Computational Linguistics and Chinese Language Processing
Vol. xx, No. x, June 2007, pp. xx-xx


# Improve Parsing Performance by Self-Learning

Yu-Ming Hsieh[1], Duen-Chi Yang[1], and Keh-Jiann Chen[1]

**Abstract**

There are many methods to improve performance of statistical parsers. Resolving structural ambiguities is a major task of these methods. In the proposed approach, the parser produces a set of n-best trees based on a feature-extended PCFG grammar and then selects the best tree structure based on association strengths of dependency word-pairs. This paper aims to provide a self-learning method to resolve these problems. The constructed structure evaluation model improved the bracketed f-score from 83.09% to 86.59%. We believe that the above iterative learning processes can improve parsing performances automatically by learning word-dependence information continuously from web.

**Keywords:** Parsing, Word Association, Knowledge Extraction, PCFG, PoS Tagging, Semantic 剖析，詞彙關聯，知識萃取

## 1. Introduction

How to solve structural ambiguity is an important task in building a high-performance statistical parser, particularly for Chinese (Black et al., 1991; Charniak and Johnson, 2005). Since Chinese is an analytic language, words can play different grammatical functions without inflection. A great deal of ambiguous structures would be produced by parsers if no structure evaluation were applied. There are three main steps in our approach that aim to disambiguate the structures. The first step is to have the parser produce n-best structures. Second, we extract word-to-word associations from large corpora and build semantic information. The last step is to build a structural evaluator to find the best tree structure from the $n$-best candidates.

## 2. Feature Extension of PCFG Grammars for Producing the N-best Trees

Treebanks provide not only instances of phrasal structures and word dependencies but also their statistical distributions...

1. Institute of Information Science, Academia Sinica, Taipei, Taiwan
E-mail: {morris, ydc, kchen}@iis.sinica.edu.tw

## 2.1 Coverage Rates of the Word Associations

Data sparseness is always a problem of statistical evaluation methods. The five levels of word associations derived from Figure 1 are...

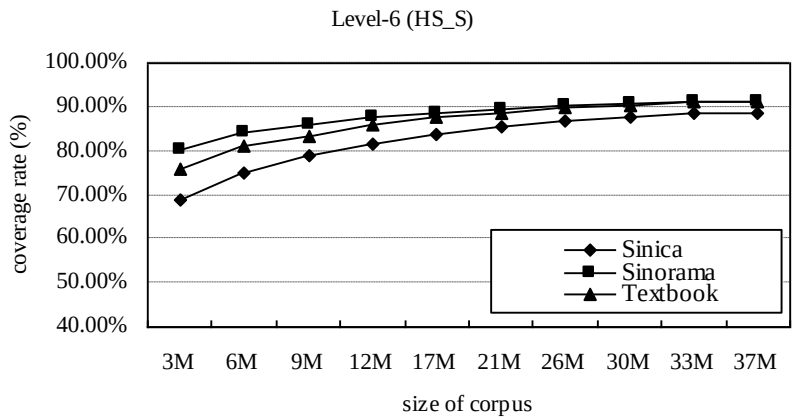


Figure 1: WA coverage rate of Level-6.

### 2.1.1 Title

From the results shown in Table 5...

| Testing Data | Sources | Hardness | Rule type-1 | Rule type-2 | Rule type-3 |
|---|---|---|---|---|---|
| Sinica | Balanced corpus | Moderate | 92.97 | 94.84 | 96.25 |
| Sinorama | Magazine | Difficult | 90.01 | 91.65 | 93.91 |
| Textbook | Elementary school | Easy | 93.65 | 95.64 | 96.81 |

Table 1: The 50-best oracle performances from the different grammars.

# Acknowledgments

This research was supported in part by National Science Council under Grant NSC 95-2422-H-001-008- and National Digital Archives Program Grant 95-0210-29-戊-13-09-00-2.

# References

Black, E., Abney, S., Flickinger, D., Gdaniec, C., Grishman, R., Harrison, P., et al. (1991). A procedure for quantitatively comparing the syntactic coverage of english grammars. In *Proceedings of the Workshop on Speech and Natural Language*, pages 306–311.

Charniak, E. and Johnson, M. (2005). Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 173–180, Ann Arbor, MI.