PARTIAL CONSTRAINTS IN CHINESE ANALYSIS

Yiming YANG, Shuji DOSHITA and Toyoaki NISHIDA

Department of Information Science
Kyoto University
Sakyo-ku, Kyoto 606, Japan

ABSTRACT

In this paper, we describe a method using semantic constraints to reduce the ambiguities and generate case structure from phrase structure in Chinese sentence analysis.

Semantic constraints written on semantic markers indicate the plausible case structure. Different sets of semantic markers are chosen according to the purpose. A priority evaluation scheme steers the analysis towards the most plausible structure first, without trying all possibilities.

1. INTRODUCTION

Chinese is written with characters that don't admit formal inflections to indicate the grammatical categories of words. Also, there are few functional words, so few cues are available to indicate the grammatical structure.

Automatic parsing of Chinese runs immediately into an explosive growth of possible structures due to ambiguity, so syntactic and semantic constraints must be introduced as soon as possible in the analysis to restrict the search.

Knowledge that can be used for this purpose is mostly of a partial nature that leads to "plausible" interpretations. As such, it is difficult to manage, because several possibilities occur at each step.

In a previous paper (Yang et al. 1984), we show how to use the knowledge associated with "characteristic words" in a preprocessor designed to precede syntactical analysis.

Here we describe a system that uses the semantic categories of words to obtain case structure.

2. CASE STRUCTURE ANALYSIS

In Chinese, we cannot derive the phrase structure from the syntactic categories of the words and phrases, because cues such as conjunctions and inflexions are for most part lacking.

Fig 1 shows three examples of Chinese sentences:

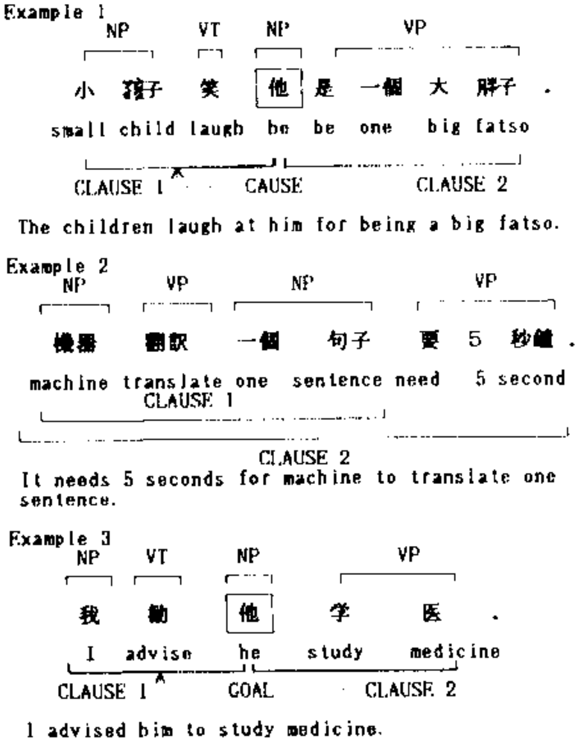
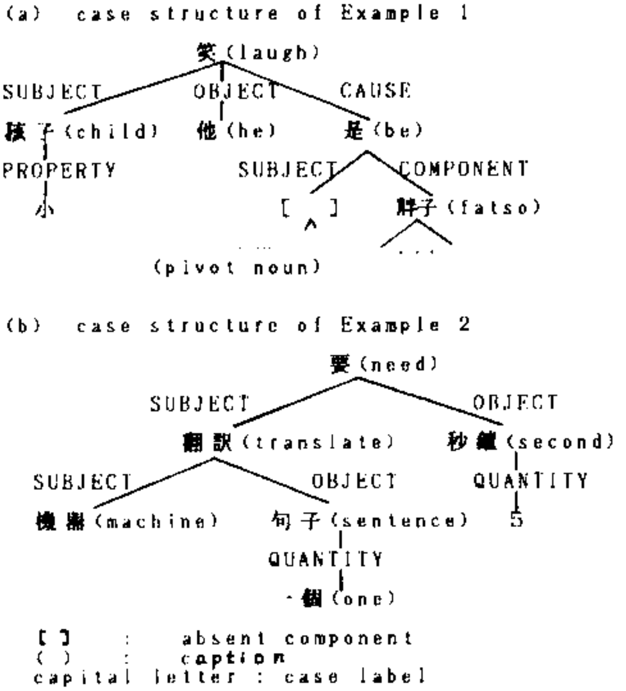

