Hierarchical Maximum Pattern Matching with Rule Induction Approach for Sentence Parsing

Yi-Syun Tan, Yuan-Cheng, Chu, Jui-Feng Yeh*
Department of Computer Science and Information Engineering,
National Chiayi University,
No.300 Syuefu Rd., Chiayi City 60004, Taiwan (R.O.C.).
Ralph@mail.ncyu.edu.tw

Abstract

Chinese parsing has been a highly active research area in recent years. This paper describes a hierarchical maximum pattern matching to integrate rule induction approach for sentence parsing on traditional Chinese parsing task. We have analyzed and extracted statistical POS (part-of-speech) tagging information from training corpus, then used the related information for labeling unknown words in test data. Finally, the rule induction regulation was applied to extract of the structure of short-term syntactic and then performed maximum pattern matching for long-term syntactic structure. On Sentence Parsing task, our system performs at 44% precision, 53% recall, and F1 is 48% in the formal testing evaluation. The proposed method can achieve the significant performance in traditional Chinese sentence parsing.

1 Introduction

Recently, natural language processing has become one of the most essential issues in computational linguistics especially in human centric computing. In Chinese text processing, it is important to distinguish words significance in syntactic analysis. In order to comprehend the word significance, sentence parsing becomes one of the important techniques in the natural language understanding. The aim of sentence parsing is assigning a Part of Speech (POS) tag to each word and recognizing the syntactic structure in a given sentence. Therefore, it will help us to understand the text by correct sentence parsing by give the structure information.

For Chinese knowledge, there was a research on Categorical analyzing (Chinese Knowledge Information Processing Group, 1993). and then developed balanced Chinese corpora (Chen et al., 1996). The Sinica Treebank has been developed and released for academic research since 2000 by Chinese Knowledge Information Processing (CKIP) group at Academia Sinica (Huang et al., 2000; Chen et al., 2003), it under the framework of the Information-based Case grammar (ICG), a lexical feature-based grammar formalism, each lexical item containing both syntactic and semantic information

In word segmentation, Hidden Markov Models were used to solve word segmentation problem (Lu, 2005). Asahara et al. (2003) combined Hidden Markov Model-based word segment and a Support Vector Machine-based chunker for Chinese word segmentation. In later research, Goh et al.(2005) used a dictionary-based approach, and then apply a machine-learning-based approach to solve the segmentation problem.

In sentence parsing, there were two kinds of general methods, one was the statistical-based and the other was the rule-based. In rule-based, it wanted Expert knowledge and needed human labeling, but human labeling would not only produce a lot of problems but spent a lot of time. In rule-based approaches, Tsai and Chen (2003) showed that used context-rule classifier for part-of-speech tagging and performed better than Markov bi-gram model. In statistical-based, recently commonly used machine learning algorithm to solve it. For example, Support Vector Machine (SVM), Hidden Markov Model (HMM), Maximum Entropy (ME) and Transformation-Based Learning Algorithm (TBL) be used widely. However, single machine learning algorithm had not enough, in order to had better performance that usually combined different machine learning algorithm , for instance (Lin et al., 2010) purposed a method that used maximum matching to upgrade accuracy of Hidden Markov Model (HMM) and conditional random fields (CRF). However, if only used statistical-based methods and machine learning algorithm was need for a

Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 237–241,
Tianjin, China, 20-21 DEC. 2012

lot of corpus to train models, and it lack for expert knowledge.

In semantic role labeling, (You and Chen, 2004.) showed that adopted dependency decision making and example-based approaches to automatic semantic roles labeling system for structured trees of Chinese sentences. It used statistical information and combined with grammar rules for role assignments (Gildea and Hockenmaier, 2003).

Unknown word extraction was an important issue in many Chinese text processing tasks. (Chen and Ma, 2002) showed that used statistical information and as much information as possible, such as morphology, syntax, semantics, and world knowledge in unknown word extraction. In 2003 research, (Ma and Chen, 2003) showed that proposed a bottom-up merging algorithm to solve a problem that superfluous character strings with strong statistical associations were extracted as well.

