Natural Language Processing. *Am J Gastroenterol*. 2014 Jun 17. doi: 10.1038/ajg.2014.147. [Epub ahead of print].
- Gobbel GT, Garvin J, Reeves R, et al. Assisted annotation of medical free text using RapTAT. *J Am Med Inform Assoc*. 2014 Jan 15.
- Goryachev S, Sordo M, Zeng QT. A suite of natural language processing tools developed for the I2B2 project. *AMIA Annu Symp Proc*. 2006:931.
- Grabar N, Rizand P, Livartowski A, Hamon T. Automatic acquisition of synonym resources and assessment of their impact on the enhanced search in EHRs. *Methods Inf Med*. 2009 48(2):149-54. doi: 10.3414/ME9213. Epub 2009 Feb 18.
- Guo H, Zhu H, Guo Z, Zhang X, Wu X, Su Z. Domain adaptation with latent semantic association for named entity recognition. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics; 2009: Association for Computational Linguistics; 2009*. p. 281-9.
- Gurulingappa H H-AM, Fluck J . . Concept identification and assertion classification in patient health records. . *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data Boston, MA, USA: i2b2*. 2010.
- Harkema H, Dowling JN, Thornblade T, Chapman WW. ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *J Biomed Inform*. 2009 Oct;42(5):839-51.
- Harris Z. *A Grammar of English on mathematical principles*. NY: Wiley & Sons. 1982.
- Harris Z. *A theory of language and information: a mathematical approach*. Oxford: Clarendon Press;. 1991.
- Harris Z. *Mathematical structures of language*. NY: Wiley Interscience. 1968.



- Haug P KS, Lau LM, Wang P, Rocha R, Huff S. A natural language understanding system combining syntactic and semantic techniques. *annu symp comput appl med care*; 1994 p. 247–51.
- Head JD, Zerner MC. A Broyden-Fletcher-Goldfarb-Shanno optimization procedure for molecular geometries.. *Chem Phys Lett* 1985 122:264–70.
- Health USD-po, Human S. Secretary Sebelius Announces Final Rules To Support Meaningful Use of Electronic Health Records.
- Heng-Li Yang AFYC. Sentiment analysis for Chinese reviews of movies in multi-genre based on morpheme-based features and collocations. *Information Systems Frontiers*. 2014.
- Hripcsak G AJ, Alderson PO, Friedman C. Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports. *Radiology*. 2002 224(1):157–63.
- Hripcsak G EN, Chen YH, et al. Using empiric semantic correlation to interpret temporal assertions in clinical texts. *J Am Med Inform Assoc*. 2009 16:220-7.
- Hripcsak G FC, Alderson PO, DuMouchel W, Johnson SB, Clayton PD. Unlocking clinical data from narrative reports. *Ann Int Med*. 1995 122(9):681–8.
- Hripcsak G, Rothschild AS. Agreement, the F-Measure, and Reliability in Information Retrieval. *J Am Med Inform Assoc*. 2005 vol. 12, no. 3, pp. 296–298.
- Huang Y LH. A Novel Hybrid Approach to Automated Negation Detection in Clinical Radiology Reports. *J Am Med Inform Assoc*. 2007 14:304-11.
- I. Tsochantaridis TJTH, Altun Y. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*. 2005 vol. 6, pp. 1453–1484.
- I2b2 2012 annotation guidelines.  
<https://www.i2b2.org/NLP/Relations/assets/Assertion%20Annotation%20Guidelines.pdf>. 2012.
- Iii HD. Frustratingly Easy Domain Adaptation. <http://www.cs.utah.edu/~hal/docs/daume07easyadapt.pdf>.

- Imler TD MJ, Imperiale TF. Clinical decision support with natural language processing facilitates determination of colonoscopy surveillance intervals. *Clin Gastroenterol Hepatol*. 2014 12(7):1130-6. doi: 10.1016/j.cgh.2013.11.025. Epub 2013 Dec 4.
- J. Lafferty AM, Pereira FCN. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Departmental Papers (CIS). 2001.
- J. P. Sequential minimal optimization: a fast algorithm for training support vector machines. *Technique Report* 1998 1-21.
- Jain NL, Friedman C. Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports. *Proc AMIA Annu Falls Symp*. 1997 829–833.
- Jain NL, Knirsch CA, Friedman C, Hripcsak G. Identification of suspected tuberculosis patients based on natural language processing of chest radiograph reports. *Proc AMIA Annu Falls Symp*. 1996 542–546. .
- Jiang M CY, Liu M, et al. Hybrid approaches to concept extraction and assertion classification - vanderbilt's systems for 2010 I2B2 NLP Challenge. . *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data* 2010 Boston, MA, USA: i2b2.
- Jiang M, Chen Y, Liu M, et al. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *J Am Med Inform Assoc*. 2011 Sep-Oct;18(5):601-6.
- Jiebai Zhou DWXL, et al. Translational medicine as a permanent glue and force of clinical medicine and public health: Perspectives (1) from 2012 Sino-American symposium on clinical and translational medicine. *Clin Transl Med*. 2012 1:21.
- Jonnalagadda S GG. Can distributional statistics aid clinical concept extraction? . *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data* Boston, MA, USA: i2b2. 2010.
- Kang N BR, Afzal Z, et al . Erasmus MC. approaches to the i2b2 Challenge. *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*. 2010 Boston, MA, USA: i2b2.

- Keerthi SS, Chang K-W, et al. A sequential dual method for large scale multi-class linear SVMs. . the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. 2008 408–16.
- Kho AN, Pacheco JA, Peissig PL, et al. Electronic medical records for genetic research: results of the eMERGE consortium. *Sci Transl Med*. 2011 Apr 20;3(79):79re1.
- Kudo T, Matsumoto Y. Chunking with support vector machines. presented at the NAACL 01: Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies YYYY:2001, 2001. 2001 pp. 1–8.
- Kudoh T, Matsumoto Y. Use of support vector learning for chunk identification. presented at the Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning. 2000 Volume 7, pp. 142–144.:vol. 20, no. 3, pp. 273–297.
- Lev Ratnov DR. Design challenges and misconceptions in named entity recognition. Proceedings of the thirteenth conference on computational natural language learning(CONLL). 1999 <http://cogcomp.cs.illinois.edu/papers/RatnovRo09.pdf>.
- Li Q, Zhai H, Deleger L, et al. A sequence labeling approach to link medications and their attributes in clinical notes and clinical trial announcements for information extraction. *J Am Med Inform Assoc*. 2013 Sep-Oct;20(5):915-21.
- Lim Mk YHZTFWZZ. Public per-ceptions of private health care in socialist China. *Health Aff (Millwood)*. 2004 23(6):222-34.
- Lin DW, Xiaoyun. Phrase clustering for discriminative learning. Annual Meeting of the ACL and IJCNLP. 2009 pp. 1030–1038.
- Lingren T, Deleger L, Molnar K, et al. Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements. *J Am Med Inform Assoc*. 2014 May-Jun;21(3):406-13.
- Liu DC NJ. On the limited memory BFGS method for large scale optimization. . *Math Programming*. 1989 45:503–28.

