USE OF HEURISTIC KNOWLEDGE IN CHINESE LANGUAGE ANALYSIS

Yiming Yang, Toyoaki Nishida and Shuji Doshita
Department of Information Science,
Kyoto University,
Sakyo-ku, Kyoto 606, JAPAN

ABSTRACT

This paper describes an analysis method which uses heuristic knowledge to find local syntactic structures of Chinese sentences. We call it a preprocessing, because we use it before we do global syntactic structure analysis[1] of the input sentence. Our purpose is to guide the global analysis through the search space, to avoid unnecessary computation.

To realize this, we use a set of special words that appear in commonly used patterns in Chinese. We call them "characteristic words". They enable us to pick out fragments that might figure in the syntactic structure of the sentence. Knowledge concerning the use of characteristic words enables us to rate alternative fragments, according to pattern statistics, fragment length, distance between characteristic words, and so on. The preprocessing system proposes to the global analysis level a most "likely" partial structure. In case this choice is rejected, backtracking looks for a second choice, and so on.

For our system, we use 200 characteristic words. Their rules are written by 101 automata. We tested them against 120 sentences taken from a Chinese physics text book. For this limited set, correct partial structures were proposed as first choice for 94% of sentences. Allowing a 2nd choice, the score is 98%, with a 3rd choice, the score is 100%.

1. THE PROBLEM OF CHINESE LANGUAGE ANALYSIS

Being a language in which only characters (ideograms) are used, Chinese language has specific problems. Compared to languages such as English, there are few formal inflections to indicate the grammatical category of a word, and the few inflections that do exist are often omitted.

In English, postfixes are often used to distinguish syntactical categories (e.g. translation, translate; difficult, dificulty), but in Chinese it is very common to use the same word (characters) for a verb, a noun, an adjective, etc.. So the ambiguity of syntactic category of words is a big problem in Chinese analysis.

In another example, in English, "-ing" is used to indicate a participle, or "-ed" can be used to distinguish passive mode from active. In Chinese, there is nothing to indicate participle, and although there is a word, "被", whose function is to indicate passive mode, it is often omitted. Thus for a verb occurring in a sentence, there is often no way of telling if it transitive or intransitive, active or passive, participle or predicate of the main sentence, so there may be many ambiguities in deciding the structure it occurs in.

If we attempt Chinese language analysis using a computer, and try to perform the syntactic analysis in a straightforward way, we run into a combinatorial explosion due to such ambiguities. What is lacking, therefore, is a simple method to decide syntactic structure.

2. REDUCING AMBIGUITIES USING CHARACTERISTIC WORDS

In the Chinese language, there is a kind of word (such as preposition, auxiliary verb, modifier verb, adverbial noun, etc..), that is used as an independant word (not an affix). They usually have key functions, they are not so numerous, their use is very frequent, and so they may be used to reduce ambiguities. Here we shall call them "characteristic words".

Several hundreds of these words have been collected by linguists[2], and they are often used to distinguish the detailed meaning in each part of a Chinese sentence. Here we selected about 200 such words, and we use them to try to pick out fragments of the sentence and figure out their syntactic structure before we attempt global syntactic analysis and deep meaning analysis.

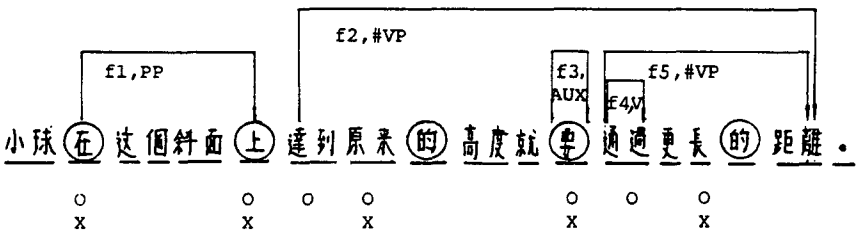
The use of the characteristic words is described below.

a) Category decision:

Some characteristic words may serve to decide the category of neighboring words. For example, words such as "了", "過", "着", "得", are rather like verb postfixes, indicating that the preceding word must be a verb, even though the same characters might spell a noun. Words like "要", "会", can be used as both verb and auxiliary. If, for example, "会" is followed by a word that could be read as either a verb or a noun, then this word is a verb and "会" is an auxiliary.

b) Fragment picking

In Chinese, many prepositional phrases start



Translation: The ball must run a longer distance before returning to the initial altitude on this slope.

