Proc. of the 1993 IEEE
Int.'l Conf. on Tools with AI,
Boston, Massachusetts, Nov. 1993

A Language Model for Parsing Very Long Chinese Sentences

Hsin-Hsi Chen

Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan 10764, R.O.C.
hh_chen@csie.ntu.edu.tw

Abstract

By corpus analyses, about seventy-five percent of Chinese sentences are composed of more than two sentence segments separated by commas or semicolons. A segment may be a sentence, a noun phrase, a verb phrase, an adjective phrase, an adverbial phrase, or a prepositional phrase. An NP segment may serve as a subject of the next segment or an object of the previous segment. The empty category pro may also appear in the VP segment. The maximal freedom of the uses of pros, the large number of segments, the various segment types, and the associativity problem make sentence parsing difficult. Few parsing systems deal with these problems. This paper regards a segment as a basic parsing unit. And it uses characteristic words, subcategories of verbs, topic chain and some heuristic rules to link the segments into meaningful units. The pro resolution and the segment linking are useful for practical applications.

1: Introduction

Punctuation marks play a significant role in natural language statements. They make texts quite clear and precise. The Chinese character string "下雨天留客天留我不留" is a famous example. It has two different interpretations shown below under different punctuations.

(1) 下雨，天留客；天留，我不留！
(As it is raining, the guest should be put up overnight; heaven wants to put up the guest, but I do not.)

(2) 下雨天，留客天，留我不？留！
(It is raining, heaven wants to put up the guest, should the guest be put up? Yes!)

English and Chinese natives have their own written styles. The following depicts the ratios of ten punctuation marks used in LOB corpus (about one million English words) and CKIP corpus (about ten million Chinese words):

	. ∘	,	、	;	:
LOB corpus	56958 36.44%	55366 35.42%	0 0.00%	3233 2.07%	2241 1.43%
CKIP corpus	105930 15.22%	422699 60.73%	55671 8.00%	5157 0.74%	9780 1.41%

	?	!	"..." 「...」	[...] (...)	–
LOB corpus	3457 2.21%	1030 0.66%	16690 10.68%	4667 2.99%	12665 8.10%
CKIP corpus	5448 0.78%	8803 1.26%	43953 6.32%	38572 5.54%	4 0.00%

The number of sentence terminators (period, question and exclamation marks) is larger than segment separators (comma and semicolon) in English. In contrast, the segment separators outnumber the sentence terminators in Chinese (7:2). It results in few and many segments in English and in Chinese sentences respectively. The following statistic verifies this point:

	1	2	3	4	5
LOB corpus	32163 52.34%	14119 22.98%	8111 13.20%	3670 5.97%	1749 2.85%
CKIP corpus	13171 10.96%	18724 15.58%	19225 16.00%	17176 14.29%	14687 12.22%
	6	7	8	9	10+
LOB corpus	789 1.28%	401 0.65%	178 0.29%	114 0.19%	151 0.25%
CKIP corpus	11604 9.66%	8610 7.16%	5860 4.88%	3922 3.26%	7202 5.99%

About seventy-five percent of Chinese sentences are composed of more than two segments. A segment may be simple as a word, complex as a phrase or a sentence. As a phrase, it may be a noun phrase (NP), a verb phrase (VP), an adjective phrase (ADJP), an adverbial phrase (ADVP), or a prepositional phrase (PP) [8]. An NP segment may serve as a subject of the next segment or an object of the previous segment. The empty category *pro* may also appear in the VP segment. These linguistic phenomena show the difficulty in parsing Chinese sentences and long English ones. Few papers touched on the effects of punctuation marks in the natural language processing systems. This paper will propose a parsing system to resolve the problems introduced by the large number of segments.

2: Associativity problem

A few parsers have been presented for Mandarin Chinese [1,4,7]. They dealt with sentences with no punctuation marks, or with one separator and one terminator. When they are used, the punctuation marks

should be dropped out before parsing. It results in ambiguous sentences or very long sentences. Thus, the segments separated by commas or semicolons should be considered as basic parsing units rather than the whole sentences. However, how to link the related segments into larger ones for further applications becomes a new problem. It is very serious when the number of segments in a sentence is large. Consider a three-segment case. There may be three possible linkages: (S1 S2 S3), (S1 (S2 S3)) and ((S1 S2) S3). The first three sentences in the following show the three linkages respectively. The last two use the same characteristic word '所以' (so) in the final segment to represent a cause-effect relationship, but they have different scopes. The selection of correct linking is called an *associativity* problem.