In Traditional Chinese Parsing Bakeoff, there are two sub-tasks: Sentence Parsing and Semantic Role Labeling. This paper focuses on Sentence Parsing task and proposes hierarchical maximum pattern matching with rule induction approach to recognize the syntactic structure. We present the bakeoff results evaluation and provide analysis on the system performance in the following sections.

In the opening section of the paper, we illustrated the research motivations and related works. The system framework is illustrated in the section 2 that is composed of rule induction regulation and maximum pattern matching. The evaluate data and results are both described in third part. Finally, some findings and future works is shown in conclusion illustrated in section 4.

2 System Overview

Figure 1 illustrates the block diagram of the proposed parsing system for traditional Chinese sentence. In preparation of starting the system, we created a dictionary by training data that the words with only one POS tagging, and also extracted the relation information according to their POS tagging. The POS tagging frequency is calculated in proceeding and cascading of each POS tagging, and used to predict the POS tagging of those token undefined in the dictionary.

2.1 Rule induction regulation

Our concern is to consider the syntactic structure of traditional Chinese sentence. Herein, a two steps method is proposed in this paper. The first step is the Part-Of-Speech tagging using the lexical dictionary. It also performs two steps for accuracy. First, the tokens with only one POS tagging are detected in dictionary, and then POS-to-POS relations are performed to modify by calculating the POS tagging of tokens those were not defined in dictionary. For instance, in Figure 2(1), after performed dictionary mapping, the words “實際(actual)” and “公佈(announcement)” were not found in the dictionary. That is to say, no corresponding with the POS tagging is matched here, so they were marked as ‘Null’. However, we performed POS-to-POS relations modification, it could be found POS tagging by calculating POS relation information to obtain ‘VH’ and ‘VE’ for those token, as shown Figure 2(2).

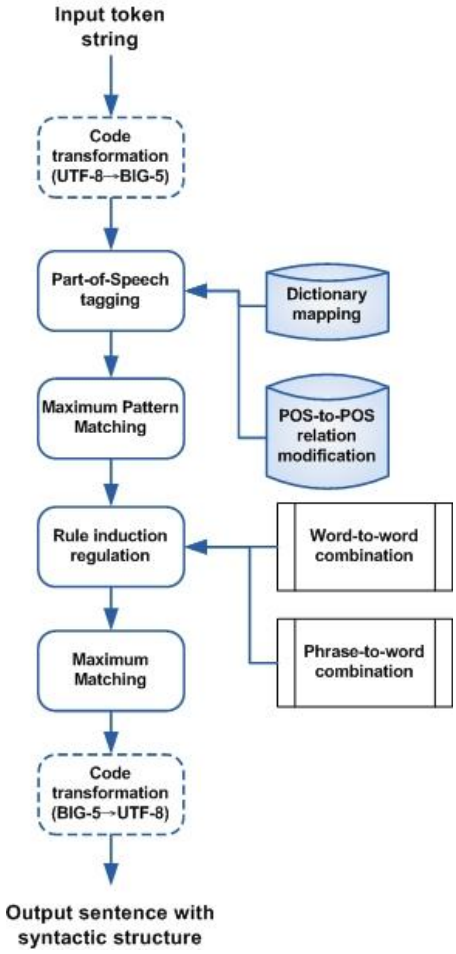
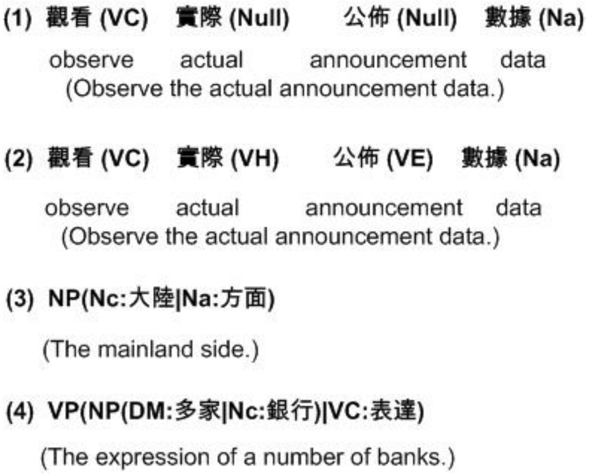

