A Hierarchical Parsing Approach with Punctuation Processing for Long Chinese Sentences

Xing Li
Institute of Automation, Chinese Academy of Sciences, Beijing 100080
xli@nlpr.ia.ac.cn

Chengqing Zong
Institute of Automation, Chinese Academy of Sciences, Beijing 100080
cqzong@nlpr.ia.ac.cn

Rile Hu
Institute of Automation, CAS, Beijing 100080
Nokia (China) Research Center, Hepingli Dongjie 11,Beijing 100013
rlhu@nlpr.ia.ac.cn

Abstract

In this paper, the usage and function of Chinese punctuations are studied in syntactic parsing and a new hierarchical approach is proposed for parsing long Chinese sentences. It differentiates from most of the previous approaches mainly in two aspects. Firstly, Chinese punctuations are classified as 'divide' punctuations and 'ordinary' ones. Long sentences which include 'divide' punctuations are broken into suitable units, so the parsing will be carried out in two stages. This 'divide-and-rule' strategy greatly reduces the difficulty of acquiring the boundaries of sub-sentences and syntactic structures of sub-sentences or phrases simultaneously in once-level parsing strategy of previous approaches. Secondly, a grammar rules system including all punctuations and probability distribution is built to be used in parsing and disambiguating. Experiments show that our approach can significantly reduce the time consumption and numbers of ambiguous edges of traditional methods, and also improve the accuracy and recall when parsing long Chinese sentences.

1 Introduction

Until recently, although punctuations are clearly important parts of the written Chinese, many Chinese parsing systems developed to date have simply ignored them. Some researches have been done on English punctuations in parsing [1, 2, 3, 4, 5], their researches have used plenty of theoretical and experimental facts to prove that it is effective to incorporate punctuation information into parsing of long complex sentences. But as far as we know, little work has been done in Chinese syntactic parsing.

Because the derivation of Chinese punctuations was referring to western language [3], they have many similarities in usage. Researches on Chinese punctuations in parsing will be valuable. However, our study shows, there are still differences between them, special research on Chinese punctuations is necessary.

In this paper, differences in English and Chinese punctuations are compared and the special difficulty and corresponding cause in parsing Chinese long sentences are analyzed. Then a new hierarchical parsing (HP) approach is proposed instead of traditional parsing (TP) method. This 'divide-and-rule' strategy greatly reduces the time consumption. Open test shows, parsing accuracy and recall of HP method are both about 7% higher than those of TP.

The remainder of this paper is organized as follows: Section 2 is related work. Section 3 mainly discusses the special difficulties and solution in parsing long Chinese sentences. Then HP method is discussed in detail in Section 4. Section 5 gives the final experiment results and corresponding analyses. Finally, the further work is expected.

2 Related Work

Nunberg's *The Linguistics of Punctuation* [2] is the foundation for most of the latter researches

in syntactic account of punctuation. In his important study, he advocates two separate grammars, operating at different levels. A lexical grammar accounts for the text-categories (text-clauses, text-phrases) occurring between the punctuation marks, and a text grammar deals with the structure of punctuation, and the relation of those marks to the lexical expressions they separate.

Based on above theory, Jones proposes his method which uses an integrated grammar. He divided punctuations into conjoining and adjoining punctuation. Conjoining punctuations can be used to indicate coordinate relationship between components. Adjoining punctuations, otherwise, only can be attached to their adjacent sentence components. In Jones' theory, in a sense, conjoining punctuation could also be treated under the adjunctive principle [3]. So, punctuations in his theory are still attached to adjacent lexical expressions. An integrated syntactic punctuation grammar is given.

Jones' method shows good modularity. However, the grammars he designed can only cover a subset of all punctuation phenomena. His experiment shows that when parsing a set of ten previously unseen punctuationally complex sentences, seven of the ten are unparsable!

In Chinese, Zhou Qiang[6] has used punctuations to do automatic acquisition of coordinate phrases. In machine translation, Chengqing Zong[7] and Huang He-yan[8] have used punctuations associating with relative pronouns to segment complex sentences into several independent simple sentences. Above all, none of previous work has carried out a thorough study on punctuations from the syntactic point of view.

