A CHINESE NATURAL LANGUAGE PROCESSING SYSTEM BASED UPON THE THEORY OF EMPTY CATEGORIES

Long-Ji Lin*, James Huang**, K.J. Chen*** and Lin-Shan Lee*

*Dept. of Electrical Engineering, National Taiwan University, Taiwan, R.O.C.
**Dept. of Modern Languages and Linguistics, Cornell University, U.S.A.
***Institute of Information Science, Academia Sinica, Taiwan, R.O.C.

ABSTRACT

In this paper, we will present a device specially designed on the basis of the theory of empty categories. This device cooperates with a bottom-up parser and is used as an elegant and efficient approach to treat the troublesome problems of the transformations of passivization,relativization; topicalization, ba-transformation and the use of zero pronouns in Chinese natural language. With the aid of the device, the grammar rules for Chinese will be much more simplified and easier to design, and the processing capability can be significantly improved.

I INTRODUCTION

Passivization, relativization, topicalization, ba-transformation and the use of zero pronouns play major roles in Chinese. To deal with those syntactic phenomena, the conventional approach is to collect a set of grammar rules to cover all the possible sentence patterns derived from those transformations. But such an approach needs a great set of grammar rules to cover all the possibilities. Especially the complexity resulting from the interactions of several transformations will make such an approach infeasible.

Another approach adopted in this paper is the use of the raise-bind mechanism based upon the theory of empty categories. It seems that the above syntactic phenomena are not related to each other. However, the sentences derived from them all involve the common use of empty categories. With the use of the raise-bind mechanism, the parser will treat the transformations in the same way.

The following sections will briefly describe our parsing algorithm first, then discuss empty categories in Chinese and how the raise-bind mechanism operates.

II THE PARSING ALGORITHM

In the SASC system presented here, Chinese sentences are syntactically analyzed from the viewpoints of generative grammar (Huang, 1982). The SASC system uses a bottom-up parser instead of a top-down parser, because the former tends to be more efficient for Chinese sentence analysis. The parser uses charts (Kay, 1973; Kaplan, 1973) as global working structures, because many natural language processing systems, such as MIND (Kay, 1973) and GSP (Kaplan, 1973), have proved the chart to be an efficient data structure to record what have been done so far in the course of parsing. A parser based on charts can avoid the inefficiency in duplicating many computations that a top-down parser often suffers when backtracking occurs.

The input Chinese sentence is submitted to a preprocessor, which segments input sentence (a sequence of Chinese characters) into words. The result of the preprocessor is represented by a chart, and is sent to the parser. The parser parses sentences in the way that phrases are built up on the chart by starting with their heads and adjoining constituents on the left or the right of the heads. For example, according to the phrase structure rule (PSR), "NP—> QP N", N (noun) is the head of NP. When encountering a noun, the parser will try to build an NP by starting with the noun and adjoining the proceding quantity phrase (QP). According to the PSR, "VP—> V-n NP", V-n (transitive verb) is the head of VP. When encountering a transitive verb, the parser's action is similar to that of "NP —> QP N", except that it tries to adjoin the following NP as its object. But if its following NP is not yet parsed by the parser, the expectation to build a VP is suspended until an NP is built up in the object position.

The parser using the above algorithm constructs syntax trees of input sentences exactly from bottom to top. The algorithm used seems to be a good combination of datad-driven parsing and hypothesis-driven parsing. The implementation of the parsing algorithm and the grammar to model Chinese syntax can be found in (Lin et al., 1986).

III EMPTY CATEGORIES

Let's consider the following Chinese sentences.

(1) 他 打傷了 張三
he hurt Chang-san

(2) ba-transformation:

他 把 張三 打傷了 e

he ba Chang-san hurt
(He hurt Chang-san)

(3) passivization:

張三 被 他 打傷了 e

Chang-san by him hurt
(Chang-san was hurt by him)

(4) topicalization:

那隻狗 我 沒 看過 e

that dog I never have seen
(I have never seen that dog)

(5) relativization:

e 玩要 的 小孩

playing de children
(the children who were playing)

(6) 張 三 設法 〔s e 逃走〕

Chang-san tried escape
(Chang-san tried to escape)

(7) pivot construction:

他 叫 小孩 〔s e 吃 飯〕

he asked children go to dinner
(He asked the children to go to dinner)

(8) using zero pronoun:

張 三 喜歡 e

Chang-san likes
(Chang-san likes someone or something)

Sentence (2)-(8) all involve a missing subject or object (indicated by "e"). But what does each missing subject or object refer to? The solid lines under sentence (2)-(7) indicate the reference of each one. The missing object in (8), however, does not refer to any element within (8). In fact, it is an omitted pronoun, which refers to someone or something understood in the situation.