Input token string
Code transformation (UTF-8→BIG-5)
Part-of-Speech tagging
Dictionary mapping
Maximum Pattern Matching
POS-to-POS relation modification
Rule induction regulation
Word-to-word combination
Maximum Matching
Phrase-to-word combination
Code transformation (BIG-5→UTF-8)
Output sentence with syntactic structure

Figure 1. Flowchart of proposed system



(1) 觀看 (VC) 實際 (Null) 公佈 (Null) 數據 (Na)
observe actual announcement data
(Observe the actual announcement data.)

(2) 觀看 (VC) 實際 (VH) 公佈 (VE) 數據 (Na)
observe actual announcement data
(Observe the actual announcement data.)

(3) NP(Nc:大陸|Na:方面)
(The mainland side.)

(4) VP(NP(DM:多家|Nc:銀行)|VC:表達)
(The expression of a number of banks.)

Figure 2. Two examples for POS-to-POS relations modification

In rule induction regulation, we were able to observe the syntactic structure in training data, and instituted syntactic structure rules of word-to-word and phrase-to-word in following:

1. **NP-Phrase structure:** It is composed of combining by noun and noun, or noun-phrase and noun.
   Na Na → NP
   NP Na → NP

2. **VP-Phrase structure:** It is composed of combining by adverb and verb, or verb and noun-phrase.
   D VC → VP
   VC NP → VP

3. **PP-Phrase structure:** It is composed of combining by preposition and noun-phrase.
   P NP → PP

4. **GP-Phrase structure:** It is composed of combining by noun-phrase and 'Ng', or verb-phrase and 'Ng'.
   VP Ng → GP
   NP Ng → GP

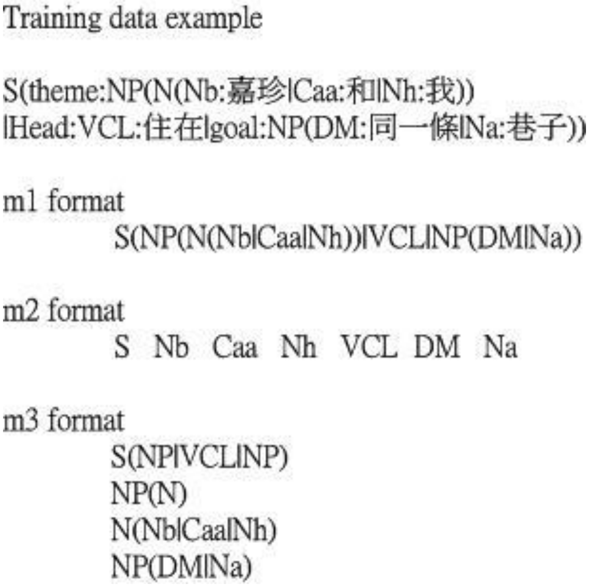
According to the rule categories defined previous, it could further be used to process the short-term syntactic structure, as shown in Figure 2 (3) and Figure2 (4).

2.2 Maximum pattern matching

In order to obtain desired information, the statistics method is used to obtain the syntactic information from training data. In the proposed method, a statistics approach used to extract the chunks is called as maximum pattern matching. The data *m1* is obtained by keeping part of speech (POS) and parser label of each word obtained from training corpus, the semantic role labeling is ignored in this stage. Furthermore, lexical text without any parse label expect the most outside parse label named *m1*, and the parse label order according to NP-VP-S-PP-GP sequence. Then utilized training data to get an only lexical text that existed everyone lexical or parse label named *m2*, and separated parse label for brackets named *m3* (see the Figure 3).

