# A Block-Based Robust Dependency Parser for Unrestricted Chinese Text[1]

**Ming Zhou**
Microsoft Research China,
Sigma Centre, 49#, Zhichun Road,
100080, Beijing, China
mingzhou@microsoft.com

**Abstract**

Although substantial efforts have been made to parse Chinese, very few have been practically used due to incapability of handling unrestricted texts. This paper realizes a practical system for Chinese parsing by using a hybrid model of phrase structure partial parsing and dependency parsing. This system showed good performance and high robustness in parsing unrestricted texts and has been applied in a successful machine translation product.

## Introduction

Substantial efforts have been made to parse western languages such as English, and many powerful computational models have been proposed (Gazdar, et al, 1987, Tomita, M, 1986). However, very limited work has been done with Chinese. This is mainly due to the fact that the structure of the Chinese language is quite different from English. Therefore the computational model in processing English may not be directly applied to the Chinese language. Lin-Shan Lee et al (1991) proposed a Chinese natural language processing system with special consideration of some typical phenomena of Chinese. Jinye Zhou et al (1986) presented a deterministic Chinese parsing methodology using formal semantics to combine syntactic and semantic analysis. However, most of the proposed approaches were realized on small-scale lexicon and rule base (usually thousands words and tens or hundreds rules). It is still an open issue whether these models will work on real texts containing various ungrammatical phenomena. A parser capable of handling real text should have not only large lexicon and big rule base, but also high robustness in coping with different kinds of ungrammatical phenomena. Therefore, it is important to design a grammar scheme which not only is capable of representing the unique grammar structures which are different with English, but also qualified of handling unrestricted text.

Phrase structure scheme is usually used in English parsing models to represent sentence structures, but it is not convenient and not strong enough to express Chinese sentence by phrase structure in some occasions. For examples:

Sentence-1 我们请她喝啤酒。

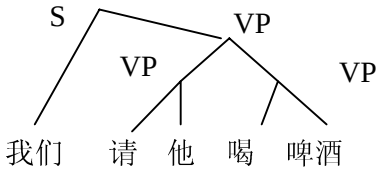


Fig. 1 phrase structure

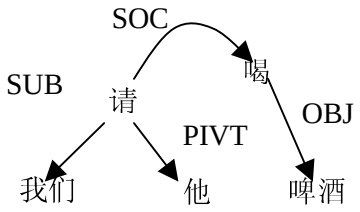


Fig.2 dependency structure

[1] This work was mainly done while the author visited Kodensha Ltd, Japan during 1996-1999

Sentence-1 is a pivot sentence(兼语句), i.e., “她” is not only the object of “请” but also the subject of ”喝”. But this phrase structure cannot indicate the relations clearly as shown in Fig.1. However, the grammar structure is clarified if it is represented in dependency structure (Fig. 2). Therefore, it is believed that dependency grammar scheme is more suitable than phrase structure to represent Chinese structures (Zhou, Huang, 1994). However, traditional Dependency grammar realizes the dependency relations between any of two specific words, then numerous word based dependency knowledge should be constructed, this is a time-consuming task. Fortunately, knowledge for phrase structure parsing has been accumulated for Chinese for many years and it should be re-used to compensate the lack of knowledge of word-based dependency parsing. Therefore, to combine the advantages of phrase structure parsing and dependency parsing, we propose a new parsing strategy, called “block-based dependency parsing”.

A “block” means a basic component of sentence, for example, there are six blocks for sentence 1:

[我们][请][她][喝][啤酒][。]

Another example:
Sentence 2: 昨天上午我们请她的全班同学喝了一箱啤酒]。
Blocks:[昨天上午] [我们] [请] [她的全班同学] [喝了] [一箱啤酒][。]

A block represents an information unit in communications. For example, in Chinese-Japanese machine translation, translations of the members within a block in a Chinese sentence usually are in a same blocks in the Japanese translation. Furthermore, it is clear to represent block with phrase structure, while it is rather complicated with dependency structure.