- Luo Xiao DW, Michael Brown, Stephan Jablonski. Information Extraction from the Web: System and Techniques. *Applied Intelligence*. 2004 Volume 21, Issue 2, pp 195-224.
- M. Jiang YCMLSTRSMJCD, Xu H. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *J Am Med Inform Assoc*. 2011 vol. 18, no. 5, pp. 601–606.
- McCarty Ca, Chisholm R. L. Chute C. G. Kullo I. J. Jarvik G. P. Larson E. B. Li R. Masys D. R. Ritchie M. D. Roden D. M. Struewing J. P. Wolf W. A. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics*. 2011 (4):13.
- Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. *J Am Med Inform Assoc*. 2005 12(4):448–457.
- Meystre Sm, Savova G. K., Kipper-Schuler K. C., Hurdle J. F. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*. Yearb Med Inform. 2008 128-144.
- Miller G HD. Maximum entropy approach to probability density estimation. 1998 Second International Conference on Knowledge-Based Intelligent Electronic Systems, 1998 Proceedings KES 1998 1:225–30.
- Mitchell KJ, Becich MJ, Berman JJ, et al. Implementation and evaluation of a negation tagger in a pipeline-based system for information extract from pathology reports. *Stud Health Technol Inform*. 2004;107(Pt 1):663-7.
- Mnih A, Hinton GE. A scalable hierarchical distributed language model. *Advances in neural information processing systems*; 2009; 2009. p. 1081-8.
- Moon S, Pakhomov S, Liu N, Ryan JO, Melton GB. A sense inventory for clinical abbreviations and acronyms created using clinical notes and medical dictionary resources. *J Am Med Inform Assoc*. 2014 Mar-Apr;21(2):299-307.
- MUC-7 Mucp. [http://www.itl.nist.gov/iaui/89402/related\\_projects/muc/proceedings/muc\\_7\\_tohtml](http://www.itl.nist.gov/iaui/89402/related_projects/muc/proceedings/muc_7_tohtml).

- Murphy SN WG, Mendis M, Chueh HC, Churchill S, Glaser JP, Kohane IS. . Serving the Enterprise and beyond with Informatics for Integrating Biology and the Bedside (i2b2). . J Am Med Inform Assoc. 2010 17(2):124-30. PMID:20190053.
- Mutalik P DA, Nadkarni P. Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the UMLS. J Am Med Inform Assoc. 2001 8:598-609.
- Nadeau D, Sekine S. A survey of named entity recognition and classification. *Linguisticae Investigationes*. 2007 vol. 30, no. 1, pp. 3–26.
- Naiwen Xue FXF-dC, Marta P. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. *Nat Lang Eng*. 2005 11, 2 (June 2005), 207-238.
- Neil Barrett, Jens Weber-Jahnke. Building a biomedical tokenizer using the token attice design pattern and the adapted Viterbi algorithm. *BMC Bioinformatics*. June 2011 12:S1.
- O. Uzuner BRSSS, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc*. 2011 18:552–6.
- O. Uzuner IS, Cadag E. Extracting medication information from clinical text. *J Am Med Inform Assoc*. 2010 vol. 17, no. 5, pp. 514–518.
- Ogren PV. Knowtator: a protégé plug-in for annotated corpus construction. 2006: Association for Computational Linguistics; 2006. p. 273-5.
- Pace WD, et al. An electronic practice-based network for observational comparative effectiveness research. *Ann Intern Med*. 2009 151(5): p. 338-40.
- Pai AK AE, Post AR, et al . The emory system for extracting medical concepts at 2010 i2b2 challenge: integrating natural language processing and machine learning techniques. *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data Boston, MA, USA: i2b2*.
- Pai VM RM, Conroy R, Luo J, Zhou R, Seto B. Workshop on using natural language processing applications for enhancing clinical decision making: an executive summary. *J Am Med Inform Assoc* 2014 21(e1):e2-5. doi: 10.1136/amiajnl-2013-001896. Epub 2013 Aug 6.

- Pakhomov S, Pedersen T, Chute CG. Abbreviation and acronym disambiguation in clinical discourse. *AMIA Annu Symp Proc*. 2005:589-93.
- Pakhomov SV, Coden A, Chute CG. Developing a corpus of clinical notes manually annotated for part-of-speech. *Int J Med Inform*. 2006 Jun;75(6):418-29.
- Patrick JD ND, Wang Y, et al . . I2b2 challenges in clinical natural language processing 2010. . *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data* Boston, MA, USA: i2b2. 2010.
- Pedersen T. Free NLP Software. <http://www.dumnedu/~tpederse/codehtml>. 2011.
- Pi-Chuan Chang MG, Christopher DM. Optimizing Chinese word segmentation for machine translation performance. In *Proceedings of the Third Workshop on Statistical Machine Translation (StatMT '08)*. Association for Computational Linguistics, Stroudsburg, PA, USA, DDD:224-232. 2008.
- Poibeau TaK, L. Proper Name Extraction from Non-Journalistic Texts. *Proc Computational Linguistics in the Netherlands*. 2001.
- Prakash M Nadkarni LO-M, Wendy W Chapman. Natural language processing: an introduction. *J Am Med Inform Assoc*. 2011 18:544-551.
- R. Subhashini VJSK. A Framework for Efficient Information Retrieval Using NLP Techniques. *Computer Networks and Information Technologies. Communications in Computer and Information Science*. 2011 Volume 142, pp 391-393.
- Report of 2011-2015 Market Survey and prediction of development of China's EMR. <http://www.chinairrorg/report/R10/R1006/201110/31-86054html>. 2011.
- Riloff E. From Manual Knowledge Engineering to Bootstrapping: Progress in Information Extraction and NLP. *Case-Based Reasoning Research and Development. Lecture Notes in Computer Science 2003* Volume 2689, p 4.
- Rindflesch TC, Aronson AR. Ambiguity resolution while mapping free text to the UMLS Metathesaurus. *Proc Annu Symp Comput Appl Med Care*. 1994 240-4.
- Roberts K RB, Harabagiu S . . Extraction of medical concepts, assertions, and relations from discharge summaries for the fourth i2b2/VA shared task. *Proceedings of the*