According to the current linguistic theory (Chomsky, 1981; Huang, 1982), sentence (2) is derived from sentence (1) by a transformation called "move α ". The transformation is performed as follows: the object, "Chang-san" in (1), is moved by carrier " 把 " ("ba") to the position indicated in (2), and then leaves behind a "trace" (indicated by "e"). The trace dominates no lexcial material, but is "bound" to its antecedant,"Chang-san". In addition to ba-transformation, passivization, topicalization and relativization can also be analyzed as involving some form of "move α ". Thus there are traces within these constructions.

Sentence (6)-(8) also contain vacant NP-positions, which are not traces, because they are not derived from "move α ". They are called "null pronominals". Null pronominals are in general free, for example, sentence (8). But those in certain constructions are bound, for example, sentence (6) and (7). Sentence (7) is called a pivot construction; that is, the object of the first verb is also the subject of the second verb. So, the null pronominal in the subject position is "bound" to the object.

Traces and null pronominals are known as empty categories (or empty NPs). The syntactic behavior of null pronominals is different from that of traces. They, however, are treated indiscriminately in our implementation.

IV THE RAISE-BIND MECHANISM

The raise-bind mechanism is used to cope with empty categories; in other words, to find out the antecedant for each empty category except those which are free (eq. sentence (8)). With the aid of the raise-bind mechanism, the parser will generate an empty NP inserted into the vacant position where an NP is expected to appear. Then the empty NP will be raised up in some way along the parsing tree, when the tree is growing up (recall that the parser works bottom-up), until its antecedant is parsed. At this point, the parser binds the empty NP by setting it to refer to its antecedant. Once being bound, the empty NP will not be raised any further — this is because an empty NP has exactly one antecedant and cannot be bound more than one time.

Not every NP position can be filled by an empty category. In Chinese, empty categories only appear in the subject position and direct object position, and never in the indirect object position, and never in the indirect object position and prepositional object position.

In our implementation, an empty NP contains three fields: (1) a field to keep the pointer to its antecedant, (2) a field to keep where it came from, and (3) a field to keep the syntactic or semantic constraints on the empty NP for later checking.

We can formulate the rules informally to treat relativization as follows: for a noun and a relative clause to be combined into an NP, the relative clause must contain an empty NP which is unbound and marked coming from either subject position or object position, and the empty NP will be bound to the (head) noun.

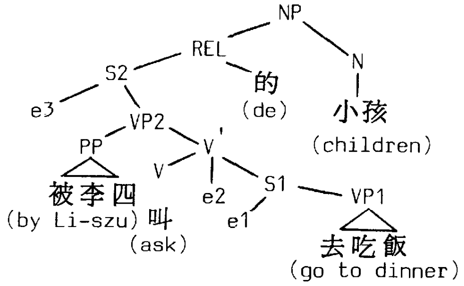
We can also state the rules for passivization as follows: once a clause is constructed, the parser checks whether the prepositional phrase, "被 +NP" (similar to "by+NP" in English) is involved in the clause. If so, there must be an empty NP which is unbound and marked coming from the object position, and it will be bound to the subject of the clause.

Rules for pivot constructions can be formulated as follows: in a pivot construction, the direct object will bind the empty NP coming from the subject position of the embedded clause.

Similarly, rules for topicalization, ba-transofmration and others can be designed. To illustrate the above rules, let's consider example (9) and its parsing tree in figure 1.

(9) 被 李四 叫 去吃飯 的 小孩

by Li-szu ask go to dinner de children
(the children who were asked by Li-szu to go to dinner)



e1 ⟶ e2

e2 ⟶ e3

e3 ⟶ 小孩 (children)

Figure 1. The parsing tree of (9)

Let's follow the bottom-up parser to parse example (9): (1) Node S1 is constructed and e1 serves as the dummy subject. (2) Node V' is constructed. V' is a pivot construction, so e1 is bound to e2. (3) Node S2 is constructed. S2 is a passive clause, because of the PP, "by Li-szu". According to the rules for passivization, e3 binds e2. (4) Node NP is constructed. According to the rules for relativization, e3 is bound to "children". Notice that only e3 was raised up across node S2, because e1 and e2 had been bound before S2 was constructed.

Once the parsing tree in figure 1 is finished, it is easy to answer who were asked and who went to dinner. Since e1 is the dummy subject of "go to dinner" and the binder of e1 is e2, whose binder is e3, whose binder is "children", we can conclude it is "children" who went to dinner. In the same way, we also conclude it is "children" who were asked.

The raise-bind mechanism also serves as a filter to rule out incorrect sentences or incorrect parsing trees. For example, if no empty NP is raised within a construction involving passivization or relativization, such a construction will be ruled out. If the mechanism is adopted for English sentence analysis, a test must be performed to rule out sentences with one or more empty categories which have no binder. But such sentences are in general grammatical in Chinese (see (8)).