This block-based dependency parsing process works like follows. For an input sentence, basic components of sentence, i.e., “blocks” are first identified by an ATN-like partial parsing procedure, which produces a clear skeleton of the sentence structure. In our phrase structure analysis, we don’t try to deduce the whole sentence into root S, instead, we only try to get the components, namely blocks. This partial parsing strategy guarantees high robustness. Then dependency parsing is applied in order to build dependency relations among blocks. The dependency parsing skips ungrammatical portions it encounters. This strategy confines ungrammatical portion and avoids errors to be propagated globally. By partial parsing and skip strategy, this parser can handle long, complicated, or even faulty sentences. The experiments show that this parser is very robust and powerful. A parser constructed based on this approach has been developed, with 220,000 words, 5,000 part-of-speech tagging rules, over 1,000 block parsing rules and 300 dependency parsing rules. This parser has been applied in a Chinese-Japanese machine translation product (Zhou, 1999). To the author’s knowledge, this parser is one of the largest scale Chinese parser ever implemented in the world.

The outline of this paper is as follows. In section 1, we present our special solution to part-of-speech tagging which significantly affects the Chinese parsing. Section 2 describes in details the block-based dependency parsing approach. We then explain the dependency parsing algorithm in section 3. The experiment and its analysis are given in section 4. The conclusion is given in section 5.

## 1 Rule-based part-of-speech tagging

The Chinese language has many special syntactic phenomena substantially different from English (Chao, 1981; Huang, 1982, Wu and Hou, 1982). One of the biggest problems is that there is no morphological change for a verb, whether the verb functions as the predicate, subject, object, or modifier of a noun. For instance:

我们要进一步加大打击力度。（[打击 V力度N] NP）

西部开发是今后五十年的主要目标（[西部N开发V] NP）

Chinese linguistics literature insists that those words are verbs, and should be marked as “V”, regardless of what context they are in. In this sense, there will be phrase structure rules for noun phrase like:

NP->N+V

NP->V+N

However, there must be some rules for VP, S like:

S->N+V

VP->V+N

Therefore the conflict of rules becomes very serious. It means that part-of-speech information in Chinese is too weak to support Chinese syntactical analysis. To solve this problem, we propose that in the part-of-speech tagging stage, the real grammar features of this kind of words are determined directly as N, instead of V. To do this, we describe all possible word category information for a word in the lexicon, for example：

努力　V/ N/F/A

//V: verb; N: noun; F: adverb; A: adjective

A set of rules with comprehensive context constraints is designed to determine the specific part-of-speech of a word in a context. For example:

1. X(N|R)+努力(V|N|F|A) + X(V*)->努力(F) + X(V)
2. 的+努力(V|N|F|A) + X(没有)->的+努力(N) + X(N)
3. 的+努力(V|N|F|A) + X(V|N)->的+努力(A) + X(N)
4. 正在+努力(V|N|F|A) + X(~V)->正在+努力(V)

X(N|R): a word X, whose word category may includes N,R, or others.

R: Pronoun;

* : any part-of-speeches;

V* having V category;

~V having no V category;

It is ideal if we have a large corpus which has been tagged with thins kind of word category information, so that we can obtain tagging rules or obtained n-gram model by training. However, at present, we can't find a Chinese corpus tagged with this kind of part-of-speech information as the training data. We had to write the part-of-speech disambiguation rules manually. Currently, over 5,000 linguistics rules have been designed.