- 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data Boston, MA, USA: i2b2. 2010.
- Roden D.M, Pulley J. M. Basford M. A. Bernard G. R. Clayton E. W. Balser J. R. Masys D.R. Development of a large-scale de-identified dna biobank to enable personalized medicine. Clin Pharmacol Ther. 2008 84(3):362–9.
- Rohini K. Srihari EP. Named Entity Recognition for Improving Retrieval and Translation of Chinese Documents. Digital Libraries: Universal and Ubiquitous Access to Information, Lecture Notes in Computer Science. 2008 Volume 5362, pp 404-405.
- Rosenbloom ST1 DJ, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. J Am Med Inform Assoc 2011 Mar-Apr;18(2):181-6. doi: 10.1136/jamia.2010.007237. Epub 2011 Jan 12.
- S. Wang SL, Chen T. Recognition of Chinese Medicine Named Entity Based on Condition Random Field. Journal of Xiamen University(Natural Science). 2009 vol. 48, no. 3, pp. 349–364.
- Sager N FC, Lyman M, et al. . Medical language processing: computer management of narrative data. Reading, MA: Addison-Wesley; . 1987.
- Sang EFTK, Meulder FD. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. in Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003. 2003 Volume 4, Stroudsburg, PA, USA, 2003, pp. 142–147.
- Sang EFTK. Introduction to the CoNLL-2002 shared task: language-independent named entity recognition. in proceedings of the 6th conference on Natural language learning. 2002 Volume 20, Stroudsburg, PA, USA, 2002, pp. 1–4.
- Sasaki Y IK, Yamamoto Y, et al. . TTI's systems for 2010 i2b2/VA challenge. Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data Boston, MA, USA: i2b2. 2010.

- Savova GK, Chapman WW, Zheng J, Crowley RS. Anaphoric relations in the clinical narrative: corpus creation. *J Am Med Inform Assoc*. 2011 Jul-Aug;18(4):459-65.
- Savova Gk, Masanz J. J. Ogren P. V., et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association : JAMIA*. 2010 Sep-Oct 17(5):507-513.
- Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc*. 2010 Sep-Oct;17(5):507-13.
- SENSEVAL. <http://www.senseval.org/>.
- Shea S, Hripcsak G. Accelerating the use of electronic health records in physician practices. *N Engl J Med*.
- Shi X CY, et al. Dependency-Based Chinese-English Statistical Machine Translation. *Computational Linguistics and Intelligent Text Processing. Lecture Notes in Computer Science*. 2007 Volume 4394, pp 385-396.
- Shivade C RP, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, Lai AM. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc*. 2014 21(2):221-30. doi: 10.1136/amiajnl-2013-001935. Epub 2013 Nov 7.
- South BR SS, Barrus R, DuVall SL, Uzuner Ö, Weir C. Qualitative analysis of workflow modifications used to generate the reference standard for the 2010 i2b2/VA Challenge. *AMIA Annu Symp Proc*. 2011 1232-1251 PMID: 22195185.
- South BR, Mowery D, Suo Y, et al. Evaluating the effects of machine pre-annotation and an interactive annotation interface on manual de-identification of clinical text. *J Biomed Inform*. 2014 Aug;50:162-72.
- Stubbs A. Mae and mai: Lightweight annotation and adjudication tools. 2011: Association for Computational Linguistics; 2011. p. 129-33



- Sun W RA, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. J Am Med Inform Assoc 2013 Sep-Oct;20(5):806-13. doi: 10.1136/amiajnl-2013-001628. Epub 2013 Apr 5.
- Sun W, Rumshisky A, Uzuner O. Annotating temporal information in clinical narratives. J Biomed Inform. 2013 Dec;46 Suppl:S5-12.
- Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. J Am Med Inform Assoc. 2013 Sep-Oct;20(5):806-13.
- Sun W, Rumshisky A, Uzuner O. Temporal reasoning over clinical text: the state of the art. J Am Med Inform Assoc. 2013 Sep-Oct;20(5):814-9.
- Suominen H, Salanterä S, Velupillai S, et al. Overview of the ShARe/CLEF eHealth evaluation lab 2013. Information Access Evaluation Multilinguality, Multimodality, and Visualization: Springer; 2013. p. 212-31.
- Tang B, Wu Y, Jiang M, Chen Y, Denny JC, Xu H. A hybrid system for temporal information extraction from clinical text. J Am Med Inform Assoc. 2013 Sep-Oct;20(5):828-35.
- Tao C SH, Deepak S, et al. . Time-Oriented Question Answering from Clinical Narratives Using Semantic-Web Techniques. Springer, Berlin: Lecture Note on Computer Science, 2011:6496.
- Terol RM M-BP, Palomar M. A knowledge based method for the medical question answering problem. Comput Biol Med. 2007 37(10):1511-21. Epub 2007 Mar 19.
- Tiejun Zhao YG, Ting Liu, Qiang Wang. Recent advances on NLP research in Harbin Institute of Technology. Frontiers of Computer Science in China. 2007 Volume 1, Issue 4, pp 413-428.
- Torii M LH. BioTagger-GM for detecting clinical concepts in electronic medical reports. Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data Boston, MA, USA: i2b2. 2010.
- Turian J, Ratnoff L, Bengio Y. Word representations: a simple and general method for semi-supervised learning. Proceedings of the 48th Annual Meeting of the

- Association for Computational Linguistics; 2010: Association for Computational Linguistics; 2010. p. 384-94.
- Uzuner O, Bodnari A, Shen S, Forbush T, Pestian J, South BR. Evaluating the state of the art in coreference resolution for electronic medical records. *J Am Med Inform Assoc*. 2012 Sep-Oct;19(5):786-91.
- Uzuner Ö, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*. 2007;14(5):550-63.
- Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. *J Am Med Inform Assoc*. 2010 Sep-Oct;17(5):514-8.
- Uzuner O, Solti I, Xia F, Cadag E. Community annotation experiment for ground truth generation for the i2b2 medication challenge. *J Am Med Inform Assoc*. 2010 Sep-Oct;17(5):519-23.
- Uzuner O, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc*. 2011 Sep-Oct;18(5):552-6.
- Uzuner O. Recognizing obesity and comorbidities in sparse data. *J Am Med Inform Assoc*. 2009 Jul-Aug;16(4):561-70.
- Uzuner O. Second i2b2 workshop on natural language processing challenges for clinical records. *AMIA Annu Symp Proc*. 2008:1252-3.
- Vishwanathan SVN, Schraudolph NN, Schmidt MW, et al. Accelerated training of conditional random fields with stochastic gradient methods. *ICML*. 2006 969–76.
- Wang H, Zhang W, Zeng Q, Li Z, Feng K, Liu L. Extracting important information from Chinese Operation Notes with natural language processing methods. *J Biomed Inform*. 2014 Apr;48:130-6.
- Wang X HG, Friedman C. Characterizing environmental and phenotypic associations using information theory and electronic health records. *BMC Bioinformatics*. 2009 10:S13.