Example 1
NP
VT
NP
VP
小孩子 笑 他 是 一個 大 胖子 .
small child laugh he be one big fatso
CLAUSE 1
CAUSE
CLAUSE 2
The children laugh at him for being a big fatso.
Example 2
NP
VP
NP
VP
機器 翻訳 一個 句子 要 5 秒鐘 .
machine translate one sentence need 5 second
CLAUSE 1
CLAUSE 2
It needs 5 seconds for machine to translate one sentence.
Example 3
NP
VT
NP
VP
我 勸 他 学 医 .
I advise he study medicine
CLAUSE 1
GOAL
CLAUSE 2
I advised him to study medicine.

Figure 1

All three examples show the same sequence NP-VT-NP-VP. The NP, VP, etc., are the partial results of a phrase structure analysis stage, but the analysis cannot proceed from here on. All the examples consist of two clauses, but the relationship between them is different in each case. In Example 2, the clause is embedded, whereas Examples 1 and 3 have both a "pivotal structure" (Li and Thompson 1981), that is, the second NP is both the object of the preceding clause and the subject of the following clause. These differences only emerge if one-considers the meaning of the words.

In our system we try to apply semantic knowledge to lift this kind of ambiguity. We do not attempt to provide at some step a complete syntactic phrase structure. Instead, we use a case stucture to represent the result of semantic analysis. We chose a set of case labels identical to that of the Mu-machine translation project (Nagao et al. 1983, 1984), for compatibility and to enable comparison.

Fig 2 shows the case structures for the two of the previous sentences:



(a) case structure of Example 1

(b) case structure of Example 2

Figure 2

3. SEMANTIC CONSTRAINTS FOR CASE STRUCTURE

We determine the case structure from the partial syntactic structure using semantic markers. The semantic markers (for nouns, verbs, etc.) used for writing the constraints are organized in hierarchies of semantic categories. Multiple hierarchies are used, according to the purpose (the particular ambiguity to be lifted) and the subject domain of the text.

Consider the examples above. We choose the case structure according to the semantic category of the verbs. Seven types of case structure are defined for the NP-VT-NP-VP sequence (including the two types shown in figure2). Likewise, verbs are classified into seven groups named LET, EMOTION, HAVE, KNOW, NEED, TEACH and OTHER, each of which corresponds to one type of structure. Note however that one verb can belong to several groups.

The rules are like the following:

"if the preceding verb belongs to the EMOTION group, then use the "pivot construction" (as in Example 1), with a CAUSE relation;
if the posterior verb (in VP) belongs to the NEED group, then the preceding part, NP-VT-NP is a clause which is the subject of VP (as in Example 2); ..."

In some cases, the rules give contradictory conclusions. In Example 1, the verb "笑" belongs to the EMOTION group, but "是" belongs to the "NEED" group. We must find a way of managing such multiple possibilities and choosing among them. This is done by means of the priority calculus scheme that we describe further on.

Consider, as another example, the sentence "A 是 B". Like the sentence "A is B" in English, "A是B" is a very common sentence in Chinese, but it is difficult to decide its scope, that is, where A begins and where B finishes. As above, we lift the ambiguity by using the semantical properties of A and B, to choose from several possible phrase structures.

For this particular problem we use a semantic hierarchy containing about 60 semantic markers for nouns, borrowed from the Mu-project. However, we add cross relationships that link some categories that are not sub-categories one of the other. For example, in the physics text-books that we tested our system with, the need appeared for relationships linking "physical phenomenon" with "relation", "standard" with "unit" etc.. This kind of add-on relationship is domain-dependant.

In the case where either A or B is an embedded clause, a quite different classification is appropriate, grouping for example the words that would most frequently stand opposite an event like "原因" (reason), "問題"(problem), "狀態 " (state), "作用 " (function), etc..

