Lexical Characteristics Analysis of Chinese Clinical Documents

Meizhi Ju
College of Biomedical Engineering and Instrument Science, Zhejiang University

Haomin Li*
The Children's Hospital of Zhejiang University
The Institute of Translational Medicine, Zhejiang University

Huilong Duan
College of Biomedical Engineering and Instrument Science, Zhejiang University

Abstract

Understanding lexical characteristics of clinical documents is the foundation of sublanguage based Medical Language Processing (MLP) approach. However, there are limited studies focused on the lexical characters of Chinese clinical documents. In this study, a lexical characteristics analysis on both syntactic and semantic levels was conducted in a clinical corpus which contains 3,500 clinical documents generated during daily practices. The analysis was based on the automatic tagging results of a lexicon-based part-of-speech (POS) and semantic tagging method. The medical lexicon contains 237,291 entries annotated with both semantic and syntactic classes. The normalized frequency of different terms, syntactic and semantic classes was calculated and visualized. Major contribution of this paper is providing a wide-coverage Chinese medical semantic lexicon and presenting the lexical characteristics of Chinese clinical documents. Both of these will lay a good foundation for sublanguage based MLP studies in China.

1 Introduction

Clinical documents which contain a tremendous amount of patient information to facilitate inter-provider communication, also the most important part of clinical data for secondary use such as clinical research and administration. Recent advance in MLP technologies (Sager et al., 1987;Sager et al., 1994; Friedman and Hripcsak, 1999; Liu et al., 2012; Irina P. Temnikova et al ,2013; Irina P. Temnikova et al ,2014; Pham et al., 2014), such as de-identification (Meystre et al., 2014; Kayaalp et al., 2014), text classification (Pestian et al., 2007; Vijay, 2012), information retrieval (Uzuner et, al., 2010; Zhu et al., 2013), etc., affords an opportunity to study and analyze clinical documents at an unprecedented scale.

In recent years, Chinese MLP topics have drawn increasing public attention as there are more and more electronic clinical data that major exist in free text format such as clinical documents and reports were accumulated in many hospitals. Some Chinese MLP studies have been reported such as information extraction (Wang et al., 2014), NER(Named Entity Recognition) (Lei et al., 2014). However, systematic studies of lexical characters of Chinese clinical documents, that is the foundation of sublanguage based MLP approach and have been widely studied in other language (Foltz, 1996; Wu and Liu, 2011; Patterson and Hurdle, 2011; Patterson et al., 2010; Friedman et al., 2002), are seldom reported. Lack of accessibility of clinical documents corpus and comprehensive lexical resources for the research community is the major obstacle.

Both syntactic and semantic lexical features are important to understand the medical language structure and grammar (Harris 1968;1991). However, studying lexical features in both syntactic and semantic levels in large scale corpus requires a comprehensive medical lexicon to support the automatic lexical tagging process (Meystre et al., 2008). Unfortunately, such lexical resources in Chinese are not available. In this study, we constructed a 237,291 entries Chinese medical lexicon using computer aided methods at first. Then a lexical analysis which aims to present syntactic and semantic features of Chinese clinical

Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015), pages 114–120,
Beijing, China, July 30, 2015.

documents was conducted in a corpus contains 3,500 clinical documents. The lexical features of clinical documents from different departments were reported. The annotated corpus was ready for further utilization such as collection of the co-occurrence patterns (Grishman et al., 1986) and sublanguage grammar.

2 Methods

To understand the lexical characters of language used in a subdomain, a large-scale corpus contains typical language samples from the real word need to be constructed at first. Then this corpus should be annotated manually or automatically with part-of-speech (POS) tags and semantic tags. Then the statistical analysis based on these tagging results will help researchers to understand the features of this type of sublanguage.

2.1 Corpus Collection

The corpus was collected from an EMR system which implemented in a 2000-beded hospital in China. More than 60,000 clinical documents were generated from 2009 to 2011 in total 35 clinical departments. Randomly selected 100 clinical documents from each department were used to construct a corpus for this study. Total 5 document types were included in the 3,500 clinical documents which contain 152,393 sentences and 2,375,909 Chinese characters. In addition, 15 clinical documents were randomly selected and manually annotated as the test set to evaluate the coverage of the lexicon as well as the performance of lexcial tagging methods.