V MORE SYNTACTIC PHENOMENA

Relativization in Chinese is a long-distance movement; that is, it can move an object across several S (sentence) ndoes. Noun phrase (10) is an example.

(10)[$_s$我 叫 李四 [$_s$ 幫 我[$_s$買 e]]] 的 書

I ask Li-szu help me buy de book
(the book which I asked Li-szu to buy for me)

(11) [$_s$ e2 喜歡 e1]的 人

like de the man

Noun phrase (11) is ambiguous. If the head noun ("the man") binds e1, the NP means "the man whom someone likes". If the head noun binds e2, it means "the man who likes someone or something". To remove the ambiguity needs semantic interactions.

Now we can formulate the rules for relativization as follows: for a noun and a relative clause to be combined into an NP, the parser checks the "empty-NP list" raised from the relative clause. And
"if no empty NP is raised, rule out the NP;
if an empty NP is raised and marked coming from subject position or object position or embedded object position (as in (10)), set the empty NP to be bound to the head noun;
if two empty NPs are raised from subject and object position (as in (11)), employ semantic analysis to determine the proper binding."

Like relativization, topicalization is also a long-distance movement and is treated in a similar way.

Another syntactic phenomena crucial to the parser is known as the Complex NP Constraint (CNPC) (Radford, 1981):
CNPC -- No transformation rule can move any element out of a complex NP.
A complex NP (CNP) is an NP containing a relative clause.

The CNPC can be easily encoded in our grammar in this way--all empty NPs can not be raised up across an NP node. Hence it is impossible for the empty NP within a CNP to be bound to any element out of that CNP.

In most cases, ba-transformation and passivization will move the direct objects of verbs. But the phenomena known as "subject-to-object raising" (Radford, 1981) makes some differences:

--The subject of an embedded clause can be moved into the subject (or ba-object) position of the higher clause by passivization (or ba-transformation).

For example, sentence (13) is derived from sentence (12) by such a movement.

(12) 大家 將來會 認爲 這個 錯誤 是 對的

people will believe this mistake is right

(13) 這個錯誤 將來會被 大家認爲 e 是對的

(This mistake will be believed to be right)

To cope with subject-to-object raising, the rules in previous section for passivization are modified as follows: the subject of a passive clause will bind the empty NP in either the object position or the subject position of an embedded clause.

VI A COMPARISON WITH THE HOLD-LIST MECHANISM

In ATN (Bates, 1978), the hold-list mechanism is used for the purpose similar to that of the raise-bind mechanism. But we object to such an approch, for (1) it is not fit for a bottom-up parser; (2) it cannot deal with null pronominals (eg. example (6)-(8)); (3) it handles left extraposition (eg. example (2)-(4)), not right extraposition (eg. example (5)). An movement is called left (right) extraposition, if it moves an NP to the position left (right) to its trace. To deal with right extraposition, ATN uses another mechanism.

In linguistic theory, bcth left extraposition and right extraposition move an NP to a position dominating its trace, and a null pronominal, if bound, is always bound to an NP dominating the null pronominal (Chomsky, 1981). So, the raise-bind mechanism is sufficient to cope with all empty categories, since its function is to raise up an empty category to be bouend to an NP which dominates this empty category.

VII CONCLUSION

We have presented how the raise-bind mechanism copes with traces and null pronominals in Chinese. With the use of the mechanism, many sophisticated syntactic phenomena can be encoded in the grammar easily.

The mechanism is simple and theoretically complete. If semantic analysis is employed to remove ambiguities, such as example (11), the correct bindings of empty categories can always be reached.

ACKNOWLEDGEMENTS

Thanks to the enlightening discussions of Chen, J.J. and Chen, J.C.

REFERENCES

[1] Bates, M. (1978) "The Theory and Practice of Augmented Transition Network Grammars",Natural Language Communication with Computers, pp.191-259.

[2] Chomsky, N. (1981) Lectures on Government and Binding, Forise, Dordrecht.

[3] Huang J. (1982) Logical Relations in Chinese and the Theory of Grammar, MIT doctoral dissertation.

[4] Kaplan, R.M. (1973) "A General Syntactic Processor", in [Rustin 1973].

[5] Kay, M. (1973) The MIND System, in [Rustin 1973].

[6] L.J. Lin, K.J. Chen, James Huang and L.S. Lee (1986) "SASC: A Syntactic Analysis System for Chinese Sentences", International Journal of Computer Processing of Chinese and Oriental Languages, Published by Chinese Language Computer Society.

[7] Radford, A. (1981) Transformational Syntax: A Student's Guide to Chomsky's Extended Standard Theory, Cambridge Univ. Press, 1981.

[8] Rustin R, ed. (1973) Natural Language Processing, Algorithm Press, N.Y.