International Journal of u- and e- Service, Science and Technology
Vol.9, No. 9 (2016), pp.305-318
http://dx.doi.org/10.14257/ijunesst.2016.9.9.29

A Lexical and Syntactic Analysis System for Chinese Electronic Medical Record

Zhipeng Jiang, Xue Dai, Yi Guan and Fangfang Zhao

Department of Computer Science and Technology,
Harbin Institute of Technology, Harbin 150001, China
hit.jiang@hotmail.com

Abstract

Lexical and syntactic analysis, including word segmentation, part-of-speech (POS) tagging, shallow parsing and full parsing, are essential for medical language processing (MLP). However, research on full parsing, even shallow parsing and POS tagging for Chinese electronic medical record (CEMR), has not been carried out because of the lack of annotated corpus on CEMR. In this paper, we built a corpus of 5,024 sentences from CEMR with word segmentation, POS tags and phrase tags, of them, 2,553 are annotated as full parsing trees. Inter-annotator agreement results: Chinese word segmentation (97.56%), POS tagging (93.34%), shallow parsing (96.5%), full parsing (91.22%). A lexical and syntactic analysis system for CEMR is developed and evaluated based on above corpus. Of its components, we proposed a joint model for word segmentation and POS tagging with the transformation-based error-driven model as correction postprocessing to alleviate the problem of error accumulation, the F1-score of word segmentation and POS tagging were 94.39% and 93.2%, respectively. A shallow parsing model under the framework of group learning we proposed was developed, which enriched word features by word embedding from large unlabeled CEMRs and achieved the F1-score of 96.3%. At last, we presented a state-of-art full parser combining the Berkeley parser and the Stanford parser to outperform the best single parser by 3.68%. The evaluation results show a substantial benefit to statistical machine learning models from the annotated CEMR. These works are the foundation for natural language processing (NLP) technologies applied to CEMR.

Keywords*: CEMR, Chinese word segmentation, part-of-speech tagging, shallow parsing, full parsing*

1. Introduction

Clinical free text in electronic medical record (EMR) contains rich medical knowledge, but difficult to be exploited because of its complexity and variety. NLP technologies, especially lexical and syntactic analysis, have been successfully used to analyze and structured free text in many domains. With the rapid development of medical informatization, more and more attention has been drawn to applications of NLP technologies in EMR. Many MLP systems appeared and mainly focused on medical information extraction, they were designed for the specific application and achieved satisfactory results on some specific tasks [1-3]. However, most of MLP systems still remain at the rule-based level, complex and general system is hard to build because of the lack of annotated corpus. This problem becomes more prominent in China since information extraction research for CEMR just started. Chinese word segmentation and POS tagging, as the basis of the framework of mainstream Chinese information extraction, are yet trained on the general domain corpus [4-6]. Syntactic analysis even has not been integrated into the Chinese medical information extraction system despite it has

ISSN: 2005-4246 IJUNESST

been proven to be able to improve the performance of medical information extraction [7-8].

To support higher-level MLP research in China, we built and evaluated a multi-stage annotated corpus. Four models based on machine learning were integrated to accomplish auto-processing from word segmentation to parsing for CEMR. To the best of our knowledge, this is the first comprehensive system designed to parse the free text in CEMR.

2. Background

Within the general domain, Chinese word segmentation and POS tagging in English under the frame of statistical machine learning have achieved high accuracy closing to manual annotation (about 98% and 97%). Syntactic analysis in English only gets F1 score of 91% due to its complexity. POS tagging and syntactic analysis in Chinese are about 5~6 percent behind English because of the variability of Chinese sentences. Nevertheless, these high-precision models usually require a large-scale annotated corpus for training. Corpus engineering plays an important role in statistical NLP development. Penn Treebank (PTB) [9] is a classic English general-domain corpus, covering technology, journalism, literature and other areas. Most English annotated guidelines of biomedical corpora are from PTB, just with varying degrees of modifications. For example, Pakhomov [10] completely follows the POS set of PTB, only specified some annotated rules of EMR, such as special symbols, drug name, dosages and foreign words. GENIA corpus [11] developed in 2005 consisting of 500 MEDLINE abstracts is a famous biomedical parsing treebank. It deletes some POS tags that do not appear in biomedical literatures, and increases prefix and suffix according to syntax.

Based on lots of annotated biomedical corpora written in English, biomedical-domain NLP systems have sprung up. Most of these systems aim at extracting clinical knowledge, but usually regarding lexical and syntactic analysis as important steps. Medical Language Extraction and Encoding System (MedLEE) [12] is an early MLP system, in which parser was designed to determine the structure of a sentence and interpret the relationships among the sentence elements. Health Information Text Extraction (HITEx) [13] is from Brigham and Women's Hospital and Harvard Medical School, containing components of sentence tokenizer, POS tagger and noun phrase finder. The clinical Text Analysis and Knowledge Extraction System (cTAKES) is trained and evaluated on the 273 clinical notes annotated with POS tags, shallow parses and named entity [14]. Xu [15] implemented a structured information extraction system and used SharpNLP, an open source NLP tool processing newspaper articles, to mark noun phrases and adjective phrases. Smith [16] generalizes the medical vocabulary to expand the POS set and presents a part-of-speech tagger based on a corpus of 5,700 manually tagged sentences. Fan [17] developed, evaluated, and shared a English clinical Treebank composed of 1,100 sentences in progress notes from 2010 i2b2/VA Clinical NLP Challenge, and achieved an accuracy of 0.811. The MiPACQ clinical corpus [18] is also taken from the Mayo Clinic EMR, but involves more comprehensive annotation, such as syntactic annotations, predicate-argument semantic annotations and the Unified Medical Language System (UMLS) entity semantic annotations. By training on this corpus, the performance of NLP components on English clinic text is boosted significantly.