2.2 Lexicon Construction

A general purpose dictionary which used in an open-source Chinese word segmenter Pangu (http://pangusegment.codeplex.com) constituted the basic of this lexicon. While most of the total 146,259 lexemes from this general purpose dictionary are irrelevant to medical concepts. ICD-10, a medication lexicon which was acquired from (http://yao.dxy.com/) using web crawler technology, and a home-grown lexicon were also compiled into this lexicon. Total 237,291 lexemes were included in this lexicon. Learning from the classical Linguistics String Project (LSP) (Grishman et al., 1973), total 24 semantic categories were designed (Listed in Table 1). POS tags were directly inherited from the Pangu systems. Semantic attribute annotations of lexcion were achieved using both statistical method and syntactic rule based method. Medical domain specialty terms such as ICD-10, medication dictionary that with known semantic class will be annotated in batch during their enrollemnts. Some semantic class with obvious morphology was assigned through matching key character of the lexeme. For example, if a character ends with "病" ("disease") with POS attribute "noun", its semantic class will be annotated as "Diagnosis" for further manual review. The ambiguity of semantic classes of many lexemes was resolved based on the most frequently usage in the corpus.

Semantic class	Example	Count
Basic Information	年龄"age"	127
Body Part	脖颈 "neck"	7,411
Nursing Care	常规护理 "nursing routine"	2,212
Chemical	硫酸 "sulfuric acid"	114
Description	交通事故 "traffic accidents"	1,282
Device	呼叫设备 "calling device"	1,618
Diagnosis	肺癌 "lung cancer"	30,209
Document Type	入院记录 "admission notes"	213
Examination	X 射线检查"X-ray examination"	2,066
Expense Name	诊疗费"medical fee"	587
Department	急诊科"emergency department"	155
Irrelevant	法案"law"	146,280
Lab Test	血清总胆固醇测定 "serum total choles-terol determination"	4,544
Medical Entity	医生"doctor"	93
Medication	阿司匹林"aspirin"	20,818
Number	多"more"	55
Organism	血吸虫"schistosome"	959
Phy Function	呼吸"breath"	281
Surgery	骨髓穿刺术"bone marrow puncture"	8,345
Symbol	$,&	303
Symptom	眩晕"dizziness"	4,681
Time	早上"morning"	1,976
Treatment	治疗方案 "therapeutic regimen"	1,340
Unit	pmol/L	236

Table 1: Semantic classes defined in the lexicon.

In addition, semantic class of lexemes with irrelevant POSes such as "Chinese idiom" was tagged as "Irrelevant". Furthermore, lexemes

which are not processed with the mentioned approaches were annotated manually.

2.3 Tokenization and Annotation

Supported by the constructed lexicon, the tokenization and annotation of the corpus were conducted in the following steps. Firstly, each clinical document in the corpus with extra space was automatically trimmed in the pre-process. Then a punctuation-driven sentence boundary detection algorithm was applied to obtain sentences and clauses. After that, all clauses were segmented into words or phrases using a Chinese lexical analyzer ICTCLAS (Zhang et al., 2003). Both the semantic and syntactic classes were annotated for each word or phrase based on the lexicon during this process. For words or phrases without semantic attributes in the lexicon will be annotated as "Unknown". To make it simple, all the symbols, Arab numbers and punctuations that without specific meanings were all removed.

2.4 Lexical Characteristics Analysis

A statistical frequencies of different lexical categories in different condition were calculated. As shown in Formula 1, a NF (Normalized Frequency) value was normalized as the count of this type of lexemes in every 10,000 lexemes used in the background. As different categories with significant difference NF values, the logarithm of NF (LoF) will be calculated to plot the diverse values easier (Shown in Formula 2).

$$NF = \frac{N_{Category} * 10000}{N_{Total}} \quad (1)$$

$$LoF = \begin{cases} \log(NF) & , \mathrm{NF} \geq 1 \\ 0 & , \mathrm{NF} < 1 \end{cases} \quad (2)$$

In Formula 1, the $N_{Category}$ indicated the count of lexemes with specific semantic or syntactic category attribute in corpus or subset of corpus. The N_{Total} represented the total number of lexemes in the same corpus. The LoF value will be set to 0, when there are seldom observation of some category in some subset of corpus.

