

A Chinese Chunk-based Parser

Samuel W. K. Chan[†]

Lawrence Y. L. Cheung

Mickey W. C. Chong

Dept. of Decision Sciences
Chinese University of Hong Kong
Shatin, Hong Kong SAR

Email: swkchan@cuhk.edu.hk[†]

Abstract

The technique of multi-level chunking has been applied to full sentence parsing in a number of previous studies. Instead of using the popular IOB chunking technique, this paper explores a novel method of locating chunks by identifying their boundaries. We investigated the use of a machine learning algorithm to classify the boundaries using the part-of-speech (POS) of words and their associated information-theoretic measures. A phrase recognizer is developed to predict the syntactic class of the identified chunks. While the training is conducted using a partial version of a treebank with less than 360,000 POS-tagged tokens, our system achieves a performance of 80.46% labelled recall, 85.70% labelled precision, and 83.00% F-score in the testing of 93,000 POS-tagged tokens, without any assumption of prior knowledge or statistical measures from any other resources.

Keywords: chunking, parsing, machine learning, treebank, natural language processing

1. Introduction

Sentence chunking was proposed by Abney (1991) as a computationally less costly and shallow text processing. Instead of full parsing, it is possible to focus on identifying major chunks to reduce the computational burden in some natural language processing applications in many cases. The output can subsequently feed into more sophisticated linguistic analyses. Pioneer works, such as Magerman & Marcus (1990) among others, focused on the identification of non-recursive base noun phrases (base NPs). Some recent studies adapt the chunking technique and turn it into a full sentence parser, and achieve reasonably good results (Abney, 1996; Sang, 2001). To enhance the performance, in this paper, we investigate a new method to identify chunk boundaries. The technique is applied to parsing Chinese sentences in a Chinese Treebank (Zhou, 2003; Zhou, 2004). A machine learning algorithm is used to assign the chunk boundaries and syntactic class to the identified chunks.

2. Related Work

Inspired by the work from Ramshaw & Marcus (1995), chunk-based parsing is often cast as a tagging problem. In the training input, each word is assigned a tag to indicate the position of the word relative to a chunk of type X. The tags include beginning (B-X), inside (I-X), or outside (O-X). The chunker learns how to assign the IOB tags to words based on the POS tags. The memory-based learning algorithm is utilized to

recognize the syntactic class (Sang, 2001). Trained on the English Penn Wall Street Journal Treebank, the parser obtains the results with labelled precision 78.72%, labelled recall 82.34%, F-score 80.49% and crossing brackets rate 1.69. Tsuruoka & Tsujii (2005) adopt a sliding-window approach to collect potential chunks. To reduce the search cost, their rules are filtered by a rule dictionary and a naive Bayes classifier, and the POS tag set is simplified. A set of candidate phrases is produced. A Maximum Entropy classifier is then applied to phrase recognition. Finally, the most probable candidate parse is selected. Using the English Penn Wall Street Journal Treebank, their system achieves better results of labeled precision (82.59%), labeled recall (87.69%) and F-score (85.06%). This chunk-based approach has also been applied to parsing Chinese sentences. Fung *et al.* (2004) utilize IOB-tagging and iterative level-based chunking to parse Chinese sentences. Based on the Maximum Entropy model and training on Penn Chinese Treebank, their system performs both POS-tagging and chunking, and achieved a performance of 78.30% labelled precision, 80.85% labelled recall and 79.56% F-score.

3. Basic Rationale of the Parser

Chunking refers to the partitioning of the sentence into segments which correspond to phrases (i.e. chunks). Chunkers accept a stream of words with the corresponding POS tags as input, and group the sub-sequences of tags that most likely form a phrase. Abney (1996) is the first to apply chunking by levels (or what he calls “cascades”) to the analysis of the hierarchical syntactic structure. Initially, the chunker identifies minimal chunks in a very local context. Then the chunks progressively merge with the neighboring tags or chunks by repeated application of the procedure. The procedure iterates till only one tag is left in the input. The following simulates the chunking of a Chinese sentence.