Thus the semantical hierarchies are both domain and problem (ambiguity) dependant.

4. PRIORITY

As we mentioned above, our semantic constraints use knowledge which is

- not 100% correct (only statistically probable),
- not complete (described for parts of structure, or only written for typical situations).

This often results in some wrong decisions when a local analysis is done, so at some stages several possibilities must be retained. We use a priority scheme to evaluate the plausibility of each case structure and find the best choice.

In calculation:
a) Partial priority is calculated by constraint rules. It is determined experimentally.
b) Positive priority indicates a likely structure, negative priority an unlikely structure, 0 is indifferent.
c) Priority of the whole structure is the sum of partial priorities.

Consider the two plausible structures in Example 1, for example. Their scores are both given as +1 when the semantic constraints are checked with NP-VT-NP-VP. However, the score for "A is B" in the structure of Figure 1 is +1 because both A ("he") and B ("fatso") belong to the same group HUMAN. The score for the other interpetation, "It is a big fatso that the childrens laugh at him.", is -1 because A, an embeded clause, and HUMAN B ("fatso"), form an unlikely pair. The two partial scores above are added together, and the correct structure(in Figure 1) has the higher score.

5. OTHER FEATURES

For describing the knowledge used in semantic analysis, a complete set of rules comprising both general and word specific rules would be clumsy to write, use and modify. Instead, we use an object-oriented scheme to separate the rules into independent modules (objects) according to their operational properties, and organize them into layers of classes. More then one parent, or a set of parents are allowed. The object-oriented interpreter supports a nondeterministic search mechanism for the multiple heritage.

A priority-driven parser is designed to make the searching efficient. The parser does a phrase structure analysis following context-free grammar rules, in a bottom-up way. It evokes an object-oriented interpreter, like calling a procedure, to generate case structure and calculate priority for each hypothesis it makes. Only the partial result with the highest priority is expanded. The others are saved, so the searching direction can be adjusted each time the priority is changed.

This system is in the course of being developed and is only partially completed. As a preliminary evaluation, we tested the system against 20 typical sentences selected from grammar books and science and technology books in Chinese (Li and Thompson 1981, Lu 1980, Zhu 1982). The correct case structures are obtained as first choice 70% of the time, and as first or second 100% of time. We also did a hand simulation with 160 sentences from a physics book in Chinese, resulting in 70% success rate in the first choice.

6. SUMMARY

In this paper, we described a technique for semantical analysis in our Chinese analysis system.

Rules of partial semantic constraints built on a limited set of concepts are used to reduce the ambiguities in case structure generation.

The priority scheme gives us a way to write incomplete knowledge into our rules. The priority-driven parser guides the global analysis through the search space heuristically, so a combinatorial explosion of computation can be avoided.

The object-oriented scheme makes it easy to modularize, access and modify different kinds of knowledge.

In conclusion, we hope this method to be useful for natural language processing, where very complex semantical information must be managed in an efficient way.

ACKNOWLEDGEMENTS

I wish to thank Associate Professor Jun-ichi Tsujii and Dr. Jun-ichi Nakamura of Kyoto University for their help and fruitful discussions, and Dr. Alain de Cheveigne for help with the English.

REFERENCES

[1] Yang, Y., Nishida, T., Doshita, S. (1984), "Use of Heuristic Knowledge in Chinese Language Analysis", COLING-84, 222-225.

[2] Nagao, M. (1983), "Summary of Machine Translation Project of Science & Technology Agency (of Japanese Government)" (科技庁機械翻訳プロジェクトの概要), technical report of WG on Natural Language Processing of IPSJ, 38-2 (in Japanese).

[3] NAKAMURA, J., TSUJII, J.,and NAGAO, M. (1984), "Grammar writing system (GRADE) of Mu-machine translation project and its characteristics", COLING-84, 338-343.

[4] Li, C., Thompson, S. (1981) "MANDARIN CHINESE --- A Functional Reference Grammar", University of California Press.

[5] Lu, S. (1980), "800 Mandar in Chinese Words" (現代漢語八百詞), Beijing (in Chinese).

[6] Zhu, D. (1982) "Lecture of grammar" (語法講義), Beijing (in Chinese).