3 Motivations

3.1 Differences between Chinese and English Punctuations

In Chinese, there are some punctuations which don't exist in English. The first one is a pair of Chinese book-name mark '《' and '》', which are obvious marks that the content between them must be name of a book. The second one is pause mark '、', which replaces comma as the separating mark between coordinate components. For instance, sentence "I like to walk, skip, and run." can be translated into Chinese one as "我喜欢走、跳、和跑。". Chinese pause mark is the evident mark with the exclusive usage is to separate coordinate words or simple phrases, so it is easier to get coordinate words or simple phrases in Chinese sentences.

3.2 Special Difficulty in Parsing Long Chinese Sentences

In essence, English is a kind of hypotaxis language, so an intact syntax structure denotes a sentence. When several simple sentences are connected to form a compound sentence, there should be obvious conjunctions between them. Differently, Chinese is a kind of parataxis language, and the language unit which expresses a complete thought is an intact Chinese sentence. Therefore, several sentences with associative meanings can be connected by come punctuations to form a compound one without any conjunctions. This type of sentence is called 'run-on sentence', and which is prevalent in Chinese. For example, we randomly selected 4431 sentences whose lengths are over 30 characters from a Chinese corpus named TCT 973.[1] There are 1830 run-on sentences, covering 41.3%. Chinese sentence "我现已步入中年，每天挤车，搞得我精疲力尽。" is this kind of sentence. The corresponding English meaning is "Now, I am not young and I still have to take bus to work everyday, which make me very tired". So, in above Chinese sentences, commas are used not only as separating marks of sub-sentences but also as separating marks of components in one sub-sentence. However, lack of connections makes methods [7, 8] of segmenting complex sentences invalid. In this situation, acquisition of the boundaries of sub-sentences and syntactic structure of sub-sentences or phrases should be done simultaneously in once-level parsing strategy, which will undoubtedly increase the difficulty of parsing long sentences.

3.3 Corresponding Solution

In order to solve this problem, a hierarchical parsing (HP) approach is proposed by us. Nunberg's theory of two categories grammars provides us the theoretical base of HP approach.

[1] Please refer to http://www.chineseldc.org/

According to his definition of two categories of grammars described in section 2, the two grammars can operate at different levels independently. Punctuations which can occur as elements of text grammar are defined by us as 'divide' punctuations. Then punctuations which can occur as elements of lexical grammar are 'ordinary' ones. The 'divide' punctuations can be used to divide the whole sentence into several parts. Then the parsing will be carried out in two steps. Thus, acquisition of syntactic structure of sub-sentences or phrases is done in the first level parsing, and acquisition of the boundaries of sub-sentences and relationship of sub-sentences or phrases can be done in second level parsing. This is the main idea of HP approach, which can reduce the difficulty of parsing run-on sentences and other types of compound sentences. The framework of HP approach is shown as following Figure 1:

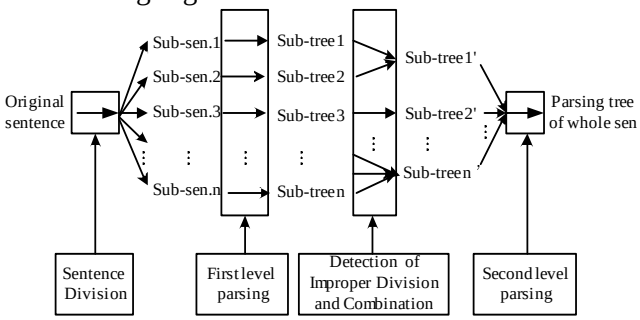

Figure 1. Framework of HP Approach

4 Hierarchical Parsing Approach

4.1 Classification of Chinese Punctuations

In this paper, the 'divide' punctuations are defined as follows: If lexical sentences or phrases which are separated by certain punctuations must be correlative to each other wholly not partly, these punctuations are in level of text grammar, which are classified as 'divide' punctuations. Punctuations in *a* and *b* of Figure 2 are examples of two categories of punctuations (P stands for punctuations).

In Chinese, the semicolon is used to separate coordinate sub-sentences. The colon is used as separation mark of interpretative phrases or sub-sentences from former sub-sentences. So, according to above definition, they can be classified as 'divide' punctuations. The comma, specially, can occur as a mark of coordinate phrases element. So, using of it as 'divide' punctuation may cause improper division problems and a compensatory solution is introduced in, which will be discussed in detail in Section 4.3.3.