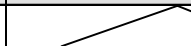
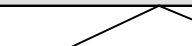
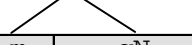
Level 4	Tag	dj				
						
Level 3	Tag	n	vp			
						
Level 2	Tag	n	vC	np		
						
Level 1	Tag	n	vC	mp	n	
						
Level 0	Tag	n	vC	m	qN	n
	Word	社会学	是	一	门	科学
	Gloss	sociology	be	one	(classifier)	science

Figure 1: Construction of the parse tree by level.

In Level 0 chunking, ‘one’ and the classifier¹ forms an mp. The POS tags in these chunks are replaced with a syntactic class (SC) tag, i.e. mp in Level 1. Essentially, POS tags are terminal nodes, and SC tags, non-terminal nodes. Similarly, the mp and ‘science’ together forms a chunk (i.e. np). The merging procedure continues and stops after four iterations.

¹ A classifier is a Chinese part of speech which is used with numerals and nouns to define the quantity, shape or nature of entities.

4. Parser Architecture

4.1 Chunker, Phrase Recognizer and Head Identifier

Our chunk-based parser has to tackle four tasks. First, a chunker must be able to locate chunks. Second, a phrase recognizer predicts the SC tag corresponding to the chunk (e.g. $m + qN \rightarrow mp$) on the basis of the tag sequence of the chunk. Many previous studies, such as Sang (2001) to name a few, tend to combine chunking and phrase recognition in one step, though our proposed parser separates them into two steps. Third, a learning algorithm is devised to learn the identification of all the possible heads. The syntactic structures encoded in the treebank form the empirical basis of our machine learning. Figure 2 shows the basic architecture of the parser.

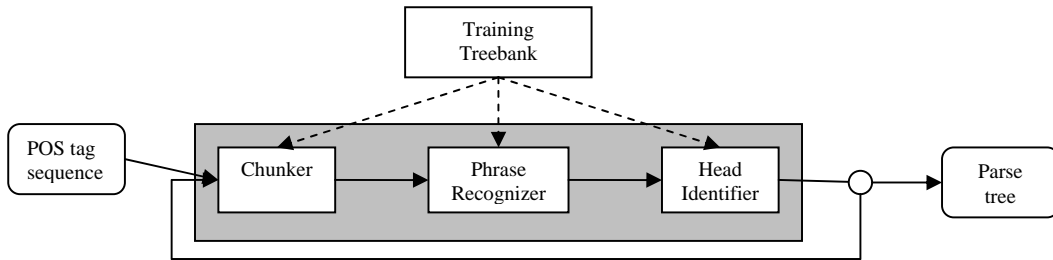


Figure 2: Architecture of the chunk-based parser

Chunk identification is one of the most critical tasks in our chunk-based parsing. A chunker isolates subsequences in the input tag stream as potential chunks. In this paper, we explore an approach that aims at identifying chunk boundaries or “chunking points.” Suppose the focus point is between two consecutive POS tags x_n and x_{n+1} . A chunking point is a focus point such that x_n and x_{n+1} that are not siblings with the same parent node in the target parse tree. We refer to all non-chunking points as “merging points” Chunking points are marked with %, and merging points, with +. For example, the output of Level 0 chunking is shown (1).

$$\% \ n \ \% \ vC \ \% \ m \ + \ qN \ \% \ n \ \% \tag{1}$$

The phrase recognizer assigns SC tags to the tag sequences that have been isolated as chunks in the output of the chunker. As shown in Sentence (1), the phrase recognizer takes the output of the chunker and replaces the chunk “ $m + qN$ ” with mp . The updated tag sequence becomes the input of the head identifier, which inspects each tag in the phrase and decides which tag(s) is the most likely head(s) of the phrase. The head information is appended to the SC tag (e.g. $mp-1$). The way we tackle phrase recognition and head identification is to use a machine learning algorithm to learn the rule patterns. The completion of the three steps concludes the processing of one syntactic level. The output is fed into the loop again for processing at the next level. The process terminates when the output contains only one node (or SC tag).

4.2 Learning Module

The training data for machine learning was derived from the Tsinghua Chinese Treebank (Zhou, 2003; Zhou, 2004). It comes with a tagging system consisting of 66 POS tags and 16 SC tags. The Chinese treebank includes 48,982 parse trees and 447,454 words.