—— : distinguish a word from others

◯ : characteristical word

⌐¬ : fragment

o : verb or adjective

x : the word can not be predicate of sentence

Fig.1 An Example of Fragment Finding

with a preposition such as "在", "到", "向", and finish on a characteristic word belonging to a subset of adverbial nouns that are often used to express position, direction, etc.. When such characteristic words are spotted in a sentence, they serve to forecast a prepositional phrase. Another example is the pattern "...是...的", used a little like "... is to ..." in English, so when we find it, we may predict a verbal phrase from "是" to "的", that is in addition the predicate VP of the sentence.

These forecasts make it more likely for the subsequent analysis system to find the correct phrase early.

c) Role deciding

The preceding rules are rather simple rules like a human might use. With a computer it is possible to use more complex rules (such as involving many exceptions or providing partial knowledge) with the same efficiency. For example, a rule can not usually with certainty decide if a given verb is the predicate of a sentence, but we know that a predicate is not likely to precede a characteristic word such as "的" or "是" or follow a word like "的", "是" or "所". We use this kind of rule to reduce the range of possible predicates. This knowledge can be used in turn to predict the partial structure in a sentence, because the verbal proposition begins with the predicate and ends at the end of the sentence.

In the example shown in Fig.1, fragments f3 and f4 are obtained through step (a) (see above), f1 through (b), and f2 and f5 through (c). The symbol "o" shows a possible predicate, and "x" means that the possibility has been ruled out. Out of 7 possibilities, only 2 remained.

3. RESOLVING CONFLICT

The rules we mentioned above are written for each characteristic word independantly. They are not absolute rules, so when they are applied to a sentence, several fragments may overlap and thus be incompatible. Several combinations of compatible fragments may exist, and from these we must choose the most "likely" one. Instead of attempting to evaluate the likelihood of every combination, we use a scheme that gives different priority scores to each fragment, and thus constructs directly the "best" combination. If this combination (partial structure) is rejected by subsequent analysis, back-tracking occurs and searches for the next possibility, and so on.

Fig.2 shows an example involving conflicting fragments. We select f3 first because it has the highest priority. We find that f2 , f4 and f5 collide with f3, so only f1 is then selected next. The resulting combination (f1,f3) is correct.

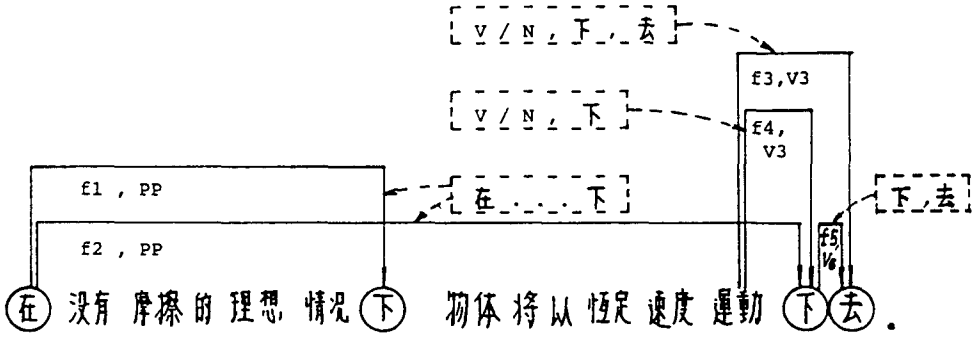
Fig.3 shows the parsing result obtained by computer in our preprocessing subsystem.

4. PRIORITY

In the preprocessing, we determine all the possible fragments that might occur in the sentence and involving the characteristic words. Then we give each one a measure of priority. This measure is a complex function, determined largely by trial and error. It is calculated by the following principles:

a) Kind of fragment

Some kinds of fragments, for example, compound verbs involving "完", occur more often than others and are accordingly given higher priority