(3) ([S1 他$_i$駕駛著太空梭]，[S2 e$_i$在太空中繞著月球飛行]，[S3 e$_i$等待這兩個人完成工作])。
([S1 He$_i$ drove the space shuttle] [S2 and e$_i$ flew around the moon], [S3 e$_i$ waiting for these two men completing their jobs]).

(4) ([S1 他不在家]，([S2 所以我們沒有找到他]，[S3 就馬上去車站]))。
([S1 He was not at home], ([S2 so we did not find out him] [S3 and went to station immediately])).

(5) (([S1 我們沒有找到他]，[S2 就馬上去車站])，[S3 所以他很生氣])。
(([S1 We did not find out him] [S2 and went to station immediately]), [S3 so he was very angry]).

(6) ([S1 他嚴肅地告訴我們]，([S2 我們沒有找到他]，[S3 所以他很生氣]))。
([S1 He seriously told us] (that [S2 we did not find out him], [S3 so he was very angry])).

3: Linguistic knowledge

Four kinds of linguistic knowledge - punctuation marks, categories of segments, linking elements and topic chains, are used to disambiguate the segment linking.

3.1: Punctuations marks

There are fourteen marks in Mandarin Chinese [8]. Only period, question mark, exclamation mark, comma, semicolon and caesura sign are discussed in this paper. The former three are sentence terminators, and the latter three are segment separators. Period is placed at the end of a sentence to indicate that the meaning of a sentence is complete. Question mark is used to express the question, doubt, argument, or even astonishment. Exclamation mark shows the writer's feeling, e.g., happy, angry, or sad. Comma has multiple functions. Its meaning is more difficult to identify. It may be used to separate some juxtaposed clauses or phrases. Chinese natives usually use it at random. Semicolon indicates the juxtaposed or the contrast clauses. Caesura sign, a Chinese specific punctuation mark, shows the shortest suspend. It may appear in the following two cases: i) the two segments neighbor to the sign have the same category (see sentence (7)); ii) the categories of a segment and part of its neighbor segment are the same (see sentence (8)).

(7) 他看見森林裡的大樹，強壯、清澈。
(He saw the great trees in the forest, which are strong and pure.)

(8) 他做不成大公雞、小綿羊，也做不成大白鵝、小鳥兒，只好躲起來。
(He can neither act as a big cock nor a little sheep. He cannot act as a big white goose or a little bird either. Thus, he just hides himself.)

3.2: Categories of segments

As we know, a segment may be a clause or a part of a clause. S and VP are clausal segments. The segments in the examples (3-6) are such ones. NP, ADJP, ADVP and PP are non-clausal segments. Below shows two NP segments act as objects of the verb in the first segment.

(9) 我們養了一隻狗，一個小猴子，一頭貓。
(We keep a dog, a monkey and a cat.)

For a non-clausal segment, we try to find the clausal segment that governs it. In the example (9), the two NP segments and the NP object in the first segment are juxtaposed. They have the same behavior, so that these two NP segments belong to the first one. Besides such a juxtaposition, a long subject or object is often written as a segment. The second segment in sentence (10) is a complex NP object.

(10) 你先要衡量，你願被人瞭解與能被人瞭解的程度。
(At first, you should measure the extents that you want to be understood and that you are able to be understood.)

The subcategory of the verb '衡量' (measure), i.e., transitive verb, tells us there is a missing object in the first segment. It provides some clue to determine to which clausal segment an NP segment belongs. The treatments of the other non-clausal segments are simple. ADJP segments are always short and juxtaposed. They and the nearest NP segment form a larger one. Then it is treated in the same way as the usual NP segment. ADVP (PP) segment modifies the following segment, which is a clause.

3.3: Linking elements

For a clausal segment, we try to find the relationship with other clausal segment(s). The explicit linking elements in the segments are important knowledge to determine these relationships. There are three kinds of linking elements [5]: forward-linking elements, backward-linking elements and couple-linking elements. A segment with a forward-linking (backward-linking) element is linked with its next (previous) segment. A couple-linking

element is a pair of words that exist in two segments. Apparently, these two segments are joined together. Sentences (11-13) show examples for each kind of linkings respectively.