The chunker training data was obtained by extracting the tag sequence by level from the parse trees. Each focus point is a training case. For each training case, an attribute vector is constructed, and the corresponding target value (i.e. chunking point vs. merging point) was also provided. The attributes for chunker training can be classified into two broad categories. The first type of attributes consists of the POS/SC tags surrounding the focus point. The second type is the information-theoretic measures of these tags.

The information measures include mutual information (MI), which reflect the likelihood of the fragment collocation.

	Training Data	Testing Data	Total
No. of word tokens	354,767 (79.3%)	92,687 (20.7%)	447,454 (100.0%)
No. of parse trees	32,771 (66.9%)	16,211 (33.1%)	48,982 (100.0%)

Table 1: Size of training and testing data

The phrase recognizer predicts the SC tag on the basis of the tag sub-sequence in a chunk. The head identifier locates the tag that is most likely the head of the phrase. The machine learning method is again used to learn to make predictions about the SC tag and the head assignment. In the training of the phrase recognizer, all phrases from the treebank were extracted as training cases. For each training case, an attribute vector is set up and the target class is the SC tag (e.g. np, pp etc). The attributes are the tags of each phrase and the target value is the SC tag.

5. Experiments

The modules described above were put together to form a parser. The parser was applied to parsing 16,211 Chinese sentences in the testing set. Table 2 summarizes the parsing performance.

Without head identification			With head identification		
LP	LR	F-Score	LP	LR	F-Score
83.5%	81.3%	82.4%	75.6%	73.6%	74.6%

Table 2: Parsing performance

6. Conclusion

This study investigates a new approach to chunk-based parsing. Through the identification of chunk boundaries, sentences are segmented based on various POS tags and their collocation measures. We have articulated a way to combine different, large and heterogeneous sets of attributes in the chunking point detection using the machine learning algorithm with high accuracy. Together with the phrase recognizer and head identifier, the parser achieves good performance, considering that the only resource available is a 350,000-word treebank. Though our evaluation was conducted using a Chinese treebank in the event, the computational method is language-independent and can be easily adapted to other languages.

7. Acknowledgements

We thank the Chinese Information Processing Society of China, Tsinghua University, Northeastern University for organizing CIPS-ParsEval-2009, and for providing the partial Tsinghua Chinese Treebank for training and testing. The work described in this paper was partially supported by the grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project Nos. CUHK470605 and CUHK440607).

8. References

- Abney, S. (1991). Parsing by Chunks. In: Berwick, R., Abney, S., Tenny, C. (eds.), *Principle-Based Parsing*. Kluwer Academic.
- Abney, S. (1996). Partial Parsing via Finite-state Cascades. *Natural Language Engineering*, 2, 337-344.

- Fung, P., Ngai, G., Yang Y.S., & Chen, B.F. (2004). A Maximum-Entropy Chinese Parser Augmented by Transformation-Based Learning. *ACM Transactions on Asian Language Information Processing*, 3, 2, 159-168.
- Magerman, D.M., & Marcus, M.P. (1990). Parsing a Natural Language Using Mutual Information Statistics. *Proceedings of AAAI-90, 8th National Conference on AI*, 984-989.
- Ramshaw, L. A., & Marcus, M.P. (1995). Text Chunking Using Transformation-based Learning. *Proceedings of the Third Workshop on Very Large Corpora*, 82-94.
- Sang, E. (2001). Transforming a Chunker to a Parser. In: Veenstra, J., Daelemans, W., Sima'an, K., Zavrel, J., (eds.), *Computational Linguistics in the Netherlands 2000*, 177-188.
- Tsuruoka, Y., & Tsujii, J. (2005). Chunk Parsing Revisited. *Proceedings of the 9th International Workshop on Parsing Technologies*, 133-140.
- Zhou, Q. (2003). Build a Large-Scale Syntactically Annotated Chinese Corpus. *Proceedings of 6th International Conference of Text, Speech and Dialogue (TSD2003)*, Czech Republic, Sept. 9 –12, 106-113.
- Zhou, Q. (2004). (in Chinese). Annotation scheme for Chinese Treebank. *Journal of Chinese Information Processing*, 18, 4, 1-8.