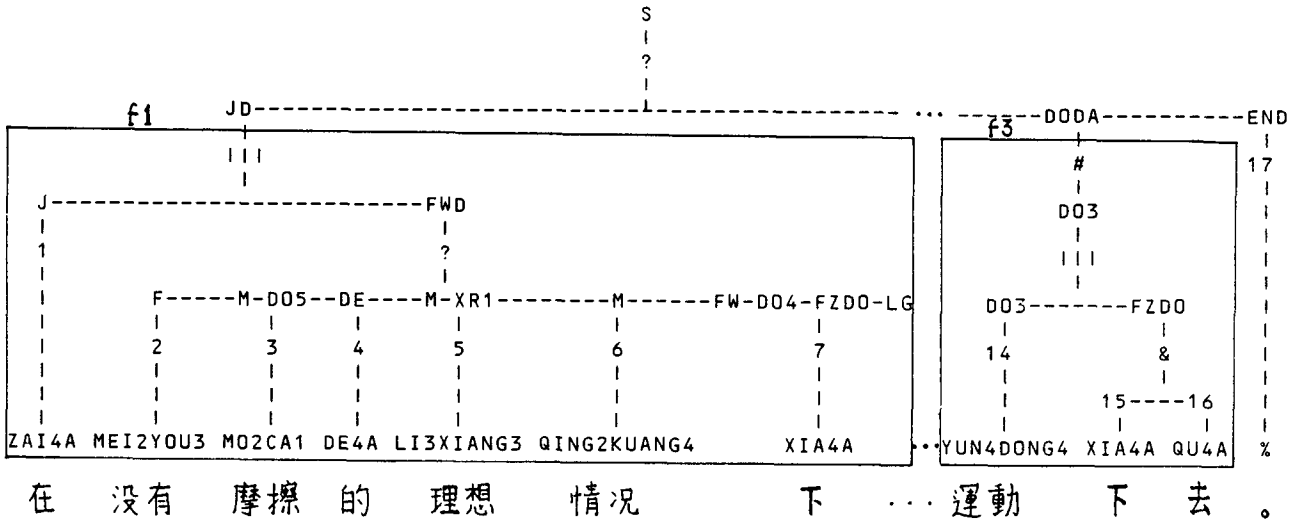


Translation : In the perfect situation without friction the object will keep moving with constant speed.

: pattern of fragment

V / N : a word which is either a verb or a noun (undetermined at this stage)

Fig.2 An Example of Conflicting Fragments



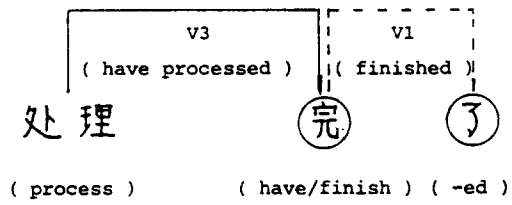
Translation : In the perfect situation without friction the object will keep moving with constant speed.

: fragment obtained by preprocessing subsystem

f1 , f3 : the names of fragments shown in Fig.2

. . . : the omitted part of the resultant structure tree

Fig.3 An Example of The Analysing Result Obtained by The Preprocessing Subsystem



Translation : had processed

: fragment given the higher priority

: fragment given the lower priority

Fig.4 An Example of Fragment Priority

(Fig.4). We distinguish 26 kinds of fragments.

b) Preciseness

We call "precise" a pattern that contains recognizable characteristic words or subpatterns, and imprecise a pattern that contains words we cannot recognize at this stage. For example, f3 of Fig.2 is more precise than f1, f2 or f4. We put the more precise patterns on a higher priority level.

c) Fragment length

Length is a useful parameter, but its effect on priority depends on the kind of fragment. Accordingly, a longer fragment gets higher priority in some cases, lower priority in other cases.

The actual rules are rather complex to state explicitly. At present we use 7 levels of priority.

5. PREPROCESSING EFFICIENCY

The preprocessing system for chinese language mentioned in the paper is in the course of development and it is partly completed. The inputs are sentences separated into words (not consecutive sequences of characters). We use 200 characteristic words and have written the rules by 101 automata for them. As a preliminary evaluation, we tested the system (partly by hand) against 120 sentences taken from a Chinese physics text book. From these 369 fragments were obtained, of which 122 were in conflict. The result of preprocessing was correct at first choice (no back-tracking) in 94% of sentences. Allowing one back-tracking yeilded 98%, two back-trackings gave 100% correctness.

In this limited set, few conflicting prepositional phrases appeared. To test the performance of our preprocessing in this case we tried the method on a set of more complex sentences. From the same textbook, out of 800 sentences containing prepositional phrases, 80 contained conflicts, involving 209 phrases. Of these conflicts, in our test 83% were resolved at first choice, 90% at second choice, 98% at third choice.

6. SUMMARY

In this paper, we outlined a preprocessing technique for Chinese language analysis.

Heuristic knowledge rules involving a limited set of characteristic words are used to forecast partial syntactic structure of sentences before global analysis, thus restricting the path through the search space in syntactic analysis. Comparative processing using knowledge about priority is introduced to resolve fragment conflict, and so we can obtain the correct result as early as possible.

In conclusion, we expect this scheme to be useful for efficient analysis of a language such as Chinese that contains a lot of syntactic ambiguities.

ACKNOWLEDGMENTS

We wish to thank the members of our laboratory for their help and fruitful discussions, and Dr. Alain de Cheveigne for help with the English.

REFERENCE

[1]. Yiming Yang:
A Study of a System for Analyzing Chinese Sentence, masters dissertation, (1982)

[2]. Shuxiang Lu:
"現代漢語八百詞", (800 Mandarin Chinese Words), Bejing, (1980)