(11) forward linking
下課之後，我要去看電影。
(After I get out of class, I go to the movies.)

(12) backward linking
我本來想去看電影，可是我沒有買到票。
(I had originally intended to go to the movies, but I didn't buy a ticket.)

(13) couple linking
因為我沒有買到票，所以我沒有去看電影。
(Because I didn't buy a ticket, I didn't go to the movies.)

Linking elements have lexical categories adverb and/or conjunctive. Some linking elements may serve as a forward linking in one case (see (14)), and serve as a backward linking in another case (see (15)).

(14) 因為今天天氣不好，我們明天才起程。
(Because the weather is bad today, we will start on journey tomorrow.)

(15) 我們明天才起程，因為今天天氣不好。
(We will start on journey tomorrow, because the weather is bad today.)

Some forward- or backward-linking elements with other words form couple-linking elements. The word '因為' (because) is a typical example (see (13)). Thus, we further classify the linking elements into four types: purely forward, purely backward, dual and couple. Current experimental parsing system adopts the following linking elements selected from [5,6].

i) 4 pure forward-linking elements
任憑 (no matter), 尚且 (still), 既然 (since), 雖 (though).

ii) 37 pure backward-linking elements
一邊 (while V-ing), 於是 (therefore), 而且 (besides), 不然 (otherwise), 反而 (instead), 以至 (result in), 以便 (so that), *etc.*

iii) 27 dual-linking elements
不過 (but), 因為 (because), 如果 (if), 的話 (if), 雖然 (though), 儘管 (even if), 由於 (since), *etc.*

iv) 108 couple-linking elements
一邊…一邊… (while V-ing, V-ing), 不但…而且… (not only, but also), 因為…所以… (because), 如果…則… (if ... then), 既然…就… (since ... then), *etc.*

The word '而且' (besides) is a pure backward-linking element and a part of a couple-linking element.

3.4: Topic chain

The topic of a clausal segment is deleted under the identity with a topic in its preceding segment. The result of such a deleting process is a *topic chain* shown as example (3). The following demonstrates the statistic of the uses of *pro* in an elementary school corpus which contains 12 texts.

pro position	antecedent position	total
subject	subject	329 (87.97%)
	object	35 (9.36%)
	prepositional object	10 (2.67%)

The table depicts that 87.97% of the zero subjects refer to the antecedents which are at the subject position of the previous segments. Thus, we have the following postulation:

given two VP segments, or one S and one VP segments, if their expected subjects (external arguments of verbs) are unifiable, then the two segments can be linked.

4: A new parsing system

A complete parsing system is composed of three major modules: preprocessing, segment-parsing and post-processing.

4.1: Preprocessing

The tasks of this module are: to divide the input sentences into a sequence of segments on the basis of punctuation marks; to assign a unique index to each segment; to check the existence of the linking elements; to identify their types; to retract the linking elements from segments. The type identification procedure of linking elements is shown as follows. Assume there are n segments.

i) Scan the segments from left to right and stop at the segment i ($1 \le i \le n$) with linking element e.

ii) The word e is a part of a couple-linking element.
Find its right couple position j from segments (i+1) to n. Do the type identification procedure on the partition (i+1) to (j-1).

iii) The word e is a pure-linking element.
Link this segment with its right (left) neighbor if it is a forward- (or backward-) linking element.

iv) The word e is a pure-linking element and a part of a couple-linking element.
Execute step ii) first. If it fails, then e is a pure-linking element.

v) The word e is a dual-linking element and a part of a couple-linking element.
The step ii) is performed first. If it fails, then e is a dual-linking element. If i is the first (or the last) segment, e is a forward- (or backward-) linking element. If segment i and segment n (i.e., the last segment) belong to the same topic chain, then e is a backward-linking element. Otherwise, it is still ambiguous.

For each segment, a six-element list [*Category*, *Index*, *Mark*, *Word*, *Attribute*, *Trace*] is used to record the necessary information. When the preprocessing is completed, four of the six items are known and shown as follows. *Category* and *Trace* are not available before segment-parsing.

i) Index: the segment identifier

ii) Mark: the punctuation mark following the segment
iii)Word: the linking element in the segment
iv)Attribute: the type of linking element and its related segment. It has the following possible forms: [lc, Number] where *lc* denotes a left couple linking element, and *Number* indicates the position of its corresponding right couple; *pf* (pure forward); *pb* (pure backward); *rc* (right couple); *none* (no linking element).