3 Results

3.1 Evaluation of the Lexicon Coverage and Lexical Tagging Methods

The quality of the lexical characters generated from statistical analysis depends on the coverage and completeness of the lexicon constructed. Comparing with the typical comprehensive medical lexical resources such as UMLS which contains millions of terms, our lexicon scale is relatively small. So we calculate the coverage and completeness of the lexicon during the tokenizing and annotation. Total 13,660 lexemes were unrecognized among all 2,375,909 lexemes in the corpus. The coverage of our lexicon in the corpus was 99.43% calculated by Formula 3. Similarly, the distinct lexemes among the unrecognized lexemes and lexemes in the corpus were 577 and 19,847 respectively. Thus, the completeness of the lexicon was 91.11% calculated by Formula 4.

$$Coverge = \frac{N_{Unrecognized\ lexemes}}{N_{Total}} * 100\% \quad (3)$$

$$Completeness = \frac{N_{Unrecognized\ distinct\ lexemes}}{N_{Total\ distinct\ lexemes}} * 100\% \quad (4)$$

Based on the manually annotated test set, we evaluated the accuracy of word segmenter performance and syntax and semantics classification. Word segmentation and annotation regarding POS and semantics were conducted on the test set with the ICTCLAS. As a result, 4,006 lexemes were obtained excluding punctuations and Arabics by the automatic tagging process. Manually checking by one of the authors, the number of error segments caused by ICTCLAS was counted. Meanwhile, the number of lexemes with error POS tag or semantic tag was picked out. The accuracy of word segmentation, POS and semantics was calculated separately by Formula 5 and demonstrated in Table 2.

Evaluate item	Accuracy
Word Segmentation(ICTCLAS)	96.03%
POS	88.09%
Semantics	90.86%

Table 2: The evaluation result of the lexicon.

3.2 Lexical Characters in Chinese Clinical Documents

The semantic class of lexemes usage frequency (NF value) in different clinical departments was plotted in Fig. 1 using heatmap.2 function gplots package in R. It is apparent from the heat map that "body part", "time", "symptom" and "diagnosis" were the top four semantic classes. We can easily distinguish the mental health department from other departments as the "body part" was used in a relatively lower frequency. Some internal medicine department such as rheumatology, hematology and nephrology more interested in the lab test result discussion.

The fluctuation of 22 POS categories in 5 typical document types in Fig. 2.A is basically consistent in general. However, there are observable

differences between semantic categories as showed in Fig. 2.B. For example, document type of informed consents has great differences compared with other types of clinical documents.