In recent years, more and more researchers start to focus on NLP research in CEMR. They created the respective Chinese annotated clinical text so as to develop named entity recognition (NER) system [19-20]. There is no natural separator in a Chinese sentence, so Chinese word segmentation is an essential annotation at the

bottom. Xu [21] proposed a joint model using dual decomposition trained in 336 labeled Chinese discharge summaries to boost both word segmentation and named entity recognition, which was demonstrated to be superior to independent models, incremental models and a joint model trained on combined labels. Lei [22] annotated 400 admission notes and 400 discharge summaries from Peking Union Medical College Hospital in China, and investigated the effects of different types of feature including bag-of-characters, word segmentation, part-of-speech, and section information, and different machine learning algorithms. The system using structural support vector machines (SSVM) achieved the highest performance by combining word segmentation and section information.

Though MLP systems have appeared in China, there is still not any syntactic analysis system for CEMR, even Chinese annotated clinical text with phrase tags or POS tags. Additionally, the English annotation guidelines of clinical text can't be directly used because of the language gaps. These create barriers to the development of MLP in China.

3. System Description

Chinese clinical sentences without separators can be entered into the system, and processed by components on different levels in a pipeline manner (an example is shown in figure 1). The system is constituted by the following components:

- Tokenizer
- Part-of-speech tagger
- Shallow parser
- Full parser

The architecture of system shows a process combining automatic analyzing and semi-automatic annotating for Chinese plain clinical text (as in figure 2). Considering the complexity of full parsing annotation, we separated shallow parser from full parser to train and test by itself, so as to reduce the cost of annotation.

3.1. Corpus

The corpus was randomly sampled from CEMR of a large comprehensive Level-A hospital in China. It is a collection of 306 records embracing discharge summaries and progress notes. Of them, 70 are from the department of neurology, 68 are from the department of general surgery, 92 are from the cardiovascular department, and 76 are from the obstetrics department. Discharge summaries contain several sections of Discharge Instructions (DI), Treatment Effects (TE), Discharge Conditions (DC), Treatment Course (TC), Admission Conditions (AC), Clinical Definite Diagnosis (CDD), Clinical Initial Diagnosis (CID), Admitted Diagnosis of Clinic (ADC), Date of Admission/Discharge (DAD) and Patient Information (PI). Progress notes contain sections of Treatment Plan (TP), Differential Diagnosis (DD), Assessment (AS), Characteristics of Cases (COC), Clinical Initial Diagnosis (CID) and Chief Complaint (CC).

3.1.1. Preprocessing Scheme: For the particularity of data acquisition of CEMR, we proposed a preprocessing flow.

- Step 1. CEMR is usually stored in the medical record department as read-only picture, we need to convert them to text with Tesseract[1], a tool of Optical Character Recognition (OCR).
- Step 2. Annotators start to correct recognition errors.
- Step 3. Corrected text is encoded in Extensible Markup Language (XML) to facilitate manual annotation with graphical tools and evaluation on various sections of CEMR.

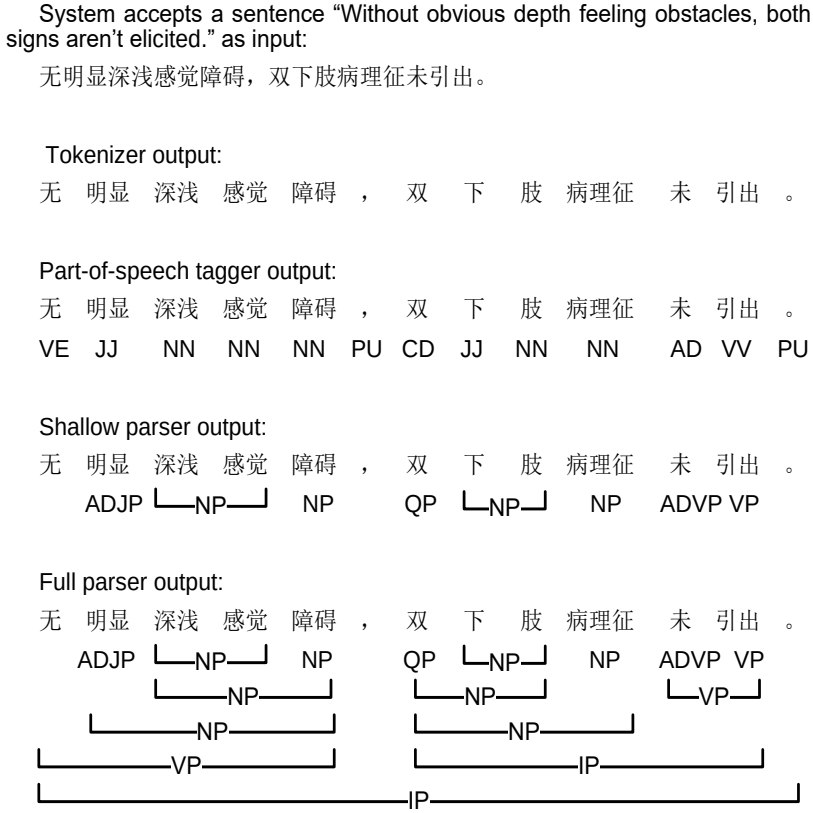


Figure 1. A Chinese Clinical Sentence Processed by the System

3.1.2. Annotation Scheme: The annotation scheme is a process adjusting guidelines for the CEMR and labeling with layers of Chinese word segmentation, POS tags and phrase tags.