4.2: Segment-parsing

A sentence segment is considered as a basic parsing unit. Because a segment may be S, NP, VP, ADJP, ADVP or PP, we adopt a Prolog-based left-corner bottom-up parsing system with top-down expectation, history-record and some mechanism for movement transformations [3]. Given a segment, the parser tries to construct a maximal projection as possible as it can. Top-down expectation is significant only in the local domain because the start-symbol is unknown before parsing. History record is helpful if a linking element has other categories than conjunctive and adverb. The mechanism for movement transformations generates a trace if an NP is absent in its proper position. The NP is empty for the following two reasons: i) it is deleted in the topic chain (see example (3)), and ii) it is regarded as a separated segment like example (10). The trace information is useful to link the related segments.

4.3: Postprocessing

Postprocessing can be divided into four modules: to expand the multiple solutions of segments to all possible combinations; to link two neighbor segments according to punctuation marks, categories, linking elements and topic chains; to link all the related segments for the whole sentence; to generate the parsing tree(s).

4.3.1: Expanding the solutions Ambiguity causes multiple solutions of a segment. Assume there are m segments with n1, n2, ..., nm solutions. There are n1 x n2 x ... x nm combinations. Each combination corresponds to a complete parsing tree.

4.3.2: Grouping the segments The four kinds of linguistic knowledge - punctuation marks, categories of segments, linking elements and topic chain, are applied in a predetermined order shown as Figure 1. These four criteria have different strengths for segment grouping. The lower the linguistic knowledge is, the weaker linkage strength it has. For instance, assume two segments can be linked by linking elements and topic chains simultaneously. Because linking elements are explicit information and topic chain is just a postulation, grouping by linking elements first is preferred. Thus, the topic chain may be ignored. In this four-pass module, each pass

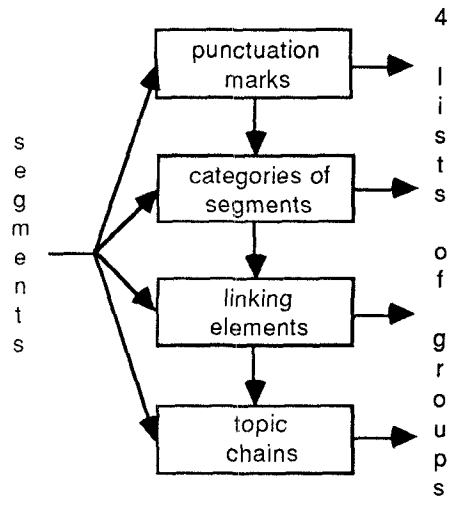


Figure 1. Control hierarchy of segment grouping

records the generated group(s). Because the linkage strength of segments of the earlier pass is stronger than that of the later one, a new group must be checked to make sure that it does not appear in the other lists.

The first pass is based on punctuations whose meanings were discussed in the previous section. Two segments beside a caesura sign must be related. Linking these two segments is reasonable. If segment *i* is followed by a caesura sign, [i,i+1] will be inserted to the list of linkage. The module just records the group no matter what categories these two segments have. The consideration of category information delays to the phase of parsing tree generation.

The second pass of linkage is to compare the categories of the two adjacent segment. The following shows some rules to link segments *i* and *i*+1.

i) If segment *i* is an NP, and segment *i*+1 is a VP,
then a sentence will be generated.
ii) If segment *i* is a VP, segment *i*+1 is an NP, and the VP has a Trace with category NP,
then a complete VP will be generated.
iii)If segment *i* is an S, segment *i*+1 is an NP, and the S has a Trace with category of NP,
then a complete S will be generated.
iv)If segment *i* is an ADVP (PP),
then it is linked with its successive segment.
v) If segment *i* is a VP, and segment *i*+1 is an ADJVB,
then they will form a new VP.
vi)If segment *i* is an NP, and segment *i*+1 is an NP,
then these two NPs will form a juxtaposed NP.