## 2 Block-based Chinese dependency analysis

As indicated in Fig. 3, block-based dependency analysis consists of four modules, i.e., word segmentation, part-of-speech tagging, block analysis and dependency analysis. A bi-directional heuristic longest matching method is applied to decide the optimal word sequence. A set of manually compiled linguistic rules is applied to decide the optimal word category sequence. In a partial parsing process, first, local structures (such as duplication, prefix and suffix) are identified by a set of word formation rules, and proper names are identified by a set of construction rules. This kind of local structures are called meta-blocks. Then frame structures (DP), which have paired starting word and ending word , such as "在"…."时", "从"…."中" etc are identified, but its internal structure analysis is delayed. Then ATN network is used to identify the basic blocks, called level-1 blocks (these blocks don't contain IP, LP and DP). Then we use a set of heuristic rules to identify the boundaries of IP and LP. Then ATN network will use again to identify the complicated blocks, called level-2 blocks, which may contain LP, DP, IP as its components. Then a sequence of blocks obtained is then transported to dependency parser, which will generate dependency relations among blocks. After that, we will recursively parse the internal parts of IP,

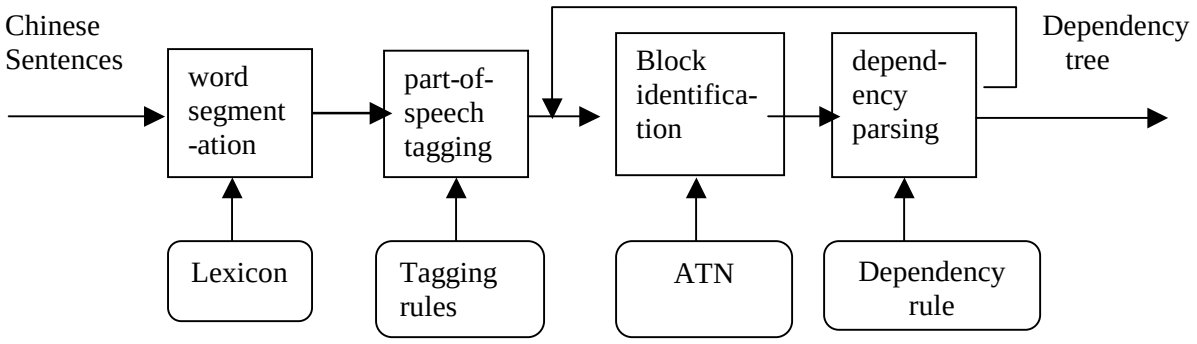


Fig. 3 Configuration of the block-based dependency parser

LP and DP to get its inner blocks and dependency relations.

We define 11 kinds of blocks as explained below.

| | | |
|---|---|---|
| NP | Noun phrase | 我的书 |
| UP | Digital phrase | 14560,三千二百 |
| UG | Digital-classifier phrase | 五斤,六十米 |
| NTL | Phrase expressing the period of time | 三十年, 60 个月 |
| NTP | Phrase expressing the exact time | 一九八九年三月 |
| AP | Adjective phrase | 美丽大方 |
| FP | Adverb phrase | 勤奋刻苦地 (学习英语) |
| VP | Verb phrase | 学习研究 |
| IP | Preposition phrase | 为人民服务 |
| LP | Post-position phrase | 仓库里 |
| DP | Frame structure | 为公司的利益起见,在打击盗版方面 |

Table 1 Blocks defined in the system

Except PP, LP and DP, each kind of block is defined by a set of rules in the form of phrase structure rule. All of these rules combined with syntactic and semantic constraints are implemented as an ATN network (Allen, 1995).

We also define 17 kinds of dependency relations for Chinese as shown in table 2.

| | | |
|---|---|---|
| 1 | SUB | Subject(主语) |
| 2 | OBJ1 | Indirect-object间接宾语) |
| 3 | OBJ2 | Direct object(直接宾语) |
| 4 | COMP | Complement(补语) |
| 4 | NUM | Amount(数量关系) |
| 5 | TOP | Topic(主题) |
| 6 | ADVN | Near adverbs（副词状语） |
| 7 | ADVF | Far adverbs（介词短语、方位短语、框式结构状语） |
| 8 | QT | miscellaneous before verbs (动词之前的闲杂成分) |
| 10 | HT | miscellaneous after verbs(动词之后的闲杂成分) |
| 11 | PUNC | Punctuation mark(标点符号) |
| 12 | PIVT | Pivot(兼語) |
| 13 | SOC | Pivot-complement(兼語补语) |
| 14 | VAA | Series of verbs after(向后连续动作) |
| 15 | VAB | Series of verbs before(向前连续动作) |
| 16 | G | 能愿动词接续关系 |
| 17 | LOG | Logical relation between sentences(假设、因果) |