- Step 1. To speed up annotation, automatic tools trained in general domain are first used in each layer of annotations, including ICTCLAS[2], Stanford POS tagger[3] and parser[4]. These tools will be replaced with our system in future annotation.
- Step 2. Annotators correct errors of automatic annotations independently.
- Step 3. Following an iterative approach combining the modification of guidelines with the evaluation of annotators' agreement, we discuss about the cases that annotators are inconsistent or uncertain, and consult doctors to confirm them.

[1] http://code.google.com/p/tesseract-ocr/
[2] http://www.ictclas.org/ictclas_download.aspx
[3] http://nlp.stanford.edu/software/tagger.shtml
[4] http://nlp.stanford.edu/software/lex-parser.shtml

- Step 4. After each correction and discussion, guidelines will be adapted until consistency stabilized at a high level. For instance, we simplified guidelines by omitting the cases do not appear in CEMR, detailed some rough strategies for patterns seen commonly in CEMR, and so on.

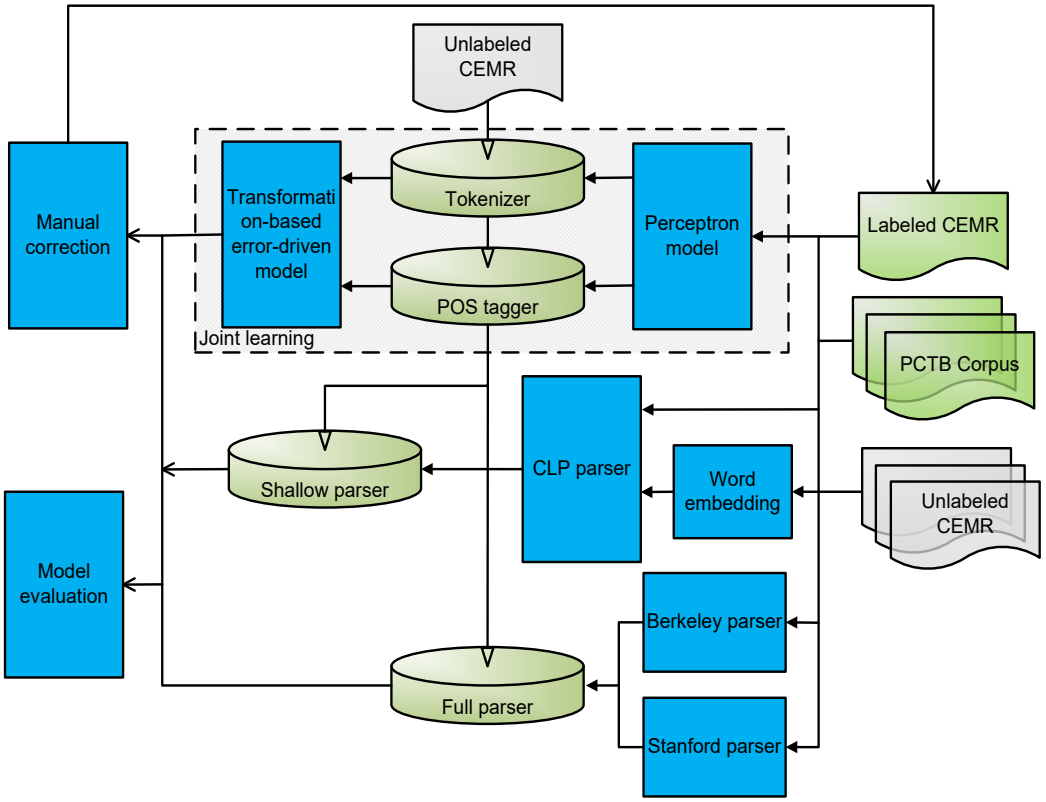


Figure 2. The Structure Diagram of the System

3.2. Tokenizer and POS Tagger

The corpus was randomly sampled from CEMR of a large comprehensive Level-A hospital in China. It is a collection of 306 records embracing discharge summaries and progress notes. The tokenizer and POS tagger are jointed by the online perceptron algorithm [23] under character-based encoding to realize Chinese word segmentation and POS tagging synchronously. The perceptron is an algorithm for learning a binary classifier: a function that maps its input x (a real-valued vector) to an output value y (a single binary value):

$$y = sign(w \cdot x + b),$$

where w is a vector of real-valued weights, $w \cdot x$ is the dot product $\sum_{i=0}^{m} w_i x_i$, and b is the bias. The online perceptron algorithm starts with an initial zero prediction vector $w = 0$. It predicts the label of a new instance x to be $\hat{y}$. If this prediction differs from the label y, it updates the prediction vector to $w = w + \phi(y) - \phi(\hat{y})$. If the prediction is correct then w is not changed. The process then repeats with the next example.

3.2.1. Joint Learning: With the character-based approach for joint Chinese word segmentation and POS tagging (S&T), each character is labeled with both its part-of-speech and BIO format (B, beginning of a word; I, inside a word; O, outside of a word).

Taking it as baseline, we first proposed the feature template for CEMR, as shown in Table 1, and used beam search algorithm to replace Viterbi search algorithm in decoding. Moreover, transformation-based error-driven model [25], a rule-based machine learning model, was introduced as a postprocessing to correct predicting errors. Similarly, we chose a more efficient training algorithm [26] instead, and proposed the transformation templates for CEMR in Table 2. In templates, 'c' and 'p' denote a character and its part-of-speech, respectively, subscripts are their relative location. For example, '$c_{-1}p_0$' represents the previous character and the part-of-speech of current character.