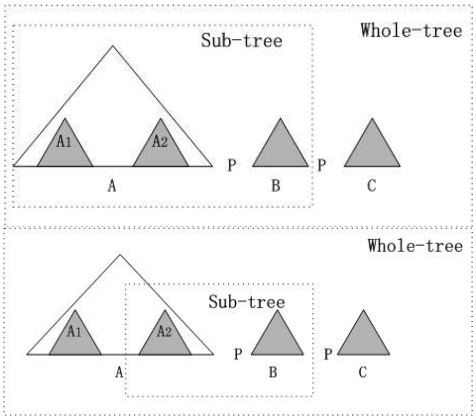

Figure 2. 'Divide' punctuations (first) and 'ordinary' punctuations (second)

4.2 Grammar Rules

The automatic extraction of grammar rules which include punctuations depends on large scales of parsed Chinese corpus which has ample syntactic phenomena and standard usage of punctuations. Fortunately, Chinese tree-bank named TCT 973 is such a corpus. It includes 1,000, 000 words and covers all kinds of text after 1990th. The average length of each sentence is 23.3 words. Long sentences of over 20 words length account for half of it.

Firstly, original grammar rules are extracted. Then generalizations are done about the use of the various punctuation marks from the rules set. For example, as mentioned before, Chinese book-name mark '《' and '》'are obvious marks that the content between '《' and '》'must be name of a book by any syntactic category. Therefore, a generalized rule can be deduced as below:

$$NP \rightarrow 《X》 \qquad X:\{NP, VP, S, PP.....\} \qquad (1)$$

In above generalized rule, *X* can be any POS of phrases or single word, so possible rules that have not been deduced from tree-bank are added into the grammar rules set with probabilities 1.

Except for above special situations, corresponding probabilities of all grammar rules are computed by Maximum Likelihood Estimate

(MLE) method.At last, all rules are combined to form an intact grammar system.

4.3 Parsing Strategy

4.3.1 Sentence Division

Depending on above classification, commas, semicolons and colons are used to divide sentences into a series of sub-sentences. Notice that quotation marks and parenthesis are treated as transparent and syntactically non-functional.

4.3.2 First Level Parsing

All sub-sentences and phrases gotten from the division processing are inputs of the first level parsing. A chart parsing algorithm is used here. The grammar rules and corresponding probabilities are used to do parsing and disambiguating. Then for all sub-sentences and phrases, their parsing trees are the highest probabilities ones of all possible trees.

4.3.3 Detection of Improper Division and Combination

Because of the specialty of comma, using of it as the division mark may cause improper divisions. The main causation is improper division between coordinate phrases which have been same component of the sentence. For example, Chinese sentence “我喜欢在春天去观赏桃花，在夏天去欣赏荷花，在秋天去观赏红叶，但更喜欢在冬天去欣赏雪景。” is a typical coordinate structure similar to “I like to do ..., to do ..., to do..., but I like better to...“ in English. So, the first three “喜欢”are coordinate predicates of the sentences. Then the improper division will break up this relationship. In this section, a detection and combination method is proposed by us to solve this problem in parsing Chinese sentences.

Because the lexical expressions surrounding punctuations are parsed in first level parsing, it is easy to get their internal syntactic structures information we need. Just a simple analysis procedure is needed to judge if there exists such a coordinate relationship between lexical expressions surrounding commas.

A description of the analysis strategy is given according to this example.

Just like Figure 3 shows, the components after the first comma are parsed as verb phrase (VP) marked as *B*. Obviously *B* is composed of a preposition phrase (PP) and a verb phrase. If there exists a minimal length of phrase immediately before the first comma and this phrase has totally the same structure to phrase *B*, then they are coordinate phrases. In Figure 3, A_2 is such a phrase. The components after other commas are analyzed similarly. Finally, A_2, *B* and *C* are coordinate phrases. Since the verb phrase *D* immediately after the second comma has obviously different structure from A_2, *B* and *C*, so they aren’t coordinate components. The part-of-speech tags throughout this paper follow the standard of TCT973.