The third pass generates another list by linking elements. Note that the information list [Category, Index, Mark, Word, Attribute, Trace] is available after segment-parsing. *Attribute* records the attributes of linking elements. The following lists the grouping rules:

i) If *Attribute* of segment *i* is *pf*,
then a group [i,i+1] will be produced.

ii) If *Attribute* of segment *i* is *pb*,
then a group [i-1,i] will be produced.
iii)If *Attribute* of segment *i* is [lc,Number],
then a group [i,Number] will be produced.
iv)If *Attribute* of segment *i* is *rc* or *none*,
then nothing will be produced.

The last pass is operated under the control of topic chain. The statistic tells us that the lacked subject of a sentence can be found at the subject position of its previous segment. This is not always true. The desired subject may be farther. Consider example (16).

(16) 他去逛街，一不小心，小狗走失了，非常傷心。
(He went shopping, and lost his dog because of carelessness, he was very sad.)

When a topic chain is constructed, only two adjacent segments which are S and VP, or VP and VP are considered. The final results are in terms of equivalence classes. Assume [i,i+1], [i+1,i+2], and [i+2,i+3] form three topic chains. These chains constitute an equivalence class {i, i+1, i+2, i+3}.

4.3.3: Integrating the groups Four lists, i.e., MarkList, CategoryList, LinkingElementList, and TopicChainList, are generated. A segment identifier may appear in more than one group. Figure 2 shows how to integrate these lists into new one(s).

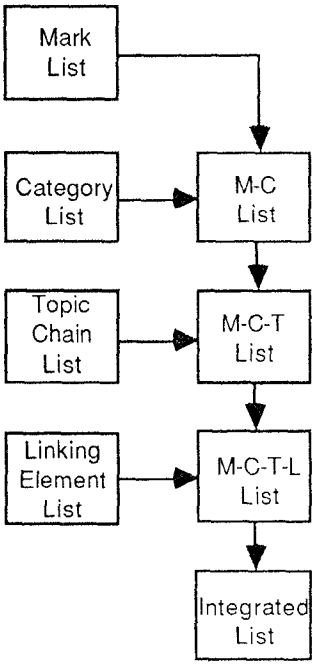
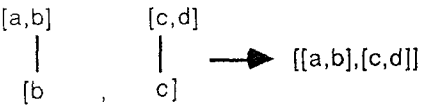


Figure 2. Hierarchy of list integration

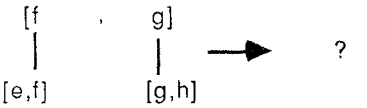
The procedure focuses on the replacement of overlapped indices. Given group [a,b] in one list and group [c,d] in another list, we will integrate these two groups into one if b=c or a=d. Let's see the case b=c. We may have two possible replacements: [a,[c,d]] or [[a,b],d]. In the former, b is substituted with [c,d]. In the latter, c is replaced by [a,b]. Which one is better? The hierarchy in Figure 2 depicts the association strength:

MarkList > CategoryList > TopicChainList > LinkingElementList

A strong association of two segments means these segments cannot be divided easily. The index in the group with weaker association will be replaced by the whole group with stronger association. Assume [a,b] is from MarkList and [c,d] is from CategoryList. When b=c, [[a,b],d] will be produced. When a=d, [c,[a,b]] will be the result. Figure 3 shows two more complicated examples. Let the upper groups have stronger association.



(a) Replacement of overlapping indices



(b) Ambiguous replacements

Figure 3. Complicated examples

The replacement in Figure 3(a) is no problem. The two upper groups [a,b] and [c,d] with stronger association replace *b* and *c* in the lower group respectively. An ambiguity occurs in Figure 3(b). Items of different lower groups can be replaced by the same group. Assume [f,g] is from TopicChainList, and [e,f] and [g,h] are from LinkingElementList. The item *f* in [e,f] or the item *g* in [g,h] may be substituted with [f,g]. All the two alternatives obey the specified criterion. A heuristic rule shown as follows is used to select the preferred solution:

The left linking element usually has a wider scope than the right linking element.

In Figure 3(b), [e,[f,g]] is produced first, and then *g* in [e,[f,g]] is replaced by [g,h]. The final solution is [e,[f,[g,h]]].

4.3.4: Generating parsing trees Each element in the integrated list shown in Figure 2 denotes an independent parsing tree. If there is more than one element in the list, then the sentence is composed of two or more independent parsing trees. Refer to example (17).