Table 1. The Feature Template of the Character-Based S&T Model

Category	Features
Unigram	$c_{-1}p_0$, c_0p_0, $c_{+1}p_0$
Bigram	$c_{-2}c_{-1}p_0$, $c_{-1}c_0\,p_0$, $c_0c_1p_0$, $c_1c_2p_0$
Transitional	$p_{-1}p_0$, $p_{-2}p_{-1}$, p_0

Table 2. The Transformation Template of the Transformation-Based Error-Driven Model

Change tag A to tag B when:
p_{-1}, $p_{-1}p_{-2}$, $p_{-1}p_1$, c_0, $c_{-1}c_0$, $c_{-2}c_{-1}$, $c_{-1}c_1$, c_0c_1, c_1c_2

3.3. Shallow Parser

The shallow parser is from the base chunking module of the CLP parser, a full parser developed in our previous study [27]. It is a model based on condition random field (CRF), improved by word-embedding features captured from large unlabeled CEMR under the framework of group learning we proposed. CRF [28] is a random field globally conditioned on the observations. By the fundamental theorem of random fields, the joint distribution over the label sequence Y given X has the form

$$p_\theta(y \mid x) \propto \exp\Big(\sum_{e\in E,k} \lambda_k f_k(e, y|_e, x) + \sum_{v\in V,k} \mu_k g_k(v, y|_v, x)\Big),$$

where x is a data sequence, y is a label sequence, and $y|_S$ is the set of components of y associated with the vertices in subgraph S, features f_k and g_k are assume to be given and fixed, the parameter estimation problem is to determine the parameters $\theta = (\lambda_1, \lambda_2, \ldots; \mu_1, \mu_2, \ldots)$ from training data.

3.3.1. Feature Expansion: Word embedding is the collective name for a set of language modeling and feature learning techniques in NLP where words are mapped to vectors of real numbers in a low dimensional space. Word2vec [29] is an open-source tool for computing continuous vector representations of words from very large data sets. These vectors have been proved to provide state-of-the-art performance on measuring syntactic and semantic word similarities.

On 3,634 CEMRs corrected and segmented by the tokenizer, we built 4,763 word representations using word2vec and clustered them using NLTK [5] toolkit. In our experiment, we took the Skip-gram algorithm to extract context features for word2vec, then clustered the word representations by the K-means algorithm to expand features to CRF model. The feature template of CRF model for CEMR was proposed in Table 3.

[5] http://www.nltk.org/

Table 3. The Feature Template of the Shallow Parseing Model

Category	Features
Unigram	w_0, w_{-1}, w_{-2}, w_{-3}, w_1, w_2, w_3, p_0, p_{-1}, p_{-2}, p_{-3}, p_1, p_2, p_3
Bigram	$w_{-3}w_{-2}$, $w_{-1}w_0$, w_0w_1, w_2w_3, $w_{-1}w_1$, $p_{-3}p_{-2}$, $p_{-1}p_0$, $p_{-2}p_{-1}$, p_0p_1, $p_{-1}p_1$, p_2p_3, w_0p_1, w_0p_2, p_0w_{-1}, p_0w_1
Trigram	$p_{-1}p_0p_1$, $p_{-2}p_{-1}p_0$, $p_0p_1p_2$, $w_{-2}p_{-1}p_0$, $p_0w_1p_1$, $p_{-1}w_0p_1$
Embedding	cn_{-2}, cn_{-1}, cn_0, cn_1, cn_2, $cn_{-1}cn_0$, cn_0cn_1
Bi-predictions	$pre_{-1}pre_0$

3.3.2. Group Learning: Taking advantage of 'section' in CEMR has become a new approach to optimize MLP model. Lei [22] utilized it as a feature and obtained a tiny improvement on NER task. Before building a shallow parser using the annotated CEMR, we trained the Stanford POS tagger and parser [30] on PCTB corpus and tested them on different sections of CEMR. The precision of POS tagging and the F1-measure of parsing as evaluation metrics were calculated using EvalB. The results are provided in Figure 3 to analyze the correlation between them, the Pearson correlation coefficient is computed as follows:

$$\rho = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}},$$

where cov(X,Y) is the covariance of X and Y, var(X) and var(Y) are their variances.

It is interesting that the two variables do not meet the positive correlation assumptions considered generally. They negatively correlate at -0.9 in several sections of the CDD, CID, ADC, DAD and PI sections, but the correlation of other sections is still high (0.99), such as the DI, TE, DC, TC and AC sections. We found that the text of sections with negative correlation are more structured, in which information is almost simply listed, lacking of rich context. This different writing style may cause the different statistical distribution. In the following experiment, we proposed group learning according to the above difference so as to improve the parser by enriching training corpus under the same distribution. Train and test corpus were both grouped by positive or negative correlation, then merged results from each group after independent test.

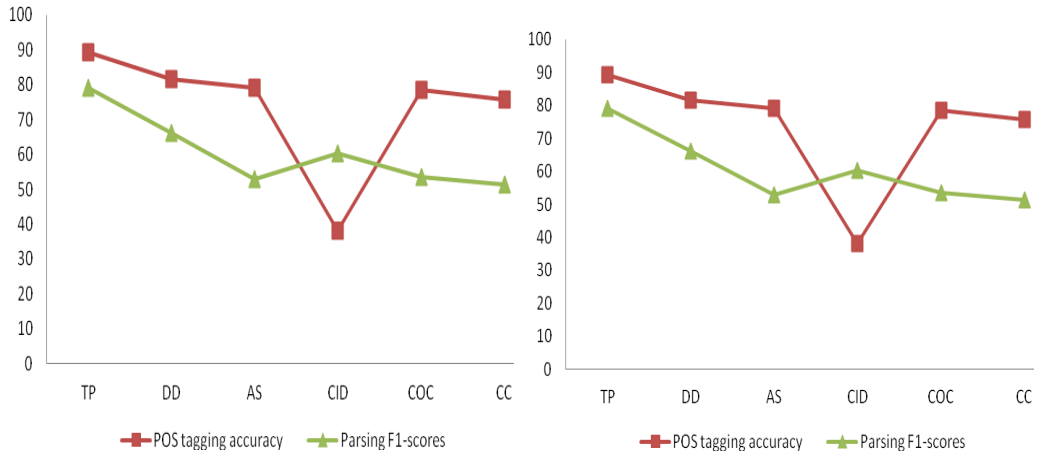


Figure 3. The Distribution of POS Tagging Accuracy and Parsing F1-Scores in Different Sections of Discharge Summaries and Progress Notes