- Wang X HG, Markatou M, et al. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *J Am Med Inform Assoc*. 2009 16:328-37. .
- Wang X. Translational Medicine is developing in China: A new venue for collaboration. *J Transl Med*. 2011 Jan 4;9:3.
- Wang Y, Yu Z, Chen L, et al. Supervised methods for symptom name recognition in free-text clinical records of traditional Chinese medicine: an empirical study. *J Biomed Inform*. 2014 Feb;47:91-104.
- Wang Y, Yu Z, Jiang Y, Xu K, Chen X. Automatic symptom name normalization in clinical records of traditional Chinese medicine. *BMC Bioinformatics*. 2010;11:40.
- Weeber M MJ, Aronson AR. Developing a Test Collection for Biomedical Word Sense Disambiguation. *proc AMIA Symp*. 2001 746-50.
- Weng C PP, Velez M, Johnson SB, Bakken S. Towards symbiosis in knowledge representation and natural language processing for structuring clinical practice guidelines. *Stud Health Technol Inform*. 2014 201:461-9.
- Wikipedia. [http://en.wikipedia.org/wiki/Natural\\_language\\_processing](http://en.wikipedia.org/wiki/Natural_language_processing).
- Wilcox A HG. Medical text representations for inductive learning. *Proc AMIA Symp* 2000 923-7.
- Wu ST, Juhn YJ, Sohn S, Liu H. Patient-level temporal aggregation for text-based asthma status ascertainment. *J Am Med Inform Assoc*. 2014 May 15.
- Xiaoshan Fang HS. A Hybrid Approach for Chinese Named Entity Recognition. *Discovery Science. Lecture Notes in Computer Science* 2002 Volume 2534, 2002, pp 297-301.
- Xu H AK, Grann VR, Friedman C. Facilitating cancer research using natural language processing of pathology reports. *Medinfo* 2004. 11(1):565–572.
- Xu H JM, Oetjens M. et al. Facilitating pharmacogenetic studies using electronic health records and natural-language processing: a case study of warfarin. *J Am Med Inform Assoc*. 2011 18(4):387-91. doi: 10.1136/amiajnl-2011-000208.

- Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. *J Am Med Inform Assoc.* 2010 Jan-Feb;17(1):19-24.
- Xu H, Stetson PD, Friedman C. A study of abbreviations in clinical notes. *AMIA Annu Symp Proc.* 2007:821-5.
- Xu Y WY, Liu T, Tsujii J, Chang EI. An end-to-end system to identify temporal relation in discharge summaries: 2012 i2b2 challenge. *J Am Med Inform Assoc.* 2013 Sep-Oct;20(5):849-58. doi: 10.1136/amiajnl-2012-001607. Epub 2013 Mar 6.
- Y. Wang YLZYL, Jiang Y. A preliminary work on symptom name recognition from free-text clinical records of traditional chinese medicine using conditional random fields and reasonable features. in *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing.* 2012 Stroudsburg, PA, USA, pp. 223–230.
- Y. Xu YWTLYFYQJT, Chang EI. Joint segmentation and named entity recognition using dual decomposition in Chinese discharge summaries. *J Am Med Inform Assoc.* 2013 vol. doi:10.1136/amiajnl-2013–001806.
- Yao K, Peng B, Zweig G, Yu D, Li X, Gao F. Recurrent conditional random field for language understanding. [http://research.microsoft.com/pubs/210167/rcrf\\_v9.pdf](http://research.microsoft.com/pubs/210167/rcrf_v9.pdf)
- Zhao J LF. Product named entity recognition in Chinese text. *Language Resources and Evaluation.* 2008 Volume 42, Issue 2, pp 197-217.
- Zheng J SW. The current status of NLP research and application in China. *Digital Library Forum.* 2008 DOI:10.3772/j.issn.1673-2286.2008.07.005.
- Zheng Lvexing LX, Liu Kun, Du Yuncheng. Recognition of Chinese Personal Names Based on CRFs and Law of Names. *Web Technologies and Applications. Lecture Notes in Computer Science 2012 Volume 7234,* pp 163-170.
- Zhu D WS, Carterette B, Liu H. Using large clinical corpora for query expansion in text-based cohort identification. *J Biomed Inform.* 2014 49:275-81. doi: 10.1016/j.jbi.2014.03.010. Epub 2014 Mar 26.

Zou Q CW, Morioka C, et al. IndexFinder: a method of extracting key concepts from clinical texts for indexing. AMIA Annu Symp Proc. 2003 763-7.

目录:

- 一， 研究目的
- 二， 研究步骤
- 三， 标记方法

## 一， 研究目的

---

研究“中文医学名词实体识别”，通过人工标注医学文档（电子病历）中的名词实体得到“标准的”“医料库”，利用已有的机器学习算法，来学习这些语料，最后用这个训练后的算法来识别临床文档中的名词，最终可以：发现和对比同一种用于英文医学术语识别的算法用在中文临床术语识别上的准确性。或者找到中文临床术语识别的特点和 NER 的挑战。

我们会选用电子病历中的入院记录部分以及出院小结部分，一方面因为这两部分内容重要，另一方面，这两部分医学术语非常集中，第三方面相对独立。

## 二， 研究计划和标注具体步骤说明

---

- 1> **Stage I--Pilot Study:** 项目负责人或 PI 熟悉软件、标注过程及规则。从所有病历中随机挑选 10 篇，然后具体负责项目的人按照标注说明中的规则来标记，主要目的是为了熟悉规则
- 2> **Stage II--Training:** 选择标注者、培训标注者、确认标注合格。标注者可以是两位医生，为了保证标注的质量，由具体负责项目的人向他们讲解规则，并从病历中随机挑选几篇进行标注，然后进行指导。标注者熟悉后，可以让标注者分别独立标注 2 篇，经项目负责人确认合格后（90%标注准确）正式开始标注。
- 3> **Stage III—Formal annotation:** 从病例中随机挑选 400 篇，并从中挑选 40 篇作为公共的文档库，并将剩下的 360 篇随即分成两份，每一份 180 篇，每位医生需要标记 180 篇加上公共的 40 篇，总共 220 篇。标记完成后，可以算出两位医生所标注的公共部分的相似度，作为衡量标注质量的指标之一

注：

1. 第二步中对医生的培训很关键，要确保医生对标注规则理解透彻。
2. 标注的内容和方法需要研究者 PI 的培训，同时参见文档“病历名称标注说明”。

## 三， 概念（名称）标记方法的说明

---

在病历文本中，我们将对下列四种概念（名称）进行标记，所有名称均为名词或名词短语，（所有概念里不包含动词）：

### (一). 病名和问题称（Problem）

Problem 是个广义的概念，不仅仅是诊断，主要包括：疾病、症状、体征、异常检查结果。要尽量体现“整体性”，即包括 modifiers，但 assertion（如 no, possible 等不标记）大的 NE 包括小的 NE。（In i2b2 guideline, articles are even included），具体标记时，凡出现下列情况之一的可被标记：

