# Building a Chinese Shallow Parsed TreeBank for Collocation Extraction

Li Baoli[1], Lu Qin[2], Li Yin[2]

[1] Department of Computer Science and Technology,
Peking University, Beijing, P.R. China, 100871
libl@pku.edu.cn
[2] Department of Computing, The Hong Kong Polytechnic University,
Hung Hom, Kowloon, Hong Kong
{csluqin, csyinli}@comp.polyu.edu.hk

**Abstract.** To automatically extract Chinese collocations and build a large-scale collocation bank, we are developing a one-million-word Chinese shallow parsed treebank. The treebank can be used not only as a training set for our shallow parser, but also as processed data from which collocations are extracted. This paper presents several issues related to this on-going project, such as our definition of shallow parsing used in Chinese collocation extraction, guideline preparation, and quality control.

## 1 Introduction

A collocation is a fixed usage of two or more words, occurring adjacently or separated by other words. The exact meaning of a collocation usually cannot be derived directly from the meaning of its components. Collocations are important for a number of applications, such as machine translation, computational lexicography, and so forth. Many studies on automatic collocation extraction have been conducted in the past decades [1][2]. The techniques in these studies are mainly based on lexical statistics, including frequency, mean and variance, hypothesis testing, and mutual information.

We consider that collocations should be restricted within grammatically bound elements that occur in a particular order. Co-occurred words like *doctor – nurse* or *plane – airport* are not regarded as collocations. To determine whether two or more words form a collocation, syntactic information must be introduced. Shallow parsing can be used effectively to identify local structure of a sentence without the need for full parsing, thus it becomes a natural choice for collocation extraction.

In order to extract Chinese collocations automatically and build a large-scale collocation bank, we are developing a Chinese shallow parsed treebank. The treebank can be used not only as a training set for the shallow parser, but also as processed data from which collocations are extracted. Several efforts were made on building large scale Chinese full parsed treebanks for general purpose [3][4], but little has been done to construct a shallow parsed treebank, especially for the purpose of collocation extraction. This motivated us to build a one-million-word shallow treebank.

In this paper, we present several issues about how to build a Chinese shallow parsed treebank for collocation extraction. Section 2 gives our definition of shallow parsing used in collocation extraction. Section 3 discusses how to build such a shallow parsed treebank. Conclusion and future plans are given in Section 4.

## 2 Shallow Parsing for Collocation Extraction

Shallow or partial parsing is usually defined as the task of obtaining only a limited amount of syntactic information from running text. But this definition is equivocal, especially when used for an engineering project. Under the context of our project, shallow parsing for collocation extraction should be able to recognize basic blocks of a sentence where the boundary of syntactic chunks can be identified. The marking of the chunks can limit the identification of collocation either within a chunk or between chunks depending on the types of collocation we are looking for. As nested chunking sometimes can make training more difficult, we limit the nesting of chunks to only 2 levels at most. Any deeper syntactic structures are ignored. Consequently, our shallow parsed tree will be a tree of no more than 2 in height. For example, a sentence (a) in Figure 1 (on next page) is fully parsed as (b) in the LDC Chinese Treebank [3], whereas it is bracketed as (c) in our corpus. The sub-structures in the phrase “具体措施和政策要点 (concrete measures and essentials on policy)” are not annotated.

## 3 Some Issues in Treebank Annotation

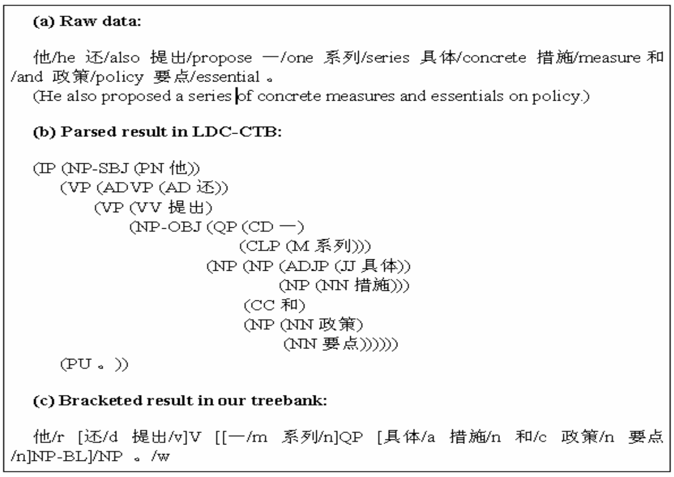
### 3.1 Guideline Preparation

One important consideration in the preparation of the annotation guideline was to make it applicable to Chinese collocation extraction. The workload of the annotation must also be manageable. Consequently, the guideline must be simple and easy to follow and the result can be of reasonable quality within a specified time frame.

### 3.2 Word Segmentation and Part of Speech Tagging

Unlike English, there is no space between Chinese words. Thus word segmentation must be done as a necessary preprocessing step. To avoid the difficulties in defining the notion of words, we simply derived the wordlist used in our annotation from a widely used Chinese syntactical lexicon [5]. Moreover, to comply with the principles adopted in contemporary linguistic theories, we defined POS tags based on syntactic distribution rather than meaning and we have about 88 different POS tags.