We could get the lexical of query sentence by part-of-speech, and used the lexical sequence to search for m1. In case all lexical of query sentence was totally matching m1, and we determine the query that to be part of m2, and we add to boundary and parse label for query sentence that utilized information of *m2*.

If lexical sequence was not complete corresponding to *m1*, the query sentence integrated by rule-based, and result that integrated with parse label by rule-based used m3 information to integrated again (see the Figure 4). It is maximum pattern matching for that integrated with parse label, because we compared lexical sequence of query sentence with *m3* information, always search for the maximum length of query sentence, and reduced length slowly until length equal to one.



Training data example

S(theme:NP(N(Nb:嘉珍|Caa:和|Nh:我))
|Head:VCL:住在|goal:NP(DM:同一條|Na:巷子))

m1 format
S(NP(N(Nb|Caa|Nh))|VCL|NP(DM|Na))

m2 format
S Nb Caa Nh VCL DM Na

m3 format
S(NP|VCL|NP)
NP(N)
N(Nb|Caa|Nh)
NP(DM|Na)

Figure 3. An example about the relationship between lexical and parse label extracted from training data

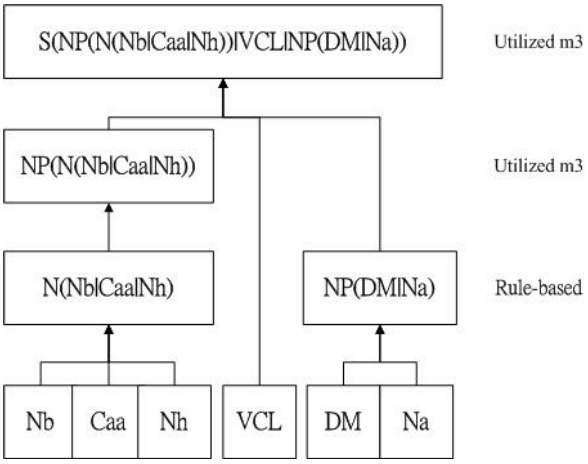


Figure 4. An example about the sentence added to boundary and parse label

3 Evaluation Results and Discussions

In training data, there are 65K token strings, we extract 39K token to create the dictionary. In testing evaluations, there are 1K token strings to be testing.

Table 1. Evaluation result

	Precision	Recall	F1
Closed	0.435	0.532	0.479

The evaluation of our system in sentence parsing sub-task is shown in table 1. Our system obtains 44% precision, 53% recall and 48% F1.

Table 2 shows the details parser ratio of each syntactic structure. For the result, it has highest ratio about 80% on sentence level parser. In test data, the token of each string are more than 6, it has more probability correspond to the syntactic structure of sentence level parser. For NP-Phrase parser, it has second rank. During we observe the training data, there are most NP-Phrase structures, and some noun of type can be NP-Phrase itself. So we focus on NP-Phrase when design the rule induction. VP-Phrase and PP-Phrase have lower ratio, some verb will combine noun to be NP-Phrase, and the rule we design on both VP-Phrase and PP-Phrase are not robustness to cause maximum pattern matching fail. GP-Phrase sample is rare in training data, it only a rule in our system.

Table 2. Evaluation result in details

Type	**Truth**	**Parser**	**Ratio(%)**
S	1233	987	80.5
VP	679	104	15.32
NP	2974	1449	48.72
GP	26	0	0
PP	96	16	16.67
XP	0	0	N/A

4 Conclusion

The evaluation results show that our system performs well in sentence level, but has lower performance in VP-Phrase and PP-Phrase, even for GP-Phrase, our system can't detect the syntactic structure.