1. 疾病名称（“糖尿病”，“慢性阑尾炎”，“结核”，“猩红热”，“老年性脑萎缩”）
2. 发病症状（“发热”，“恶心”，“咳嗽”，“呕吐”，“寒战”）
3. 身体功能不正常（“双眼黑蒙”，“大小便失禁”，“呼吸困难”，“乏力”）
4. 精神方面的不正常状态（“焦虑”）
5. 身体部位 + (形容词) + 症状（“上腹持续性胀痛”，“下肢肿大”，“鼻中隔偏曲”）
6. 病毒和细菌阳性，（如“MRSA”，乙肝携带者，等）
7. 物理损伤（“胳膊脱臼”，“骨折”）
8. 缺陷（“新生儿缺陷”）
9. 不正常检验结果（“血糖偏高”，“血 Rt 高”，“WBC 高”）

注意事项：

1. 否定词后的 **problem** 要标记？：症状或者疾病的名称前面出现否定词，虽然该症状或者疾病并没有发生（“无发热”，“无腹痛”，“无双眼黑蒙”，“无意识障碍”，“无大小便失禁”），仍然要标记，但要去除否定词，暨标记为：“发热”，“腹痛”，“双眼黑蒙”，“意识障碍”，大小便失禁”。容易忽略的是“未闻及干湿罗音”，应标记成“未闻及<干湿罗音-**problem**>”。“未闻及病理性音”，应该“未闻及<病理性杂音-**problem**>”
2. **Boundary** 问题--前修饰词或定语是否标记在 **problem** 中：这里可以分成几种情况：
  - a. “疾病诊断前”的修饰词或定语，因为已经俗成，应该全标，不能只标记“诊断”如：结核性胸膜炎、冠状动脉粥样硬化性心脏病、不稳定性心绞痛，等



- b. “异常症状体征”前的修饰词或定语，标准的如“部位 location+程度、性质 severity+频率 frequency+problem”（“上腹持续性胀痛”，“下肢肿大”，“鼻中隔偏曲”，“间断干咳”，“右侧胸腔积液”），要是没有标点符号隔开的前置定语，应该全标记。
- c. “异常症状体征”前的修饰词或定语，“非标准的，口水话式的修饰词”。特殊的修饰词，比如不是 location, severity, frequency 这种比较标准类型，例如诱因+problem：“非喷射性呕吐”、“活动后出现-喘憋”、“活动后胸闷”、“间断出现-夜间平卧时胸闷”、“静息时仍间断感胸闷”、“平卧时咳嗽”、“老年性主动脉退行性变”等诱因+problem，只要没有标点符合相隔仍然需要全部标记。
- d. 极端的例子，“<骶尾部可见约 4×4cm 皮肤破溃-problem>，已结痂”
- e. “<偶有干呕>” “该表体位时伴<腹部不适>”

3. Boundary 问题--后修饰词或补语是否标记在 problem 中（考虑整体性，problem 里可包含诸多属性，有标点分割则明确不用标记在一起）：

- a. “明确的 problem+变化”，只标记 problem。如“<右侧胸腔积液>减少”、“<喘憋>减轻”、“<咳嗽><咳痰><喘憋>进行性加重”、“<咳嗽>频繁伴<肩胛区疼痛>”、“<慢性阻塞性肺病>急性加重”、“<疼痛>无向肩背部放射”、“<胰腺炎>较重”等，以前是标记在一起，建议不标记后修饰词。另一个好例子“<脑栓塞病>史”，-史不被标记。“<糖尿病>病史”，病史不被标记
- b. “非 problem 暨正常症状体征脏器+异常变化”，而是“体征+异常变化”，一定需要整体标注，因为是合起来才是 problem，跟上述例子“problem+减弱”，本身就是 problem 了，不一样，比如“<右肺呼吸音低>”、“<心脏稍增大>”、“<左房大>”、“<动脉硬化>”、“<室间隔厚>”、“<左心输出功能降低>”、“<左室松弛功能减低>”、“<双肾实质内多发低密

度结节>”、“尿中有少量蛋白”、“患者<小便疼痛>逐渐加重”、“血糖控制不佳”

- c. 修饰词置后，对 **problem** 的某些属性进行描述，即使是关于 **location**，严重程度，发生频率等特征，均不标记在 **problem** 内（多个补语时，只标第一个，并且不能包含标点符号）。比如：“<甲状腺右叶明显增大-**problem**>，密度不均”，“<呕吐-**problem**>，<呕吐>物为胃内容物，非喷射性。”。

4. 描述一般病情的时候，我们经常会说：“<精神差>，<精神弱>，<食欲差>”，建议应该标记。
5. “患者述<平静或睡眠时也有心前区疼>”。这里诱因,且前置，为一“整体”。
6. “夜间不能平卧”，整体是个“**problem**”
7. 病史部分问题：“否认<心脑血管肾等慢性病>史”，“史”不被标记；“否认<药物及食物过敏>病史”，过敏不是特定的病（如“糖尿病”），“病史”不被标记；“否认明确<毒物接触>史”“<火碱接触>”；“预防接种史不详”，是正常预防接种，不是 **problem**，不被标记。否认<手术-**procedure**><外伤-**problem**>史。“过敏史：<双黄连-**medicine**>”
8. 体格检查部分问题：还有一个语序和体格检查的问题（否定词“无”在前面，还是语句中）：“<双肺可闻及广泛哮鸣音及湿罗音>”；“双肺未可闻及广泛<哮鸣音>及<湿罗音>”阳性问题整体标注，阴性只标注问题！“<双肺叩诊过清音>!”，“<双下肢膝关节以下水肿>”
9. “无反跳痛及肌紧张”，应该标记“无<反跳痛-**problem**>及<肌紧张-**problem**>”；“未见胃肠型及蠕动波”，因标记为“未见<胃肠型-**problem**>及<蠕动波-**problem**>”；“无移动性浊音”，应该“无<移动性浊音-**problem**>”右下腹麦氏点无<压痛>，“<双下肺可闻及细湿罗音>”，“双下肺未闻及<明显湿罗音>”，“双下肢未见<明显水肿>”“无<畸形>”，“双下肢不<肿>”，“<病理反射>未引出”，“<体型肥胖>”，“各浅表淋巴结未触及<肿大>”，“心前区无<隆起>”，“全腹

无<包块>”，“肠鸣音无<亢进>或<减退>”，“<双肺叩诊过清音>!”，“<双下肢膝关节以下水肿>”“肥胖体型”。---体格检查有很多模棱两可地方，需要确定一下，否则影响一致性。否定词作为分界线用。

10. 有些词，不管出现在哪里都是 **problem**，只不过实际标记时候有可能标记更大的范围，如<血栓>，<包块>，<结节>，<出血>，（巩膜无）<黄染>，（双下肢无）<水肿>，（双唇无）<紫绀>，（心前区无）<隆起>，（胸廓无）<畸形>，（肠鸣音无）<亢进>或<减退>，（全腹未扪及）<包块>，<移动性浊音>，<胃振水音>，

11. “否认明确毒物接触史”->“否认明确<毒物接触-problem>史”，“史”不标记！

12. 否认<吸烟>嗜好，<吸烟>60 年，<饮白酒>20 年？？

1. 并发症状描述，每个症状均需标注。“<呕吐-problem>时<腹部牵涉性痛-problem>”  
“<肺部感染-problem>合并<心衰-problem>”
2. 当一个 NE 被另一个 NE 包含时，取最长的 NE 做标注。

例：“葡萄糖溶液输注”，Medicine“葡萄糖”被包含在Procedure“葡萄糖溶液输注”中时，只标注Procedure。

13. “<反复咳嗽><喘憋>”，“开始出现<咳嗽><咳痰>”→应该分开，是两个 problems!