Table 2 Dependency relations used in the system

For an Input: S= $w_1, w_2, \ldots, w_n$, the expected parse result includes two parts as described below:

① T : a set of sub-trees, each sub-tree represents a block.

$$T=\{T_1, T_2, T_3 \ldots\ldots T_n\}$$

② D: a set of 3-tuple in the form of {governor, dependant, dependency-relation}, which represents dependency relations between blocks.

$D=\{<gov_1, dep_1, rel_1>, <gov_2, dep_2, rel_2>, \ldots <gov_m, dep_m, rel_m>\}$

**Algorithm 1: The block-based parsing algorithm**

1) Identification DP by matching the starting word and ending word;
2) Identification of meta-blocks by bottom-up analysis;
3) Identification of NP, UP, UG, NTL, NTP, AP, FP, VP of level 1 by bottom-up analysis;
4) Identification LP, PP by looking for left boundary for LP and right boundary for IP, by using a set of Chinese linguistic rules;
5) Identification of NP, UP, UG, NTL, NTP, AP, FP, VP of level 2 by bottom-up analysis;
6) Dependency parsing with the blocks identified;
7) For blocks LP, DP and LP, recursively do 1 thorough 6.

In the following, we will illustrate the parsing process with an example.

Sentence 3: 实行不实行质量管理对中小运输企业提高经营水平来说非常重要。

**(1)Word Segmentation & Part-of-speech tagging**

/实行 V/不 F/实行 V/*质量 N/管理 N*/对 I/中小 A/运输 N/企业 N/提高 V/*经营 N/水平 N*/来说 L/非常 F/重要 A/。 P/

**(2) Meta-blocks identification**

[/实行 V/不 F/实行 V/]VP

**(3) Frame structure identification**

[对 I/中小 A/运输 N/企业 N/提高 V/*经营N/水平 N*/来说 L/]DP

**(4) Block identification**

block1: [/实行 V/不 F/实行 V/]VP

block2:[*质量 N/管理 N*/]NP

block3: [对 I/中小 A/运输 N/企业 N/提高 V/经营 N/水平 N/来说 L/]DP
block4: [非常 F/重要 A/]AP
block5: [。 P/]

**(5)Predicate Identification**
Block4 is determined as the predicate.

**(6) Dependency parsing**
(block2, block1, OBJ1)
(block3, block4, ADVF)
(block1, block4, SUB)
(block5, bolck4, PUNC)

**(7)Repeat the above parsing process to analyze the internal structure of DP, IP and LP**
Analyze block3 recursively (The detailed process is omitted).

Lots of efforts have been made to parse languages into phrase structure, and many powerful computational models have been proposed (Gazdar, *et al,* 1987, Tomita, M, 1986). We build up an ATN like network to identify these blocks. Since the ATN approaches can be found in the literatures (Allen, 1995), we will not describe this algorithm in details here. In the next section, we will focus on a new efficient algorithm for Chinese dependency parsing.

## 3 Dependency analysis

Text For an Input: S= $block_1, block_2, \ldots, block_n$, the dependency parsing will generate a set of 3-tuple in the form of {governor, dependant, dependency-relation}, which represents dependency relations between blocks in the given sentence.