(17) [S1 有一個年約四歲的小女孩]，[S2 穿著紅色的衣服]，[S3 昨天在街上走失了]，[S4 她的父母十分著急]，[S5 請大家幫忙找尋]。
(A little girl of about four years age who wore red clothes was got lost on the street yesterday. Her parents were very worried, so they asked us to find her.)

The integrated list will be [[S1,S2,S3],[S4,S5]]. Incorrect use of segment separator, comma in particular,

is one major reason. The other comes from: the linking information is over syntactic level.

For those juxtaposed segments like sentence (8), additional operations shown in Figure 4 are needed.

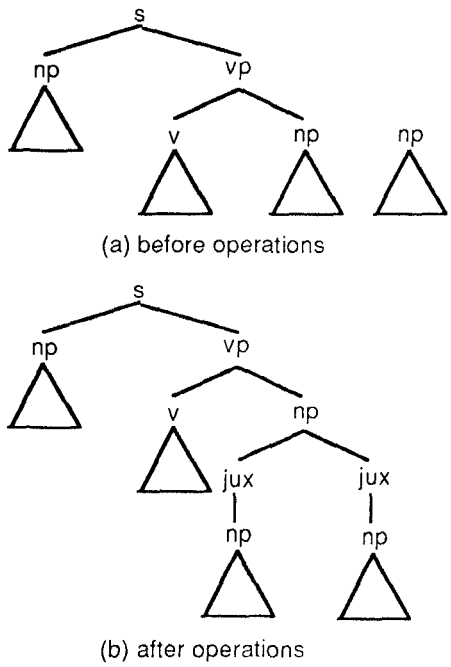


Figure 4. Partial parsing trees for sentence (8)

5: Experimental results

The following sections will demonstrate some typical examples. For each example, the results of preprocessing, grouping and integrating are shown. The system is implemented with Prolog and C, and runs on Sun Sparc Station. Prolog codes form the kernel, and C programs support the dictionary maintenance and window the controls.

5.1: Topic chains

Sentence (18) is composed of one clause and six VPs without any explicit linking elements. The subjects of the six VPs refer to the same pronoun.

(18) 他還是照常耕作，照常割草，照常灌溉，照常施肥，盡自己的力，做自己的事，並沒有特別注意這顆種子。

(He worked, cut grass, irrigated, and applied fertilizer as usual, he did his best to the work, he did not pay any special attention to this seed.)

The result of preprocessing is:

[[[1, ', e, none], 他, 還是, 照常, 耕作], [[2, ', e, none], 照常, 割草], [[3, ', e, none], 照常, 灌溉], [[4, ', e, none], 照常, 施肥], [[5, ', e, none], 盡, 自己, 的, 力], [[6, ', e, none], 做, 自己, 的, 事], [[7, 。, e, none], 並, 沒有, 特別, 注意, 這, 顆, 種子]]

The results of grouping and integrating are:

PunctuationList: []
CategoryList: []
LinkingElementList: []
TopicChainList: [1,2,3,4,5,6,7]
IntegratedList: [[1,2,3,4,5,6,7]]

The original topic chains are: [1,2], [2,3], [3,4], [4,5], [5,6], and [6,7]. They form an equivalence class [1,2,3,4,5,6,7].

5.2: Punctuation marks

Consider sentence (8) again. The result after preprocessing is:

[[[1, 、, e, none], 他, 做不成, 大, 公雞], [[2, ', e, none], 小, 綿羊], [[3, 、, e, none], 也, 做不成, 大, 白鵝], [[4, ', e, none], 小, 鳥兒], [[5, 。, e, none], 只好, 躲起來]]

The two nominal elements '大公雞' (a big cock) and '小綿羊' (a little sheep) are juxtaposed because of caesura sign, so are '大白鵝' (a big white goose) and '小鳥兒' (a little bird). They all are the objects of verb '做不成' (cannot act as). The results of grouping and integrating are:

PunctuationList: [[1,2],[3,4]]
CategoryList: []
LinkingElementList: []
TopicChainList: [[1,3,5]]
IntegratedList: [[[1,2],[3,4],5]]

5.3: Categories of segments

Sentence (9) contains three segments. The first one is an S segment, and the others are NP segments juxtaposed with the object in the first segment. The results are shown as follows:

[[[1, ', e, none], 我們, 養, 了, 一, 隻, 狗], [[2, ', e, none], 一, 個, 小, 猴子], [[1, 。, e, none], 一, 頭, 貓]]

PunctuationList: []
CategoryList: [[1,2],[2,3]]
LinkingElementList: []
TopicChainList: []
IntegratedList: [[1,[2,3]]]

5.4: Linking elements

Sentence (19) is a very long sentence, including seven segments.