14. “<咳少量粘痰>”

15. 关于过敏：“<对溴隐亭过敏>”整体标记！

16. “血酮体（+）”是一个 Problem，还是强调整体！不是 labtest

17. “<神志明显好转>”是一个 Problem，“<呕吐>缓解”是一个 Problem “<肝肾功能进行性恶化>”，“<胸闷><憋气>无明显好转”

18. “<进食后突发中上腹痛-p>，为<持续性胀痛-p>，伴<呕吐>数次，<呕吐>后症状无缓解，<<腹痛>范围逐渐扩大至全腹部”

19. “<肺部可见大片阴影>，呈颗粒状”视为标记一个 problem。但顿号连接的句子要区别对待（比如“头晕”、“眼花”、“耳鸣”则分开标记；“大腿、小腿、上腹多处烧伤”则标记为一个），逗号的不标记在一起（原则：有标点的尽量不标记在一起！）“<双肺纹理增粗、模糊-problem>”

20. 广义的 problem，也都加吧：病情危重，预后极差，生命危险，死亡，去世

## (二). 药名称(Medicine)

只包含药物名称，不包含剂量，服用方法等（“阿司匹林”，“格华止”）

注意：

- 1, 中药冠心舒合，只标记“冠心舒合”
- 2, 予+药物治疗，标记为药物，而不是治疗！

## (三). 检验名称(Test)

只包含检验名称，不包含结果（“B超”，“血常规”，“肝功能”）

如果检查名称后面带有检查结果，无论正常与否，都只需要标出名称即可，如“血糖偏高”要标记成 problem（“Labtest 异常”规则），而“肝功能检查无异常”中标记“肝功能检查”

注意：

- 1, 检验结果中的项目如何标记？如“TG

18.8mmol/l”中“TG”是一个Labtest“尿蛋白：0.3g/L”中“尿蛋白”是一个Labtest，都需要标注成labtest

- 2, 查血常规，查尿常规，等中的“查”字不包括。

- 3, <血压> 120 80; <脉搏> 80次

<肝功能>检查？

#### (四).手术和处置名称(TREATMENT: PROCEDURE, INTERVENTION)

治疗过程中除药物，检验外的部分（“补液”，“解痉”，“抗感染”）

注意：

1. 除了手术治疗方法，很多治疗措施procedure说法都不标准，比如给予某药治疗，你是标记整个治疗，还是标记药物？两者的语义应该完全不一样的。比如：“予<头孢呋辛钠-medicine>、<抗感染-procedure>、氨茶碱、<平喘-procedure>、<降压-procedure>、<降脂-procedure>、<抗血小板-procedure>、<扩冠-procedure>及<利尿>治疗-procedure”中的“予头孢呋辛钠”是整体标记为procedure？这是真实含义，还是只标记“头孢呋辛钠”为medicine？若为前一种标记方法，则这个例子中的氨茶碱，也不好标记为procedure？真实的语义和句式结构无法都照顾。建议：有“予”的，还是整体标记，后面分开的“氨茶碱”，只能标记为药物，这样虽然与语义不完全吻合，但是计算机容易一致，容易学习。--  
还是不要包括予，药物就是药物，药物就暗含是一种治疗，不在这个层面上解决歧义。
2. “予”，“舌下含服”，“治疗”等不标记在procedure内！如：“<抗血小板>治疗”、“<抗炎>治疗”、“<平喘>治疗”、“<抗心衰>治疗”，“<支气管扩张药-medicine><雾化-procedure>治疗”、“<四联抗痨>治疗”，尤其是“院外继续<雾化>治疗”，“治疗”要包括吗？
3. “予<抗感染药物-medicine>后好转”，medince是一种procedure所以，能细到药物就药物。
4. 除了“予+药物治疗”外（“予”，“行”，“治疗”不包括在内？，但却是很重要的提示词，比如“予<头孢呋辛钠>”，“予<利尿>等治疗”），予<限盐><限水>，“舌下含服+<硝酸甘油>”，怎么办？

5. Boundary的确定：上述例子中“扩冠及利尿治疗”，治疗不要标记在内：正确方法为“<扩冠-procedure>及<利尿-procedure>治疗”
6. “行<左足趾外翻截骨矫形术>”中的行不标记，因为后面的手术名称是标准的；相反，“予头孢呋辛钠治疗”，不要标记“予”，不标记治疗，也即“<头孢呋辛钠-medicine>治疗”
7. 和problem一样，治疗方案中的一些性质是否包含在“procedure”中，如“院外患者家庭<氧疗>及规律<雾化>治疗”
8. “行<空肠营养管置入术-procedure>，逐步过渡<肠内营养-procedure>。<胰功-labtest>逐渐降至正常”
9. “继续<肠内营养-procedure>，检测<血糖-labtest>，根据<血糖-labtest>调整<胰岛素-medicine>使用。”！
10. “停用<中药-medicine>”
11. “<诊断性腹穿-procedure>，抽出咖啡色浑浊液体，化验提示<炎性渗出液-problem>”。
12. “用<善宁-medicine>对症治疗”中的动词“用”是否需要标注？
13. “平素服<阿斯匹林-medicine>、<中成药-medicine>。”如何标注？
14. “<舌干燥-problem>、<脱屑-problem>”
15. “肝区外敷敷料，可见<PTCD引流液红色-problem>”
16. “<心外心脏按压-procedure>数分钟”
17. “查<心肌酶-labtest>”不标注动词“查”？
18. “予以<可达龙-medicine>控制心率”如何标注？“予以<极化液-medicine><稳定心肌细胞>”呢？