3.4. Full Parser

The full parser is a wrapper around the Berkeley parser [31] and the Stanford parser, which combines two parsers trained on CEMR, respectively. The Berkeley parser and the Stanford parser are famous open parsing tools, they both are multilingual parsers and accept training corpus in PTB format. The Berkeley parser is based on unlexicalized parsing model, but the Stanford parser is an implementation of the factored model. The two parsers have been reported to get state-of-the-art performance in Chinese general domain and the Berkeley parser is better than the Stanford parser in the Tsinghua Chinese Treebank (TCT) [32]. Besides, the CLP parser [27], a rapid hierarchical parsing model with the collaborative correcting algorithm based on CRF, has also achieved almost the same parsing precision of the two parsers in TCT, in which full parsing is regarded as a recursive sequence-labeling process.

3.4.1. Parser Combination: Though the Berkeley parser in the best performance is claimed to be language-independent, it is unfriendly to Chinese language. Null outputs may occur for a Chinese sentence containing a full-width dash or some Chinese symbols, while the less training data are, the more null outputs are. This problem becomes even more severe, because special symbols are used frequently in CEMR. In order to compensate null outputs, we compared different strategies for double parsers combination, including the Berkeley parser, the Stanford parser and the CLP parser, to choose the best one to boost our full parsing module.

4. Experiments and Analysis

4.1. Corpus

The annotated corpus was created to evaluate our system, and train it as a automatic tool for future annotation, including 306 Chinese clinical notes (5,024 sentences) with POS tags and phrase tags and 138 Chinese clinical notes (2,553 sentences) in the format of full parsing tree.

To guarantee the quality of the corpus, annotators with linguistic background firstly needed to correct the same small corpus in each round, respectively, during all the stages of Chinese word segmentation, POS tagging and syntax annotation. Then, based on the gold corpus generated after discussing, IAA on Chinese word segmentation and POS tagging was computed according to the following formula in (1). EvalB[6], a famous tool on evaluation of parsing tree, was used to calculate IAA on syntax annotation.

$$IAA = \frac{\#\ of\ \ agreed\ \ tags}{\#\ of\ \ all\ \ tags} \times 100\% \tag{1}$$

In addition to the ambiguity of guidelines, the major source of disagreements on Chinese word segmentation is the lack of clinical knowledge. Our solutions are recording the uncertain cases to consult doctors and segmenting the terminology using large-grained size. For example, the word "上颌窦炎(maxillary sinusitis)" is not segmented to guarantee its integrity.

Disagreements on syntax annotation are mostly caused by the ellipsis of grammatical constituent. It is comforting that only a few ambiguities rely on clinical knowledge. For example, when the sentence "腹部平坦，无胃肠型及蠕动波(stomach is flat, without gastrointestinal type and peristaltic wave)" was annotated, the first annotator deemed "腹部(stomach)" as a subject of the whole sentence to annotate. However, the second annotator considered the clause "无胃肠型及蠕动波(without gastrointestinal type and

[6] http://nlp.cs.nyu.edu/evalb/

peristaltic wave)" with the ellipsis of subject , which may be patient. This choice depended entirely on annotators' knowledge background.

Disagreements caused by domain knowledge did affect the IAA to some extent. But, with increasing of iterations, IAAs were still on the rising trend. The averages shown in Table 4 are satisfactory and even closed to English clinical Treebank [18]. It implies that annotated experience can make up for the lack of domain knowledge. Annotators are ability to build a corpus with high quality.

Table 4. Average IAA in the First Three Iterations

Annotation Layer	Average IAA(%)
Chinese word segmentation	*97.56*
POS tagging	*93.34*
Shallow parsing	96.5
Full parsing	91.22

4.2. Tokenizer and POS Tagger

We compared our S&T system with related works in Table 5. To enlarge training data, we combined the PCTB corpus with randomly 80% of CEMR to train all models and remained 20% as the test data. It is clear that our system obviously outperforms other systems, particularly surpasses the word-lattice based model [33], a state-of-the-art S&T model in general domain, about 4% in both Chinese word segmentation and POS tagging. The encouraging results illustrate that the rule-based method is more applicable to CEMR. In other words, the assumption about CEMR with stronger grammatical regularity can be verified. Comparing to the pipe-line annotated framework, the joint annotated framework is still superior in CEMR.

Table 5. Comparison of Our S&T System and Related Works

POS tagger	Chinese word segmentation (F1-score %)	POS tagging (F1-score %)
Character-based	*90.15*	*88.73*
Word-lattice	*90.45*	*89.05*
Ours pipeline	*84.15*	*82.11*
Ours joint annotated	***94.39***	***93.20***

4.3. Shallow Parser

The annotated 5,024 sentences were split at random in 80/20 manner where 80% of the sentences were training set and 20% were testing set. For comparing the performance of sequence-labeling format and PCFG-based format, we evaluated shallow parsers on gold word segmentation. Results are reported in table 6. Because the Stanford parser and the Berkeley parser usually needs multistage PCFGs, they do not apply to shallow parsing, whose performances are much worse than the CLP parser, specially 210 null outputs generated to bring the greatest punishment to the Berkeley parser.