One difference between our guideline and those of others is in the processing of repetitive structures, such as AA (e.g., 看看/see), ABAB (e.g., 研究研究/research), and A-一-A (e.g., 想一想/think). We regard such a repetitive structure as a single segmentation unit. In addition to POS tag, we further annotate its internal structure (e.g. 看看/v-AA). From this detailed information, we can easily obtain its stem, which will be helpful to identify collocations even in a small corpus.



**(a) Raw data:**

他/he 还/also 提出/propose 一/one 系列/series 具体/concrete 措施/measure 和/and 政策/policy 要点/essential 。

(He also proposed a series of concrete measures and essentials on policy.)

**(b) Parsed result in LDC-CTB:**

```
(IP (NP-SBJ (PN 他))
    (VP (ADVP (AD 还))
        (VP (VV 提出)
            (NP-OBJ (QP (CD 一)
                        (CLP (M 系列)))
                    (NP (NP (ADJP (JJ 具体))
                                (NP (NN 措施)))
                        (CC 和)
                        (NP (NN 政策)
                              (NN 要点))))))
    (PU 。))
```

**(c) Bracketed result in our treebank:**

他/r [还/d 提出/v]V [[一/m 系列/n]QP [具体/a 措施/n 和/c 政策/n 要点/n]NP-BL]/NP 。/w

**Fig. 1.** A sample sentence and its parsed result

### 3.3 Syntactic Bracketing

In our annotation, each clause ended with punctuations such as period (。), comma (，), semicolon (；), exclamation point (！), colon (：), and interrogation mark (？), is taken as a processing unit. The annotation is conducted in a Top-Down manner. When we process a clause, the main predicate is first recognized. Then each phrase with maximum length, which plays a distinct semantic role of the predicate, is bracketed. The concept of "phrase with maximum length" depends on contexts[1]. After recognizing the first level chunks, we further parse the chunks with nested structures. When extracting candidate collocations, we first consider the headwords of the first level chunks, and then the words within a chunk. We believe that when a chunk or a clause is too short, further annotation may not bring more benefits. In fact, simple statistical methods can filter false collocations derived from such chunks and small chunks can be easily parsed.

In our annotation, each bracket has zero or more structural or functional tags as well as a syntactic label (like [3]). For example, a noun phrase should be further annotated with its internal structure. If the internal structure is modifier-modified (e.g.,

[1] For example, "香港理工大学计算机系的学生" is a noun phrase with maximum length in the sentence "香港理工大学计算机系的学生很受欢迎", whereas "计算机系的学生" is such a maximum phrase in the sentence "计算机系的学生表演了精彩的节目".

漂亮的姑娘/beautiful girl), we only care about the headword when extracting collocations between different chunks. If the internal structure is parallel (e.g., 桌子椅子 /tables and chairs), each component in the chunk will be considered.

### 3.4 Annotation Process and Quality Control

Our annotation is conducted in two phases: first, word segmentation and POS tagging, and second, syntactic bracketing. The treebank annotation is an iterative process, in which incremental refinement of guidelines, corpus, and tools, is done step by step. One supervisor and two annotators with linguistic background are involved in our project. The one-million-word original text corpus was selected from news reports, and mainly contains articles about economic development.

We take some measures to maintain the quality, i.e. annotation accuracy and inter-annotator consistency. Every annotator is assigned about 60% of the entire data to process. Thus about 20% of the data is randomly selected for double annotation. These cross-annotated texts could be used to measure the inter-annotator consistency. Moreover, the supervisor must check and re-annotate the same part based on the two annotators' work. The final result is used as gold standard to evaluate the annotation accuracy of each annotator.

## 4 Conclusion and Future Plans

We have discussed some issues in building a Chinese shallow parsed treebank, which include our definition of partial parsing used in collocation extraction, guideline preparation, and quality control. Till now, we have finished the first annotation phase, i.e. word segmentation and POS tagging, and a small part of texts have been syntactically bracketed. The final treebank is expected to be completed in June 2003.

As indicated above, our shallow parsed treebank contains very limited syntactic and semantic information. We plan to annotate the syntactic function of each chunk and the relationship between predicate chunk and other chunks in the future.

## References

1. Christopher D. Manning, Hinrich Schutze: Foundations of Statistical Natural Language Processing. MIT Press (1999)
2. Smadja F.: Retrieving Collocations from text: Xtract, Computational Linguistics. 19:1(1994) 143-177
3. Fei Xia, et al.: Developing Guidelines and Ensuring Consistency for Chinese Text Annotation. In the Proceedings of the second International Conference on Language Resources and Evaluation (LREC-2000), Athens, Greece (2000)
4. Keh-Jiann Chen, et al.: the CKIP Chinese Treebank: Guidelines for Annotation. In: Building and Using Syntactically Annoted Corpora, Dordrecht: Kluwer (2000)
5. Yu Shiwen, et al.: the Grammatical Knowledge-base of Contemporary Chinese: a Complete Specification. Beijing: Tsinghua University Press (1998)