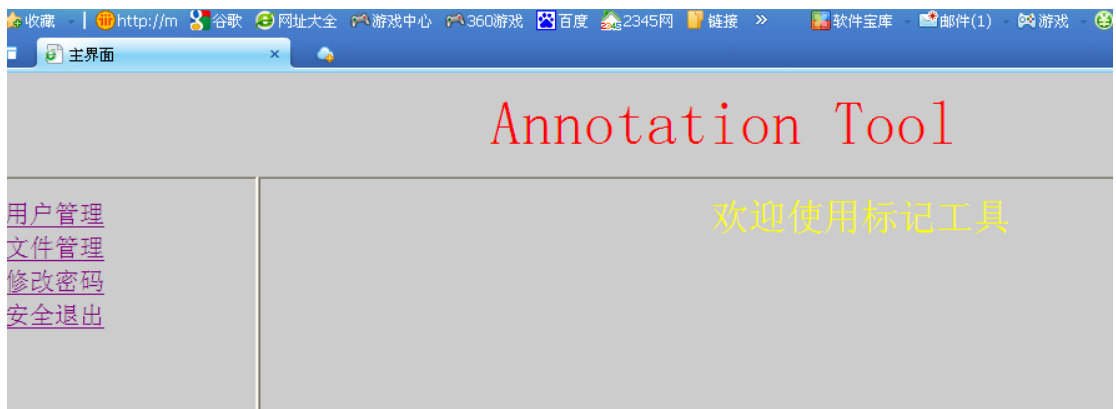
19. “加用<可达龙-medicine>0.2bid\*1周→0.2qd<控制心率>”如何标注？
20. “<心电图-labtest>示<胸前导联T波倒置-problem>，<QT间期延长-problem>”如何标注？
21. “<摔倒致右髋部及右下肢疼痛-problem>、<后动障碍-problem>”如何标注？
22. “爱人<因心血管病去世-problem>”，虽然看似不很正式，但可归纳为“诱因+problem主题”
23. “<心音弱-problem>，<率不齐-problem>”如何标注？
24. “继续<抗癫痫-procedure>、<抗感染>等对症支持治疗”如何标注？
25. “检测<电解质-labtest>及<血气-labtest>情况”，监测，查，予，用等如何标注？
26. “不适随诊”，是常见的治疗方案的一部分，整体标记为“procedure”。

## Appendix B: Major functions and screenshots of annotation tool

### 1). Login

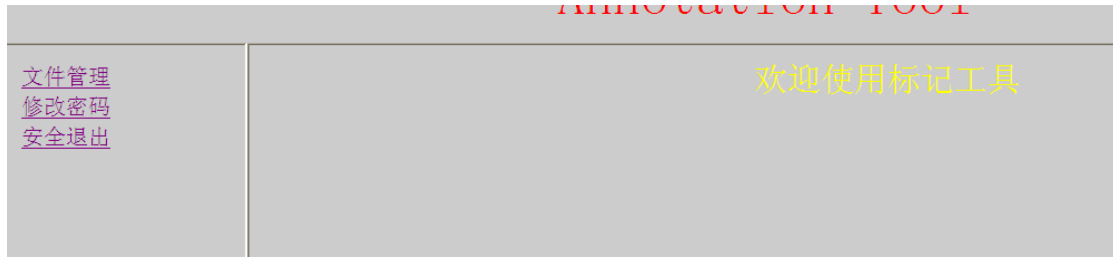


### 2). Page after administrator login:





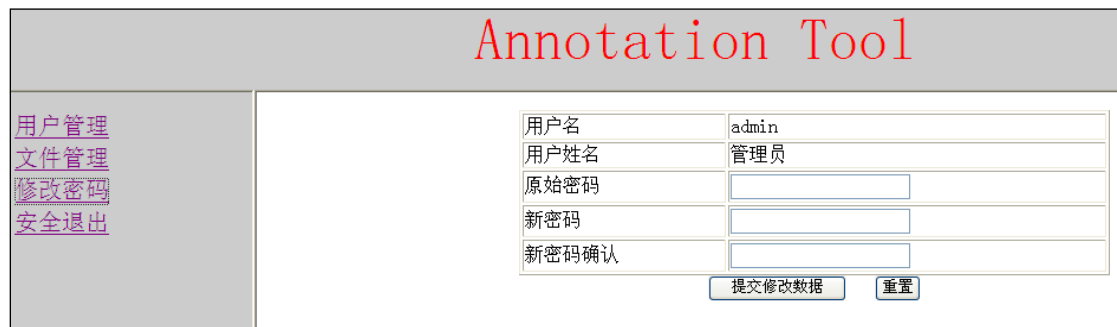
### 3). Page after common users login



### 4). Add, modify and delete users



### 5). Change users' password



## 6). File management(uploading files to be annotated, assigning users to files, etc)

Annotation Tool

[用户管理](#)  
[文件管理](#)  
[修改密码](#)  
[安全退出](#)

选择上传文件:

文件ID:  文件名:  标记状态:  标记人:

共找到6条记录, 显示所有记录.

文件编号	文件名	文件路径	标记人	标记状态	操作
23	新建 文本文档 (4).txt	D:\apache-tomcat-5.5.28\webapps\Annotation\Upload\新建 文本文档 (4).txt	test	已标记	分配 标记 删除
22	记录.txt	D:\apache-tomcat-5.5.28\webapps\Annotation\Upload\记录.txt	user	已标记	分配 标记 删除
21	新建 文本文档.txt	D:\apache-tomcat-5.5.28\webapps\Annotation\Upload\新建 文本文档.txt	My	已标记	分配 标记 删除
20	新建 文本文档 (3).txt	D:\apache-tomcat-5.5.28\webapps\Annotation\Upload\新建 文本文档 (3).txt	My	已标记	分配 标记 删除
19	新建 文本文档 (2).txt	D:\apache-tomcat-5.5.28\webapps\Annotation\Upload\新建 文本文档 (2).txt	test	已标记	分配 标记 删除
18	test.txt	D:\apache-tomcat-5.5.28\webapps\Annotation\Upload\test.txt	user	已标记	分配 标记 删除

## 7). Common annotator can only view his files to be annotated

Annotation Tool

[文件管理](#)  
[修改密码](#)  
[安全退出](#)

选择上传文件:

文件ID:  文件名:  标记状态:  标记人:

共找到3条记录, 显示所有记录.

文件编号	文件名	文件路径	标记人	标记状态	操作
22	记录.txt	D:\apache-tomcat-5.5.28\webapps\Annotation\Upload\记录.txt	user	已标记	标记
21	新建 文本文档.txt	D:\apache-tomcat-5.5.28\webapps\Annotation\Upload\新建 文本文档.txt	user	已标记	标记
18	test.txt	D:\apache-tomcat-5.5.28\webapps\Annotation\Upload\test.txt	user	已标记	标记

导出方式: ☒ Excel ☐ XML

## 8). Uploading files

Annotation Tool

[用户管理](#)  
[文件管理](#)  
[修改密码](#)  
[安全退出](#)

选择上传文件:

文件ID:  文件名:  标记状态:  标记人:

共找到6条记录, 显示所有记录.