$\{< gov_1, dep_1, rela_1 >, < gov_2, dep_2, rela_2 >, \ldots < gov_m, dep_m, rela_m >\}$

**Algorithm 2: The dependency parsing**

1) Count the number of block qualifying of acting as a predicate, denoted as *s*. These kind of blocks are called "predicate candidates".
2) Decide the predicate from these s blocks, denoted as $block_i$
3) If *s*=0, return; //need not analysis;
4) For any case of S, S=1,2,… (S>0), do dependency parsing respectively;

A sentence may contain *s* predicate candidates. For each case, we defined a detailed analysis algorithm. Up to now, the parser is designed to have ability to treat with sentences containing up to 7 predicate candidates. In case a sentence has more than 7 predicate candidates, it will be partitioned into two parts, and then doing analysis in turn.

Suppose the predicate block is $block_i$, the number of "predicate candidates" is denoted as *s*. We explain the dependency parsing by the following two simple cases.

**Case 1: s=1**

- For all $block_k$ before $block_i$, builds dependency relations of ($block_k$, $block_i$,SUB),($block_k$, $block_i$,ADV), ($block_k$, $block_i$,G), ($block_k$, $block_i$,TOP);
- For all $block_k$ after $block_i$ ,builds dependency relations of ($block_k$, $block_i$,COMP),($block_k$, $block_i$,OBJ1), ($block_k$, $block_i$,OBJ2), etc.

**Case 2: s=2,** Let's say the another predicate candidate is $block_j$

- For all $block_k$ before $block_i$ ,builds dependency relations of ($block_k$, $block_i$,SUB), ($block_k$, $block_i$,ADV), ($block_k$, $block_i$,G), ($block_k$, $block_i$,TOP);
- For all $block_k$ after $block_j$ ,builds dependency relations of ($block_k$, $block_j$,COMP), ($block_k$, $block_j$,OBJ1), ($block_k$, $block_j$,OBJ2), etc.
- For blocks between $block_i$ and $block_j$, Conducts detailed analysis based on the verb categories of $block_i$ and $block_j$
- Determines the dependency relation between $block_i$ and $block_j$

## 4 Experiments

A parsing system was implemented and extensive experiments have been performed. The system is written in C and tested on Pentium PC. A total of over 1,000 phrase structure rules and over 3,00 dependency rules were used for block-based parsing. We built a large lexicon of 220,000 word entries, with word category information and necessary syntactical and semantic features. This approach has been incorporated as Chinese parsing model in a successful commercial Chinese-Japanese machine translation system J-Beijing (Zhou, 1999).

This system accepts Chinese text and output the parsing result for each sentence. Each input sentence is defined as a word string ending with period, comma, question mark, semicolon, exclamation mark.

We evaluated the parsing result with two corpus: ① "primary school textbook of Singapore"(新加坡小学课本), a corpus consists of single sentences of modern Chinese, including 1842 sentences, which not only covers most Chinese sentence types, but also includes various of morphological phenomena, such as word duplication, affix, suffix, etc. ②Some news articles collected from People's Daily(1998,1999,2000). The sentences are real text, so there are lots of unknown words (mainly proper nouns), long sentences, complicated sentences, ellipsis, etc. The evaluation results are listed in table 3.

| Test corpus | #sentence | Average sentence length (words) | Analysis precision |
|---|---|---|---|
| Primary school textbook, Singapore | 1842 | 7.34 | 90.4% |
| People's Daily | 1400 | 14.52 | 67.7% |

Table 3 Evaluation result

Although this model has produced satisfactory initial results, some natural difficulties for the Chinese language still remain, such that further improvement is highly desired. Through mistake analysis, we found that some of main issues affecting the system performance seriously, as is listed below.

**1) Word segmentation**

■ 将来
/克林顿/总统/将来/上海/。/
/克林顿/总统/将/来/上海/。/

"将来" can not only function as single word, but also function as two words with totally different meaning.