(19) 由於他知道亡國的恥辱，一心一意要雪恥復國，所以他刻苦自勵，發憤圖強，親自率領軍隊操作，而且和人民一起奮鬥，對一切事情都以身作則。

(He knew the shame of national doom, he was bent on recovering his country. So, he endured hardship and strived for progress with determination. He leaded the army to work by himself. And he struggled with his citizens. He showed the best demonstration of everything.)

The preprocessing result is:

[[[1, ' , 由於, [lc, 3]], 他, 知道, 亡國, 的, 恥辱], [[2, ' , e, none], 一心一意, 要, 雪恥, 復國], [[3, ' , 所以, rc], 他, 刻苦自勵], [[4, ' , e, none], 發憤圖強], [[5, ' , e, none], 親自, 率領, 軍隊, 操作], [[6, ' , 而且, pb], 和, 人民, 一起, 奮鬥], [[7, ° , e, none], 對, 一切, 事情, 都, 以身作則]]

The first two segments are the reasons of the next five segments. The couple linking element '由於...所以' (because ... so) generates group [1,3]. Another pure backward linking element '而且' (and) generates group [5,6]. The groups [1,2], [3,4], [4,5] and [6,7] are generated by topic chains. Of these, [3,4] and [4,5] form an equivalence class [3,4,5]. They are shown as follows:

PunctuationList: []
CategoryList: []
LinkingElementList: [[1,3],[5,6]]
TopicChainList: [[1,2],[3,4,5],[6,7]]
IntegratedList: [[[1,2],[3,4,[5,[6,7]]]]]

Figure 5 demonstrates a complete parsing tree of this sentence.

6: Concluding remarks

This paper proposes a new parsing system for Mandarin Chinese. It considers the effects of the punctuations marks in Chinese sentences. The retrieval of characteristic words in advance makes the grammar rules simpler and facilitate the segment parsing. The segment linking groups the related segments and attaches the parsing trees to proper positions. They are useful to practical natural language applications. Machine translation is a typical example.

From the corpus analyses, English and Chinese have different written styles. English sentences consist of small number of segments. On the contrary, Chinese sentences are often very long. It is no problem in English-Chinese machine translation. The style of source sentence (i.e. English) has an effect on the generation of the target sentence (i.e. Chinese). In Chinese-English machine translation, this difference is critical. Because the parsing system can compose the related segments into meaningful units, they rather than the whole sentence can be considered as basic translation units during Chinese-English machine translation. The application of topic chain rule not only links the related segments, but identifys the co-referential relationship between an anaphor and its antecedent. Because Chinese demonstrates the maximal freedom of the uses of empty anaphors, it can use empty anaphor to refer to some element mentioned in the context. In English, if we do not place some overt pronoun at the empty site, the sentence may be unacceptable [2]. Therefore, the co-referential relationship is also useful for machine translation.

Acknowledgement

Research on this paper was partially supported by National Science Council grant NSC-82-0408-E002-405. The CKIP corpus is provided by CKIP, Academia Sinica, Taipei, Taiwan, R.O.C.

References

[1] Hsin-Hsi Chen, "A Logic-Based Government-Binding Parser for Mandarin Chinese," *Proceedings of the 13th International Conference on Computational Linguistics*, Vol. 2, 1990, Helsinki, Finland, pp. 48-53.

[2] Hsin-Hsi Chen, "The Transfer of Anaphors in Translation," *Literary and Linguistic Computing*, Vol. 7, No. 4, Oxford University Press, 1992, pp. 231-238.

[3] Hsin-Hsi Chen, I-Peng Lin and Chien-Ping Wu, "A New Design of Prolog-Based Bottom-UP Parsing System with Government-Binding Theory," *Proceedings of the 12th International Conference on Computational Linguistics*, 1988, Budapest, Hungary, pp. 112-116.

[4] Lin-Shan Lee, Lee-Feng Chien, *et al.*, "An Efficient Natural Language Processing System Specially Designed for the Chinese Language," *Computational Linguistics*, Vol. 17, No. 4, 1991, pp. 347-374.