Table 6. Comparison of the Three State-of-the-Art Parsers in General Domain

Parser	POS tagging	Shallow parsing		
	Precision (%)	*Precision (%)*	*Recall (%)*	*F1-score (%)*
Berkeley parser	*92.9*	*62.29*	*64.76*	*63.50*
Stanford parser	*95.17*	*89.38*	*89.06*	*89.22*
CLP parser	97.85	95.73	95.82	**95.78**

Based on the CLP parser, word embedding features from clusters in different granularity are used for CEMR. As Table 7, shows, the numbers following 'K' represent the clustering number in K-means algorithm, combination of different word embedding features is written as 'K-combine'. All kinds of word embedding features slightly improved the performance of CLP parser except the combination method, 'K-800' provided the best increase (0.1%). The combined usage of different clustering granularity produced more noise in the model. Furthermore, bigger is not better in the clustering number. For example, the performance of K-1000 is worse than K-800.

Table 7. Performance of the CLP Parser with Different Word-Cluster Features

Parser	POS tagging	Shallow parsing		
	Precision (%)	*Precision (%)*	*Recall (%)*	*F1-score (%)*
CLP parser	97.85	95.73	95.82	95.78
CLP parser + K-500	*97.80*	*95.80*	*95.85*	*95.83*
CLP parser + K-800	97.85	95.88	95.88	**95.88**
CLP parser + K-1000	97.76	95.86	95.87	95.87
CLP parser + K-combine	97.71	95.69	95.76	95.72

Table 8 shows the performance of the CLP parser with best word embedding features in different usage of section information. Different from Lei [22], our results demonstrated that section information as feature ('CLP parser + K-800 + section') can't help the shallow parser. More important, group learning we proposed ('CLP parser + K-800 + group') can further enhance the model's performance greatly, finally outperformed the baseline by 0.52%.

Table 8. Performance of the CLP Parser in Different Usages of Section Information

Parser	POS tagging	Shallow parsing		
	Precision (%)	*Precision (%)*	*Recall (%)*	*F1-score (%)*
CLP parser	97.85	95.73	95.82	95.78
CLP parser + K-800	97.85	95.88	95.88	95.88
CLP parser + K-800 + section	97.77	95.853	95.86	95.86
CLP parser + K-800 + group	98.11	96.32	96.29	**96.30**

4.4. Full Parser

On building a full-parsing model, we took an initial study to train the Stanford parser and the Berkeley parser in the PCTB corpus and random 80% CEMRs, respectively, and test them on 20% CEMRs. Table 9, presents details of using different corpus to train models. The results using CEMR are significantly higher than PCTB even though the number of annotated sentences is few, indicating that in-domain data can bring a substantial amelioration.

Table 9. Comparison of Parsers in Different Training Data

Parsing (F1-score %)	PCTB	CEMR
Stanford parser	*53.83*	*80.38*
Berkeley parser	*53.95*	*80.63*

In above 80/20 data split, the results of single parser and combination parser were reported in Table 10. On small training corpus, the greater a parser relies on vocabulary, the worse its performance is. Among the three single parsers, the Berkley parser benefiting from unlexicalized model achieved the best F1-score of 80.63%, even it generated 38 null outputs. On the contrary, few training data leads to the worst performance of the CLP parser because it needs lexical features in most layers. We also grouped training set by the correlation assumption, but parsers' results were not much better due to fewer training data. In addition, to solve the problem of the Berkeley parser's null outputs, we compared different parser combination approaches to fill its null outputs. The best result is from the combination of the Berkeley parser and the Stanford parser, further improving the state-of-art full parsing F-score from 80.63% to 84.31%.

Table 10. Comparison of Different Parser Combination

Parser	POS tagging	Full parsing		
	Precision (%)	*Precision (%)*	*Recall (%)*	*F1-score (%)*
CLP parser	95.13	78.69	78.58	78.63
Stanford parser	*93.8*	*80.7*	*80.07*	*80.38*
Berkeley parser	*88.47*	*84.62*	*77*	*80.63*
Berkeley parser + CLP parser	*94.19*	*81.66*	*81.05*	*81.35*
Berkeley parser + Berkeley parser (PCTB)	*94.25*	*82.81*	*80.22*	*81.5*
Berkeley parser + Stanford parser	*95.23*	*84.49*	*84.14*	***84.31***

5. Discussion

In this paper, we investigated NLP models at different levels to develop a comprehensive lexical and syntactic analysis system for CEMR. We manually annotated 306 Chinese clinical notes (5,024 sentences) with POS tags and phrase tags and 138 Chinese clinical notes (2,553 sentences) in the format of full parsing tree, and achieved excellent IAA at various annotation levels, indicating linguists can annotate most of syntactic structure without much clinical knowledge. This conclusion is consistent with a previous study on the GENIA corpus [11].

Chinese word segmentation is the foundation of senior Chinese clinical NLP. Research on Chinese word segmentation for CEMR has started, for instance, Lei [22] took word segmentation as features, Xu [21] used dual decomposition for joint Chinese word segmentation and NER. Different from existed works, we combined a

character-based S&T model with transformation-based error-driven model to increase the F1-score of POS tagging based on auto-segmentation to 93.2%. For several main error sources, like "NN(noun) vs VV(verb)" and "NN(noun) vs VA(predicative adjective)", the error-driven model can resolve ambiguities effectively. For example, of the sentences "无头痛(without a headache)" and "头痛3天(have a headache for 3 days)", the POS of word "头痛(headache)" are NN and VV, respectively. A transformation rule "if the previous two characters are '无(without)' and '头(head)', the POS of word will be changed from VV to NN" can be applied to differentiate them, it will make the POS of word "头痛(headache)"in the first sentence to be a NN.