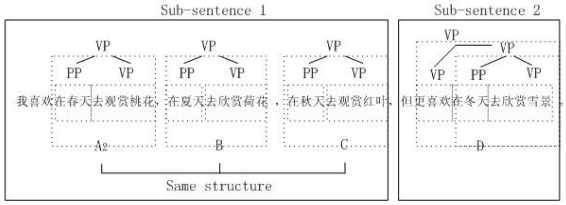

Figure 3. Syntactic structure of example sentence

Through the above analysis, we can see that the first and second commas are actually in level of lexical grammar, using them as ‘divide’ punctuations will cause the improper division as shown in Figure 2 of b. Therefore, we present a method to use sub-tree adjoining operation, firstly combine the sub-tree A_2 with tree *B* and *C*, then use the new tree A_2’ to replace original A_2 without changing original structure of *A* . Figure 4 shows the adjoining procedure.

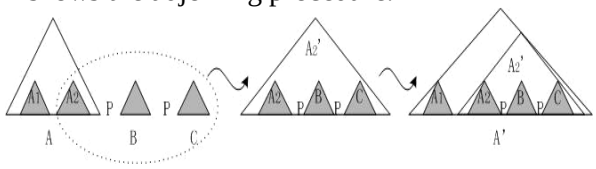

Figure 4. Sub-tree adjoining operation

Then the execution conditions and results of such adjoining operation are summarized as following rules:

$$[X,]^{+}S[X \quad Y...Y] \Rightarrow S[X[[X,]^{+}X] \quad Y...Y] \quad \textbf{(2)}$$

X = {NP, VP, AP, DP}, S stands for sentence,

Y = *(any legal POS)

$$S[Y...Y \quad X][, \ X]^{+} \Rightarrow S[Y...Y \quad X[X[, \ X]^{+}]] \quad \textbf{(3)}$$

X = {NP, VP, AP, DP}, S stands for sentence,

Y = *(any legal POS)

The execution conditions of both Rule (2) and (3) are defined as follows: all X should be coordinate phrases with the same syntactic categories.

4.3.4 Second Level Parsing

The parsing algorithm of this module is totally the same to the first level parsing; with the difference is the input string. At the first parsing stage, inputs are POS sequence of words, but at the second parsing stage, inputs are POS sequence of all sub-tree root nodes. After this stage of parsing, the best parsing trees of whole sentences will be constructed.

5 Performance Evaluation

5.1 Test Sentences

The primary aim of the HP strategy is to take use of the punctuation information to help to conquer the difficulty of parsing long sentences. Chinese sentences with over 20 words are generally regarded as long sentences. Therefore, we conduct experiments on the sentences with the length over 20 words.

Firstly, 8,059 sentences were chosen randomly from TCT 973 as train set. The 3,795 PCFG rules used in our system are extracted from the train set after generalizing. Then, for other 847 sentences, whose lengths are less than 20 words are filtered and 420 sentences are finally conserved as our open test data set. Distribution of these sentences is shown in Table 1 below:

Text Type	Number of Sen	Length of Sen (Words)	Average Length of Sen (Words)
Literature	116	21～123	36.06
News	123	22～100	37.73
Science	114	21～131	39.47
Practical writing	67	20～98	38.36
Total	420	20～131	37.84

Table 1. Distribution of test sentences

5.2 Efficiency Evaluation

In order to compare our HP approach with TP method of once-parsing algorithm, we do compared experiments using same data set in Table 1 and same grammar rules set.

5.2.1 Time Consumption Evaluation

Running two systems on a PC (Pentium 4, 1.20GHz, 256M of RAM), their time consumptions are shown in Figure 5.

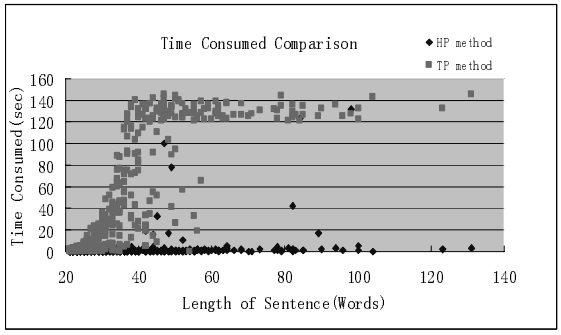

Figure 5. Time consumption

In our experiment system, we set the upper limit execution time as 120 seconds per sentence, judging at the end of every algorithm cycle. When parsing time of the sentence is overtime, the system will exit without getting final result.

Experiment results shown in Fig.5 prove that time efficiency of HP method is greatly superior to TP, especially when the sentence has more than 40 words. With the increasing of sentence length, it is more difficult for TP method to parse successfully.

