Resolving structural ambiguity is a major task in building high-performance statistical parsers, particularly for Chinese. Since Chinese is an analytic language, words can play different grammatical functions without inflections. A great deal of ambiguous structures would be produced by parsers if no structure evaluation were applied. There are three strategies in our approach that aim to evaluate and disambiguate sentential structures. The first strategy is to have a parser produce n-best structures instead of single structure of an input sentence. Second, we extract word-to-word associations from large corpora and build semantic information. The last strategy is to build a structural evaluator to find the best tree structure from the n-best candidates. The series of repetitive parsing, knowledge extraction and evaluation form a “self-learning” mechanism to the machine. Conventionally probabilistic preferences for grammar rules and feature dependencies were incorporated to resolve structure-ambiguities and had achieved improvement on parsing performance. However, the automatically extracted grammars and feature-dependence pairs suffer the problem of low coverage. We had proposed different approaches to solve these two different types of low coverage problems. For the low coverage of extracted grammar, a linguistically-motivated grammar generalization method is proposed. The linguistically-motivated generalized grammars are derived from probabilistic context-free grammars (PCFG) by right-associated binarization and feature embedding. The binarized grammars have better coverage than the original grammars directly extracted from Treebank. Then features are embedded into the lexical and phrasal categories to improve the precision of generalized grammar. Finally, we proposed a method to modify our grammars to increase the oracle scores in producing n-best sentences.

The low coverage of word association pairs is resolved by a self-learning method of automatic parsing and extracting word dependency pairs from very large corpora. It is possible to extract world knowledge from Treebanks, but the availability of a very large set of trees with rich linguistic annotations has long been a problem. A cheaper way to extract world knowledge is to automatically parse large amounts of text. A good parser could produce valuable information regarding dependence word-pairs between head words and their arguments or modifiers automatically. The following sample in figure 1 illustrates the process.

In the last step, we construct an evaluation model to evaluate n-best trees. A sentence structure is evaluated according to its syntactic and semantic plausibility. The syntactic plausibility is modeled by products of the probabilities of rules applied in the derivation tree. The semantic plausibility is modeled by the word association strengths between head words and their arguments or modifiers. We design a general syntactic and semantic evaluation model to select the best parse from the set of
parses. The system we described, using the standard PARSEVAL framework, achieves a bracketed f-score of 86.59%, which is higher than the original 1-best.

As mentioned above, parsers of any language aim to correctly analyze the syntactic structure of a sentence, often with the help of semantic information. This paper shows a self-learning method (see Figure 2) to produce imperfect (due to errors produced by automatic parsing) but unlimited amount of word association data to evaluate the n-best trees produced by a feature-extended PCFG grammar. We prove that, although the statistical association strengths produced by automatic parsing are not perfect, the extracted data is reliable enough in measuring plausibility of ambiguous structures. The parser with this word-association evaluation is considerably superior than those without evaluation. We believe that the above iterative learning processes can improve parsing performances automatically by learning word-dependence knowledge continuously from web.

The original paper was published in “Improve Parsing Performance by Self-Learning”