[5] Charles N. Li and Sandra A. Thompson, *Mandarin Chinese: A Functional Reference Grammar*, University of California Press, 1981.

[6] Shuxiang Lu (呂叔湘), *Eight Hundred Words in Contemporary Mandarin Chinese* (現代漢語八百詞), Business Publishing Company (商務印書館), Hong Kong, 1984.

[7] Yiming Yang, "Combining Prediction, Syntactic Analysis and Semantic Analysis in Chinese Sentence Analysis," *Proceedings of the 10th International Joint Conference on Artificial Intelligence*, Vol. 2, 1987, pp. 679-681.

[8] Yuan Yang (楊遠), *The Research on Punctuation Marks* (標點符號研究), Tien-Chien Publishing Company (天健出版社), Hong Kong, 1981.

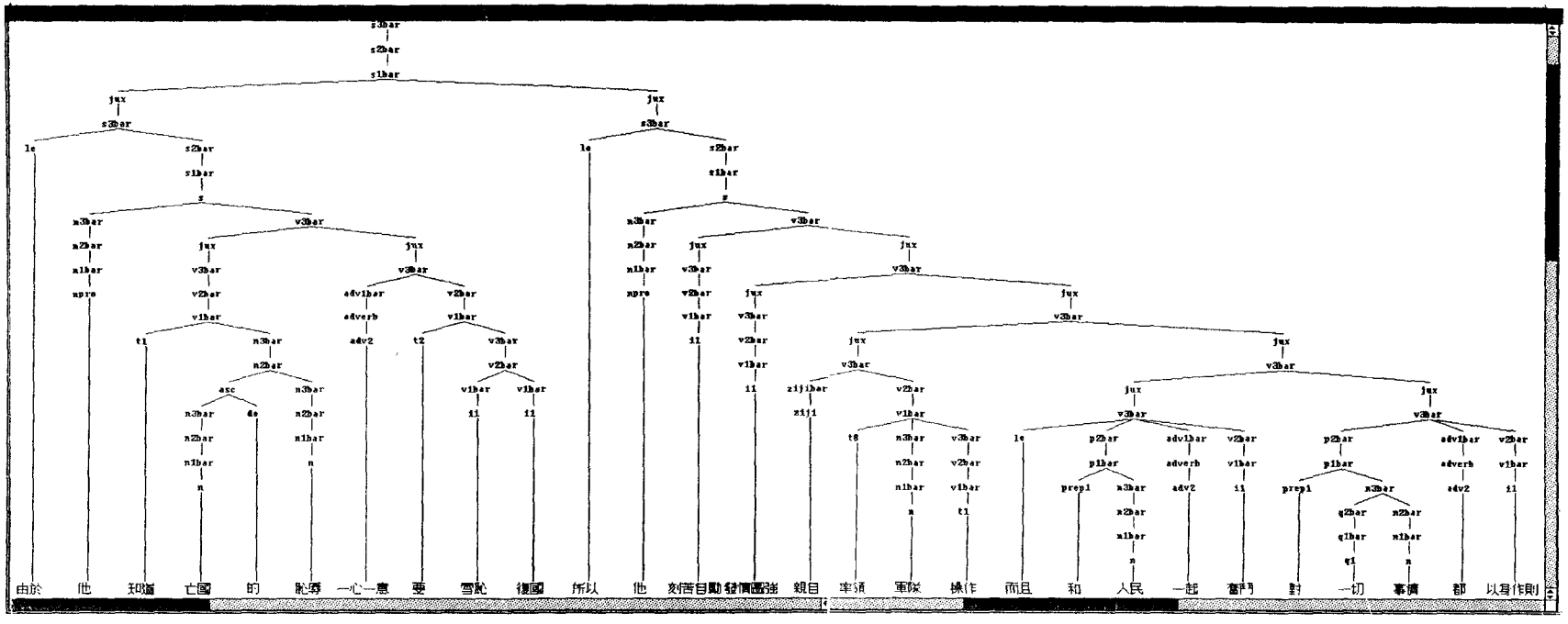


Figure 5. Parsing tree of sentence (19): "由於他知道亡國的恥辱，一心一意要雪恥復國，所以他刻苦自勵，發憤圖強，親自率領軍隊操作，而且和人民一起奮鬥，對一切事情都以身作則。"