Lexical information is important for POS tagging and parsing models, however, our limited annotated corpus can't cover a number of clinical terms used frequently in CEMR. Word embedding is considered as an approach to relieve the lexical sparse. For example, the word " 扁桃体炎(amygdalitis)" does not exist in our annotated corpus, but the word " 甲**状**腺炎(thyroiditis)" can be found, and they are both belonged to the 127 cluster according to word embedding. This clustering feature can help to improve our POS tagging and parsing models. Moreover, the content of CEMR is split according to different sections, these section information could be represented in feature format, and used to enhance NER system [22]. Nevertheless, this usage of section information is no benefit to shallow parsing model. We proposed the framework of group learning by the correlation of different sections to achieve the best performance (F1-score of 96.3%) in shallow parsing.

Development of a clinical full parsing treebank is a very formidable task, we built the first syntactic treebank in CEMR, just consisted of 2,553 sentences. Under conditions that training data are not enough, we proposed the method of combination parser to make up the loss of parser's performance. The F1-score of our full parser has reached 84.31%, closed to the stat-of-art parser trained in Chinese general domain [34]. Parsing of long and flat sentences always is a difficulty in general domain, it has becomes one of the most frequent error sources in CEMR, because this sentence pattern is commonly used to describe symptoms. Segmentation on long sentence with punctuations and rules may be a solution to this problem.

6. Conclusion

In this study, we first presented a complete annotation scheme for CEMR, including text preprocessing, guidelines adaptation and agreement evaluation. To our knowledge, we built the first corpus consisted of CEMR both with lexical and syntactic annotations. Based on the corpus, the first lexical and syntactic analysis system for CEMR was implemented. In this system, an accurate S&T module with the transformation-based error-driven model was developed. The outperformance of the error-driven model implied the sublanguage domain of Chinese clinical text can make the rule-based method thrive again. Group learning by correlation and word embedding features was introduced to improve the performance of the shallow parsing module. Furthermore, we combined the Berkeley parser and the Stanford parser to be a state-of-art combination parser for CEMR.

In future, to resolve the problem of adaptation to different departments, we are planning to port semi-supervised model or transfer learning model to our system, and explore the semi-automatic annotation model like the active learning model to build corpus with larger scale and higher quality on CEMR.