文件编号	文件名	文件路径	标记人	标记状态	操作
23	新建 文本文档 (4).txt	D:\apache-tomcat-5.5.28\webapps\Annotation\Upload\新建 文本文档 (4).txt	test	已标记	分配 标记 删除
22	记录.txt	D:\apache-tomcat-5.5.28\webapps\Annotation\Upload\记录.txt	user	已标记	分配 标记 删除
21	新建 文本文档.txt	D:\apache-tomcat-5.5.28\webapps\Annotation\Upload\新建 文本文档.txt	My	已标记	分配 标记 删除
20	新建 文本文档 (3).txt	D:\apache-tomcat-5.5.28\webapps\Annotation\Upload\新建 文本文档 (3).txt	My	已标记	分配 标记 删除
19	新建 文本文档 (2).txt	D:\apache-tomcat-5.5.28\webapps\Annotation\Upload\新建 文本文档 (2).txt	test	已标记	分配 标记 删除
18	test.txt	D:\apache-tomcat-5.5.28\webapps\Annotation\Upload\test.txt	user	已标记	分配 标记 删除

选择要加载的文件

查找范围 (Q): 桌面

Recent  
 桌面  
 我的文档  
 我的电脑  
 网上邻居

千千静听  
 无标题.bmp  
 新建 Microsoft Word 文档 (2).doc  
 新建 Microsoft Word 文档 (3).doc  
 新建 Microsoft Word 文档 (4).doc  
 新建 Microsoft Word 文档.doc  
 新建 文本文档 (2).txt  
 新建 文本文档 (3).txt  
 新建 文本文档 (4).txt  
 新建 文本文档.txt  
 需解决的技术点.txt  
 字体背景.html

文件名 (N): 新建 文本文档 (4).txt

## 7). Assigning files to different annotators.

共找到4条记录, 显示所有记录.

1

用户编号	用户姓名	用户类型	操作
admin	管理员	admin	选择
My	我	admin	选择
test	测试	user	选择
user	标记员	user	选择

## 8). Click “annotation”

选择上传文件:

文件ID:  文件名:  标记状态:  标记人:

共找到7条记录, 显示所有记录.

1

文件编号	文件名	文件路径	标记人	标记状态	操作
24	mytest.txt	D:\apache-tomcat-5.5.28\webapps\Annotation\Upload\mytest.txt		新文件	分配 标记 册
23	新建 文本文档 (4).txt	D:\apache-tomcat-5.5.28\webapps\Annotation\Upload\新建 文本文档 (4).txt	test	已标记	分配 标记 册

## 7). Annotating concepts

[管理](#)  
[管理](#)  
[密码](#)  
[退出](#)

×

Welcome

这是测试文档请选择要标记的文字

Tag Name:

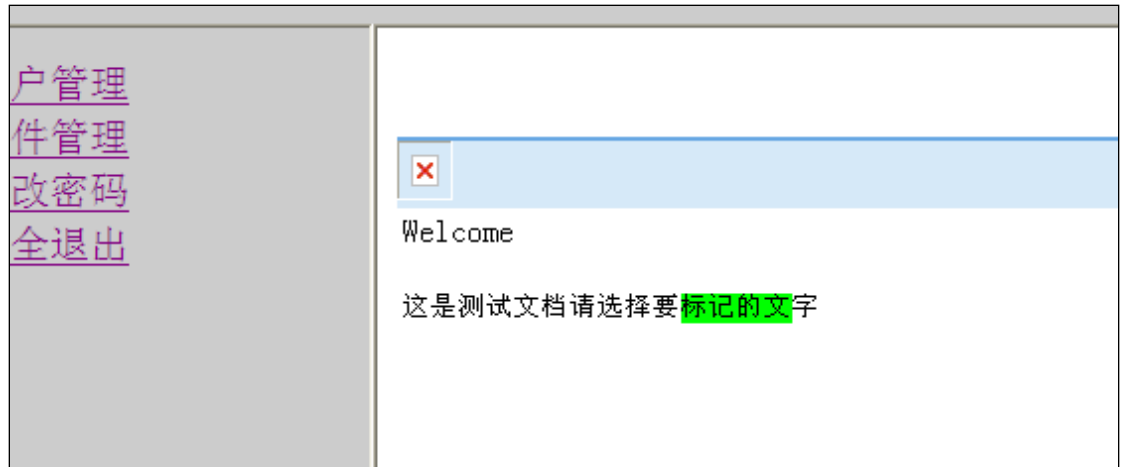
☐ Problem
☒ Medicine
☐ Labtest
☐ Procedure

Selection:

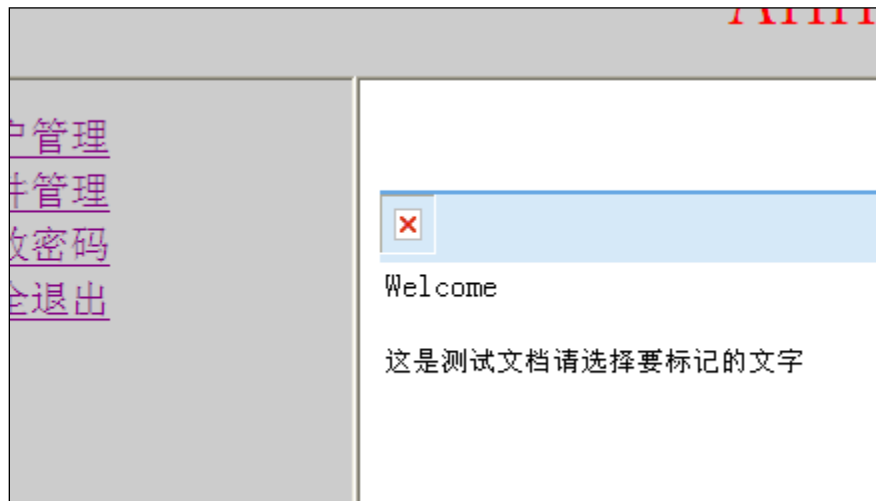
Position:

LineNumber:

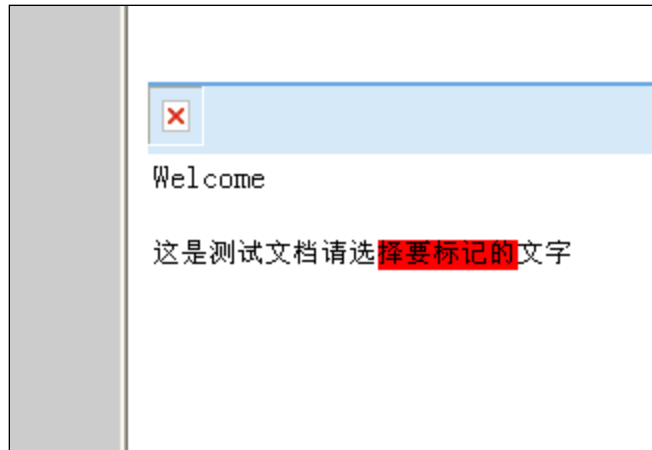
## 8). Texts are highlighted after annotation



#### 9). Deleting annotation



#### 10). Change annotation type



11). deleting