5.2.2 Accuracy and Recall Evaluation

Firstly, Table 2 shows numbers of sentences failed to be parsed in two methods with the time limitation of 120 seconds.

Methods	Numbers of Test Sen	Numbers of Failed Sen	Ratio
TP	420	97	23.1%
HP	420	16	3.8%

Table 2. Ratio of failed sentences

It is evident that HP method can largely reduce failed sentences in given time limitation.

Then, except for failed sentences, only considering the successfully parsed sentences, the parsing accuracy and recall of the two methods should be compared. The standard PARSEVAL measures [9] are used to evaluate two methods. Results are shown in Table 3.

From Table 3, we can see that the total parsing accuracy and recall of HP method are both almost 7% higher than those of TP method. Amounts of average crossing brackets are also reduced greatly.

Analyzing data in Table 3, to different text types, the accuracy and improvement effect of TP method are slightly different. Sentences of literature text have the highest parsing accuracy and recall. Studied show that there are 97 'run-on sentences' in the 116 literature text sentences, covering 84%. The comparatively higher accuracy and recall of these sentences prove that our HP approach is effective.

Text type	Method	LP%	LR%	CBs	0CB%	≤2CBs%
Literature	TP	67.31	66.76	6.97	19.77	48.84
	HP	73.57	73.77	3.24	40.74	62.09
News	TP	61.05	61.69	5.80	10.47	34.88
	HP	70.66	70.58	3.52	28.33	61.83
Science	TP	61.20	60.89	5.63	12.66	37.97
	HP	68.74	68.98	4.14	23.37	59.10
Practical writing	TP	64.10	64.61	6.17	6.25	27.08
	HP	66.55	67.81	4.68	21.54	50.77
Total	TP	63.38	63.41	6.14	13.04	38.46
	HP	70.06	70.03	3.80	30.24	61.01

Table 3. Results using standard PARSEVAL measures

Sentences of application have lowest parsing accuracy and smallest improvement. Because comparing to other three types, sentences of this type have more long nested noun phrases or coordinate components, such as long organization names and commodity names, which will cause noun phrase combination disambiguation.

6 Conclusion and Future Work

This paper studies the usage and function of Chinese punctuations in syntactic parsing. A new hierarchical parsing approach is proposed. Besides, a grammar rules system including all punctuations and probability distribution is built to be used in parsing and disambiguation. Compared experiments prove that HP method is effective in long Chinese sentences parsing.
In future work, theories of punctuations should be studied further to get a more formal point of view.

Acknowledges

The research work is supported by the national natural science foundation of China under grant No.60375018 and 60121302, and also supported by the outstanding oversea scholar foundation of CAS.

References

1. Benard Jones, Towards a Syntactic Account of Punctuation. In Proceedings of the 16th International Conference on Computational Linguistics (COLING-96), Copenhagen, Denmark, August . (1996b)

2. Geoffrey Nunberg. The Linguistics of Punctuation. CSLI Lecture Notes, No. 18, Stanford CA, (1990)

3. Benard Jones,What's the Point? A (Computational) Theory of Punctuations. PhD thesis, Centre for Cognitive Science, Universito of Edinburgh, Edinburgh, UK, (1997)

4. Edward Briscoe. The Syntax and Semantics of Punctuation and its Use in Interpretation. In Proceedings of the ACL/SIGPARSE International Meeting on Punctuation in Computational Linguistics, Santa Cruz, California. (1996) 1–7.

5. Charles Meyer. A Linguistic Study of American Punctuation. Peter Lang: New York. 1987.

6. Zhou Qiang. The Chunk Parsing Algorithm for Chinese Language. In Proceedings of JSCL'99, (1999) 242-247

7. Chengqing Zong, Yujie Zhang, Kazuhide Yamamoto, Masashi Sakamoto,etc. Chinese Utterance Paraphrasing for Spoken Language Translation, In Journal of Chinese Language Computing, Singapore, 2002,12 (1): 63-77.

8. Huang He-yan, Chen Zhao-xiong, The Hybrid Strategy Processing Approach of Complex Long Sentence, In Journal of Chinese Information processing, 2002,16(3):1-7.

9. E.Charniak, "Statistical parsing with a context-free grammar and word statistics". In Proc of AAAI'97, 1997.