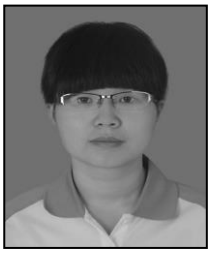
References

[1] Ö. Uzuner, I. Goldstein, Y. Luo and S. K. Isaac, "Identifying patient smoking status from medical discharge records", Journal of the American Medical Informatics Association, vol. 15, no. 1, **(2008)**, pp. 14-24.
[2] H. Xu, K. Anderson, V. R. Grann and C. Friedman, "Facilitating cancer research using natural language processing of pathology reports", Stud Health Technol Inform, vol. 107, no. 11, **(2004)**, pp. 565-72.
[3] M. Fiszman, W. W. Chapman, D. Aronsky, R. S. Evans and P. J. Haug, "Automatic detection of acute bacterial pneumonia from chest X -ray reports", Journal of the American Medical Informatics Association, vol. 7, no. 6, **(2000)**, pp. 593-604.
[4] F. Ye, Y. Y. Chen, G. G. Zhou, H. W. Li and Y. Li, "Intelligent Recognition of Named Entity in Electronic Medical Records", Chinese Journal of Biomedical Engineering, vol. 30, no. 2, **(2011)**, pp. 256 -262.
[5] X. Han and R. Ruonan, "The Method of Medical Named Entity Recognition Based on Semantic Model and Improved SVM-KNN Algorithm", Semantics Knowledge and Grid (SKG), 2011 Seventh International Conference on, IEEE, **(2011)**, pp. 21-27.
[6] L. Yi, B. Pengfei and X. Wanguo, "Research on Information Extraction of Electronic Medical Records in Chinese", Journal of Biomedical Engineering, (in Chinese), vol. 27, no. 4, **(2010)**, pp. 757 -762.
[7] B. D. Bruijn, C. Cherry, S. Kiritchenko, J. Martin and X. Zhu, "Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010", Journal of the American Medical Informatics Association, vol. 18, no. 5, **(2011)**, pp. 557–562.
[8] P. G. Mutalik, A. Deshpande and P. M. Nadkarni, "Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the UMLS", Journal of the American Medical Informatics Association : JAMIA, vol. 8, no. 6, **(2001)**, pp. 598–609.
[9] M. P. Marcus, M. A. Marcinkiewicz and B. Santorini, "Building a large annotated corpus of English: The Penn Treebank", Computational linguistics, vol. 19, no. 2, **(1993)**, pp. 313-330.
[10] S. V. Pakhomov, A. Coden and C. G. Chute, "Developing a corpus of clinical notes manually annotated for part -of-speech", International journal of medical informatics, vol. 75, no. 6, **(2006)**, pp. 418 -429.
[11] Y. Tateisi, A. Yakushiji and T. Ohta, "Syntax annotation for the GENIA corpus", Proceedings of Ijcnlp, **(2005)**.
[12] C. Friedman, "Towards a comprehensive medical language processing system: methods and issues", Proc AMIA Annu Fall Symp, **(1997)**, pp. 595-599.
[13] Q. T. Zeng, S. Goryachev, S. Weiss, M. Sordo, S. N. Murphy and R. Lazarus, "Extracting principal diagnosis, comorbidity, and smoking status for asthma research: evaluation of a natural language processing system", BMC Med Inform Decis Mak, vol. 6, no. 1, **(2006)**.
[14] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. K. Schuler and C. G. Chute, "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications", Journal of the American Medical Informatics As sociation, vol. 17, no. 5, **(2010)**, pp. 507-513.
[15] X. Yan, H. Kai, J. Tsujii and E. I. Chang, "Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries", Journal of the American Medical Informatics Association Jamia, vol. 19, no. 15, **(2012)**, pp. 824-832.
[16] L. Smith, T. Rindflesch and W. J. Wilbur, "MedPost: a part-of-speech tagger for bioMedical text", Bioinformatics, vol. 20, no.14, **(2004)**, pp. 2320-2321.
[17] J. W. Fan, E. W. Yang, M. Jiang, R. Prasad, R. M. Loomis, D. S. Zisook, J. C. Denny, H. Xu and Y. Huang, "Syntactic parsing of clinical text: guideline and corpus development with handling ill-formed sentences", Journal of the American Medical Informatics Association, vol. 20, no. 6, **(1900)**, pp. 1168-1177.
[18] D Albright, A Lanfranchi, A Fredriksen, W. F. Styler IV, C. Warner, J. D. Hwang, J. D. Choi, D. Dligach, R. Nielsen, J. Martin, W. Ward, M. Palmer and G. K. Savova, "Towards comprehensive syntactic and semantic annotations of the clinical narrative", Am Med Inform Assoc, vol. 20, no. 5, **(2013)**, pp. 922-930.
[19] S. Wang, S. Li and T. Chen, "Recognition of Chinese Medicine Named Entity Based on Condition Random Field", J Xiamen University (Natural Science), vol. 48, no. 3, **(2009)**, pp. 349–364.
[20] W. Hui, Z. Weide, Z. Qiang, L. Zuofeng, F. Kaiyan and L. Lei, "Extracting important information from chinese operation notes with natural language processing methods", Journal of Biomedical Informatics, vol. 48, no. 2, **(2014)**, pp.130–136.
[21] Y. Xu, Y. Wang, T. Liu, J. Liu, Y. Fan, Y. Qian, J. Tsujii and E. I. Chang, "Joint segmentation and named entity recognition using dual decomposition in Chinese discharge summaries", Journal of the American Medical Informatics Association, vol. 21, no. e1, **(2014)**, pp. e84-e92.
[22] J. Lei, B. Tang, X. Lu, K. Gao, J. Min and X. Hua, "A comprehensive study of named entity recognition in chinese clinical text", Journal of the American Medical Informatics Association, vol. 21, no. 5, **(2014)**, pp. 808-814.
[23] Y. Freund and R. E. Schapire, "Large margin classification using the perceptron algorithm". Machine Learning, vol. 37, no. 3, **(1999)**, pp. 277–296.

[24] D. R. Liou, J. W. Liou and C. Y. Liou, "Learning Behaviors of Perceptron", iConcept Press, **(2013)**.
[25] E. Brill "Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging", Computational Linguistic, vol. 21, no. 4, **(1995)**, pp. 543-565.
[26] Z. Ming, W. Jin and H. C. Ning, "A fast learing algorithm for part of speech tagging: an improvement on Brill's transformation-based alogrithm", Chinese, Computers, vol. 21, no. 4, **(1998)**, pp. 358-366.
[27] Z. Jiang, Y. Guan and X. Dong, "A Chinese Hierarchical Parsing Approach Based on Multi-layer Collaborative Correction", Journal of Chinese Information Processing, vol. 28, no. 4, **(2014)**, pp. 29-36.
[28] J. Lafferty, A. McCallum and F. C. N. Pereira, "Conditional random fields: probabilistic models for segmenting and labeling sequence data", Proceedings of the 18th International Conference on Machine Learning. ACM, **(2001)**, pp. 282–289.
[29] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient estimation of word representations in vector space", arXiv preprint arXiv:1301.3781, **(2013)**.
[30] D. Klein and C. Manning, "Fast exact inference with a factored model for natural language parsing", In Advances in Neural Information Processing Systems 15 (NIPS 2002), **(2003)**, pp. 3-10.
[31] S. Petrov and D. Klein, "Improved inference for unlexicalized parsing." In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, **(2007)**, pp. 404–411.
[32] X. Chen, C. Huang, M. Li and C. Kit. "Better Parser Combination." CIPS-ParsEval-2009, **(2009)**.
[33] K. Zhang and M. Sun, "Reduce Meaningless Words for Joint Chinese Word Segmentation and Part-of-speech Tagging," In Proceedings of CoRR, **(2013)**.
[34] Z. G. Wang and C. Q. Zong, "Phrase parses reranking based on higher-order lexical dependencies." Ruanjian Xuebao/Journal of Software, vol. 23, no. 10, **(2012)**, pp. 2628-2642.

Authors

Zhipeng Jiang, is a Ph.D candidate of Harbin Institute of Technology. His scientific interests include Chinese segmentation, part-of-speech tagging, parsing, and transfer learning for Chinese electronic medical record.



Xue Dai, is currently a master student of Harbin Institute of Technology. Her main research field is lexical and syntactic analysis for Chinese electronic medical record.

Yi Guan, is a professor and doctorial supervisor of Harbin Institute of Technology. His scientific interests include health informatics, intelligent information retrieval, natural language processing and cognitive linguistics.