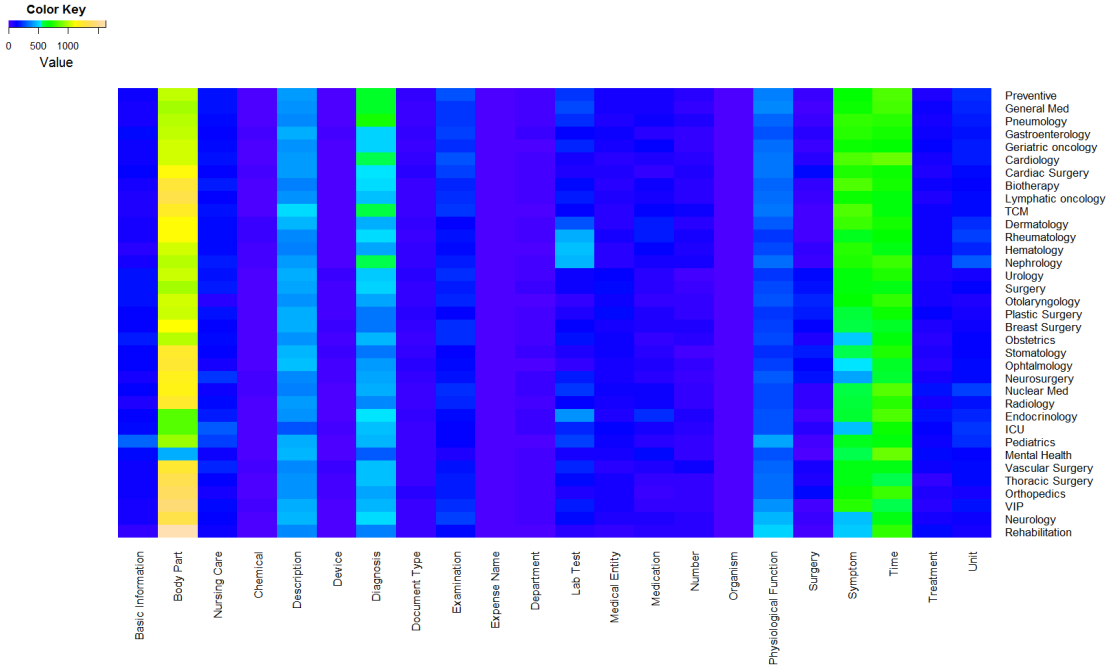

Fig.1. Heat map of original NF value.

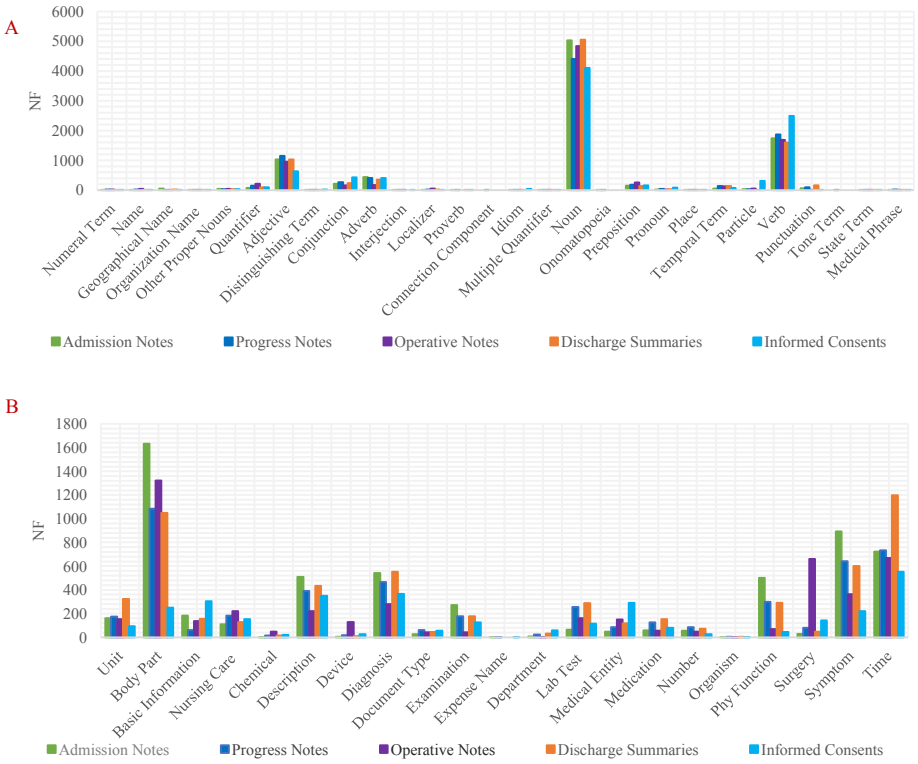

Fig. 2. Sublanguage (A) and POS (B) features of 5 document types in corpus.

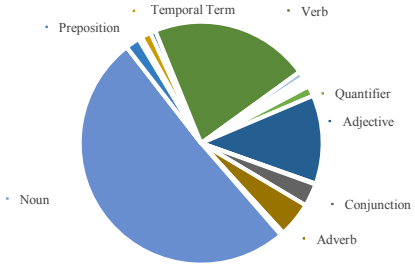

Fig. 3. The POS proportion of Chinese clinical documents.

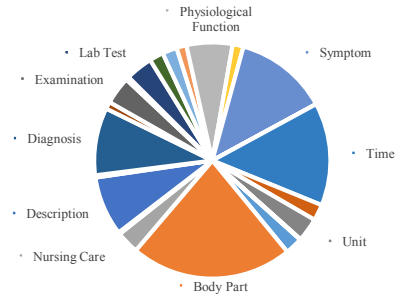

Fig. 4. The sublanguage proportion of Chinese clinical document.

We can also notice that large number of phrases related to "time" were used in discharge summaries, implying that these retrospective documents record many temporal information. Fig. 3 and Fig. 4 show the overall LoF proportion of semantic and POS types in the corpus. All the figures lead us to the conclusion that body part, symptom and diagnosis sublanguages account for the largest portion of Chinese clinical documents.

3.3 Co-occurrence patterns in Chinese Clinical Documents

Furthermore, more than 168,823 nonrepeating clauses were obtained in the corpus including total 565,630 clauses. To count the sematic patterns among these clauses, some frequently used co-occurrence patterns were summarized in Table 3. For each pattern, the example clause was highlighted with different font colors and styles to show corresponding semantic component.

These co-occurrence patterns will lay a foundation to create sublanguage grammars for the Chinese medical language.

Co-occurrence pattern	Sample	Count
Body Part +Irrelevant +Symptom	*心前区*无**隆起**"no **uplift** in *precordium*"，	12,740
Irrelevant +Symptom	为白色粘痰" is white sticky sputum"，	8,588
Body Part +Description	*颈部*对称"*the neck* is symmetrical"，	7,679
Irrelevant +Diagnosis	考虑**脑瘤**"possibly suffer **brain tumor**"，	4,877
Body Part +Diagnosis	*颈椎***肿瘤**"*Cervical* **Cancer**"：	4,278
Number +Body Part+Description +Symptom	双 *下肢*轻度水肿"two *lower extremities* mild edema"；	3,161

Table 3: Top co-occurrence patterns in the corpus.

4 Discussion and Future Work

In this paper, through constructing a comprehensive medical semantic lexicon, the lexical characteristics of clinical documents both in semantics and syntactic level were analyzed separately. In addition, a number of the most frequent sublanguage co-occurrence patterns of Chinese clinical documents were discovered.

The quality of the lexicon constructed in this study is the major challenge of current analysis. As a mature and high-quality lexical resource such as UMLS will take years and cost millions of dollars to maintain. A Chinese counterpart is urgently needed and its value should be well recognized by governments and funding agencies.

Our future work includes improving the coverage and quality of the lexicon based on the corpus using more computer aided approaches. The accuracy of the automatic tagging process still has plenty of room to improve. Currently most of the errors were caused by ambiguous of semantic type or POS. But the results of this lexical analysis still provide much useful information to Chinese medical language researchers.

Lack accessibility of corpus is one of the obstacles for current Chinese medical language processing studies due to current regulation and privacy concerns. As the automatic de-identification methods already widely accepted in many countries, we will evaluate it in our corpus in the future. After that this annotated corpus will open to the community.

Reference

Anne-Dominique Pham, Aurélie Névéol, Thomas Lavergne, Daisuke Yasunaga, Olivier Clément, Guy Meyer, Rémy Morello and Anita Burgun. 2014. *Natural Language Processing of Radiology Reports for the Detection of Thromboembolic Diseases and Clinically Relevant Incident Findings.* BMC Bioinformatics,15:266.

Carol Friedman and George Hripcsak. 1999. *Natural Language Processing and its Future in Medicine*. Journal of the Association of American Medical Colleges.

Carol Friedman, Pauline Kra and Andrey Rzhetsky. 2002. *Two Biomedical Sublanguages: A Description Based on the Theories of Zellig Harris.* Journal of Biomedical Informatics. 222-235.

Dongqing Zhu, Wu Stephen, Masanz James, Ben Carterette and Hongfang Liu. 2013. *Using Discharge Summaries to Improve Information Retrieval in Clinical Domain.*

Garia, Vijay. 2012. *Kernel Methods and Semantic Techniques for Clinical Text Classification*.

Huaping Zhang, Hongkui Yu, Deyi Xiong and Qun Liu. 2003. *HHMM-based Chinese Lexical Analyzer ICTCLAS.* 2nd SIGHAN workshop affiliated with 41th ACL, Sapporo Japan.

Hui Wang, Weide Zhang, Qiang Zenf, Zuofeng Li, Kaiyan Feng and Lei Liu. *Extracting Important Information from Chinese Operation Notes with Natural Language Processing Methods.* Journal of Biomedical Informatics 48 (2014) 130-136.

Hongfang Liu, Stephen T. Wu, Dingchen Li, Siddhartha Jonnalagadda, Sunghwan Sohn, Kavishwar Wagholikar, Peter J. Hang, Stanley M. Huff and Christopher G Chute. *Towards a Semantic Lexicon for Clinical Natural Language Processing.* AMIA Annual Symposium, 2012.

Irina P. Temnikova, Ivelina Nikolova and William A. Baumgartner Jr. *Closure Properities of Bulgarian Clinical Text.* In Proceedings of RANLP. 2013, 667-675.

Irina P. Temnikova, William A. Baumgartner Jr., Negacy D. Hailu, Ivelina Nikolova, Tony McEnery , Adam Kilgarriff, Galia Angelova and K. Bretonnel Cohen. *Sublanguage Corpus Analysis Toolkit: A Tool for Assessing the Representativeness and Sublanguage Characteristics of Corpora.* In Proceedings of LREC. 2014, 1714-1718.

John P. Pestian, Christopher Brew, Pawel Matykiewicz, Dj J. Hovermale, Neil Johnson, Kevin B. Cohen and Wlodzislaw Duch. 2007. *A Shared Task Involving Multi-label Classification of Clinical Free Text.* Biological, translational, and clinical language processing, pages 97-144, Prague, Association for Computational Linguistics.

Jianbo Lei, Buzhou Tang, Xueqin Lu, Kaihua Gao, Min Jiang and Hua Xu. 2014. *A Comprehensive Study of Named Entity Recognition in Chinese Clinical Text.* J Am Med Inform Assoc, 21:808-814.

Mehmet Kayaalp, Allen C. Browne, Zeyno A. Dodd, Pamela Sagan and Clement J. McDonald. 2014. *De-identification of Address, Date, and Alphanumeric Identifiers in Narrative Clinical Reports.* AMIA Fall Symposium.

Naomi Sager, Carol Friedman and Margaret S. Lyman. 1987. *Medical Language Processing: Computer Management of Narrative Data*.

Naomi Sager, Margaret S. Lyman, Christine Bucknall, Ngo T. Nhan and Leo J. Tick. 1994. *Natural Language Processing and the Representation of Clinical Data.* .

O¨zlem Uzuner, Imre Solti and Eithon Cadag. 2010. *Extracting Medication Information from Clinical Text.* J Am Med Infor Assoc, 17:514-518.

Olga Patterson and John F. Hurdle. 2011. *Document Clustering of Clinical Narratives: A Systematic Study of Clinical Sublanguages.* AMIA Annual Symposium Proceedings, 1099–1107.

Olga Patterson, Sean Igo and John F. Hurdle. 2010. *Automatic Acquisition of Sublanguage Semantic Schema: Towards the Word Sense Disambiguation of Clinical Narratives.* AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium, 612–6.

Peter W. Foltz. 1996. *Latent Semantic Analysis for Text-based Research.* Behavior Research Methods, Instruments &Computers, 28(2),197-202.

Ralph Grishman, Lynette Hirschman and Ngo T. Nhan. 1986. *Discovery Procedures for SubLanguage Selectional Patterns: Initial Experiments.* Compusational Lingustics.

Ralph Grishman, Naomi Sager, C. Raze and B. Bookchin. 1973. *The Linguistic String Parser.* Proceedings of national computer conference and exposition p427-434.

Stéphane M. Meystre, Óscar Ferrández, F. Jeffrey Friedlin , Brett R. South, Shuying Shen and Matthew H. Samore. 2014. *Text De-identification for Privacy Protection: A Study of its Impact on Clinical Text Information Content.* Journal of Biomedical Informatics, 50: 142-150.

Stephane M. Meystre, Guergana K. Savova, Karin C. Kipper-Schuler and John F. Hurdle. 2008. *Extracting Information from Textual Documents in the Electronic Health Record: A review of Recent Research.* IMIA

Stephen Wu and Hongfang Liu. 2011. *Semantic Characteristics of NLP-extracted Concepts in Clinical Notes vs.Biomedical Literature.* AMIA Annu Symp Proc. 1550–1558.

Zellig S. Harris. 1968. *Mathematical Structures of Language.* Interscience Publishers.

Zellig S. Harris. 1991. *A Theory of Language and Information: A Mathematical Approach.* Clarendon Press.