By observing the evaluation result, we discover that have much errors in the POS tagging due to the out of vocabulary (OOV). For instance, proper noun such as personal names "張蘭(Zhang Lan)" and "寶來(Polaris)" that are not defined in the dictionary. During POS tagging step, it usually causes errors by using the POS-to-POS relation modification. The wrong POS labeling affects the performance in rule induction regulation step significantly and maximum pattern matching. In maximum pattern matching, the parse labeling is ordered according to NP-VP-S-PP-GP sequence. Maximum pattern matching was possible to correct the wrong structure and labeling of the parsing because it always searches for NP first.

In future works, we will focus on improving the POS tagging methods and enhance the unknown word tagging. For rule induction, there are more robustness rule we can design and achieve the improvement in the performance of maximum pattern matching

Reference

Chu-Ren Huang, Keh-Jiann Chen, Feng-Yi Chen, Keh-Jiann Chen, Zhao-Ming Gao, Kuang-Yu Chen . 2000. Sinica Treebank: Design Criteria, Annotation Guide-lines, and On-line Interface. In Proceedings of 2nd Chinese Language Processing Workshop (Held in conjunction with ACL-2000). 29-37.

Keh-Jiann Chen, Chu-Ren Huang, Feng-Yi Chen, Chi-Ching Luo, Ming-Chung Chang, Chao-Jan Chen, Zhao-Ming Gao. 2003. Sinica Treebank: Design Cri-teria, Representational Issues and Implementation. In Anne Abeille (Ed.) Treebanks Building and Using Parsed Corpora. Language and Speech series. Dor-drecht:Kluwer, 231-248.

Keh-Jiann Chen, Wei-Yun Ma. 2002. Unknown word extraction for Chinese documents. In Proceedings of COLING 2002, pages 169-175.

Wei-Yun Ma, Keh-Jiann Chen. 2003. A bottom-up merging algorithm for Chinese unknown word extraction. In Proceedings of the second SIGHAN workshop on Chinese language processing, Pages 31-38.

Chen Keh-Jiann, Chu-Ren Huang, Li-Ping Chang, Hui-Li Hsu. 1996. Sinica Corpus: Design Methodology for Balanced Corpra. Proceedings of the 11th Pacific Asia Conference on Language, Information, and Computation (PACLIC II), SeoulKorea, pp.167-176.

Chinese Knowledge Information Processing Group. 1993. Categorical Analysis of Chinese. ACLCLP Technical Report # 93-05, Academia Sinica.

Jia-Ming You, Keh-Jiann Chen. 2004. Automatic Semantic Role Assignment for a Tree Structure. Proceedings of SIGHAN workshop.

Qian-Xiang Lin, Chia-Hui Chang, Chen-Ling Chen. 2010. A Simple and Effective Closed Test for Chinese Word Segmentation Based on Sequence Labeling. International Journal of Computational Linguistics & Chinese Language Processing, Vol. 15, No. 3-4, September/December 2010.

Tsai Yu-Fang and Keh-Jiann Chen. 2003, Context-rule Model for POS Tagging. Proceedings of PACLIC 17, pp146-151.

Asahara, M., C.L. Goh, X.J. Wang, Y. Matsumoto. 2003. Combining Segmenter and Chunker for Chinese Word Segmentation. In Proceedings of Second SIGHAN Workshop on Chinese Language Processing, pp. 144–147.

Chooi-Ling Goh, Masayuki Asahara, Yuji Matsumoto. 2005. Chinese Word Segmentation by Classification of Characters. Computational Linguistics and Chinese Language Processing, 10(3), pp. 381-96.

Daniel Gildea and Julia Hockenmaier. 2003. Identifying Semantic Roles Using Combinatory Categorial Grammar. Conference on Empirical Methods in Natural Language Processing (EMNLP), Pages 57-64.

Lu, X. 2005. Towards a Hybrid Model for Chinese Word Segmentation. In Proceedings of Fourth SIGHAN Workshop on Chinese Language Processing, 189-192.