**2) Part-of-speech tagging**

(1,我,R) (2,对,I) (3,你,R) (4,为,I) *(5,发展,N)* (6,中俄关系,N) (7,所,E) (8,作,V) (9,的,E) *(10,积极,F) (11,努力,F)* (12,表示,V,P) (13,高度,N) (14,赞赏,V,C) (15,，,P)

**3) Compound noun**

■ 研究工作
/我们/正在/研究/工作/
/这个/研究工作/十分/重要/

■ 研究过程
/我们/正在/研究/过程/参数/
/在/这个/研究/过程/中/,/

Since compound nouns cannot exhaustively numerated, errors will be inevitable.

**4) Identification of proper noun**

(1,江主席,N) (2,电贺,V,P) *(3,普,A) (4,京,N)* (5,当选,V,C) (6,俄,N) (7,联邦,N) (8,总统,N)

**5) Syntactical ambiguity**

(1,祝愿,V,P) (2,中俄,N) (3,两国,N) (4,世代,N) (5,做,V,C) (6,好,A, ) (6,邻居,N) (7,好,A) (8,伙伴,N) (9,好,A) (10,朋友,N) (11,。,P)

For pattern of "V+A+N", there are usually two kinds of reduction methods:

[[V+A]vp+N] *做好/邻居*

[V+[A+N]np] *做/好邻居*

All of these problems need further improvements in the future.

**Conclusion**

In this paper, a practical Chinese parser is presented. The block-based dependency parsing strategy is a novel integration of phrase structure partial approach and dependency parsing approach. The partial parsing approach and dependency parsing approach can cope with ungrammatical or faulty, or complicated sentences, therefore making the system highly

robust. Furthermore, our top-down strategy of identifying the Chinese special structures such as frame structures, preposition structures, post-preposition structures produces a simplified sentence skeleton, thereby improving the efficiency of parsing.

Although this model has shown satisfactory initial results, some natural difficulties for the Chinese language still remain, and further work will be needed. We currently determine the word category by a set of linguistics rules compiled by human which limits the precision of identification precision. Therefore, other approaches such as statistical approach or some kind of hybrid approach will be adopted in the future. In addition, new methods in handling ambiguous word segmentation, proper noun and compound noun identification, block analysis, predicate identification and dependency analysis will be studied.

**Acknowledgements**

Our thanks go to Dr. Kai-Fu Lee and Prof. Changning Huang of Microsoft Research China for their valuable suggestions. Also thanks all the members of Chinese-Japanese MT group of Kodensha for their great efforts in testing the parsing system and improving the dictionary.

**References**

Gazdar, G.,Franz.,A., Osborne, K., and Evans, R. (1987), Natural Language Processing in the 1980s.", CSLI, Stanford University.

Tomita, M. (1986). Efficient Parsing for Natural Language: A Fast Algorithm for Practical Systems, Boston: Kluwer.

Jinye Zhou, Shi-kuo Chang (1986), A Methodology for Deterministic Chinese Parsing, Computer Processing of Chinese & Oriental Languages, Vol. 2, No. 3 May 1986.

Lin-Shan Lee, Lee-Feng Chien, Longj-ji Lin, James Huang, K.-J. Chen (1991), An Efficient Natural Language Processing System Specially Designed for the Chinese Language, Computational Linguistics, Vol.17, No. 4, 1991

M. Zhou (1999), J-Beijing Chinese-Japanese Machine Translation System, Proceedings of JSCL, 312-319, Beijing, 1-3, Nov, 1999

Jingcun Wu, Xuechao Hou, Modern Chinese Syntactical Analysis, Beijing University Press, 1982.

Zhengsheng Luo, Changjian Sun, Cai Sun (1995), An Approach to the Recognition of predicated in the automatic analysis of Chinese sentence patterns, Advances and applications on Computational Linguistics, Tsinghua University Press

Chao, Y.R. (1968). A Grammar of Spoken Chinese, Berkeley, CA: University of California Press

M. Zhou, C.N.Huang, (1994) An Efficient Syntactic Tagging Toll for Corpora. Proc. COLING 94, Kyoto, pp. 945-955.

Huang, J. (1982). Logical relations in Chinese and the theory of grammar, Doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA.