Combining Prediction, Syntactic Analysis and Semantic Analysis in Chinese Sentence Analysis

Yiming Yang

Computing Research Laboratory, New Mexico State University
Box 30001, Las Cruces, NM 88003_0001, USA

ABSTRACT

This paper describes a system for the analysis of Chinese language that incorporates several heuristic techniques for reducing ambiguities.

Heuristic knowledge about "characteristic" words is used to predict a partial syntactic structure of sentence before doing global analysis. More than one prediction may exist, the system chooses the best combination of compatible partial predictions. Backtracking is evoked if acontradiction is later detected, producing a different choice of predictions.

Semantic constraints are used for translating phrase structure into case structure Rules are written to independent modules for word classes. An object-oriented scheme organizes the rules into layers, according to their priorities of application.

A preference score is calculated at each step of processing, giving a synthetic evaluation of both syntactic plausibility and semantic plausibility. The scores of partial results steer a priority-driven parser towards the most plausible sentence structure, instead of generating all the possible results and afterwards making the choice.

An experimental system based on these techniques has been built and tested on over one hundred sentences selected from published material. The results were very successful. The number of trees required to be produced in order to obtain a correct analysis was typically reduced from several hundred to under ten.

1 INTRODUCTION

The studies concerned with automatic analysis of Chinese language have been done mostly in the 1980's. A number of systems have been built. For example, Fan and Xu [1981] used an ATN grammar in a prototype Chinese understanding system. Feng [1983] experimented with translation of Chinese text on the GETA system. Huang (1986] built a system that performed a bidirectional translation between English and Chinese. In these studies, however, the emphasis was not on resolving ambiguities.

Phrase structure analysis of Chinese runs immediately into an explosive growth of possible structures because there are very few morphological variations of word roots to indicate category or sentence structure. We must find a way to introduce both syntactic and semantic constraints as early as possible in order to restrict the growth of ambiguities.

The Chinese analysis system presented here is characterized by two heuristic approaches to disambiguation: predicting syntactic structure from the presence of characteristic words, and evaluating the plausibility of possible results by semantic preferences on case structures The basic principles *of* these two approaches were presented in two papers by the current author [Yang, 1981, 1985]. The emphasis here is on incorporating these approaches into an effective analysis system.

2 PROCESSING FLOW AND CONTROL METHOD

This system consists of four mapr components which perform segmentation, preprocessing, phrase structure analysis, and case structure analysis These are illustrated in figure 1 and explained below

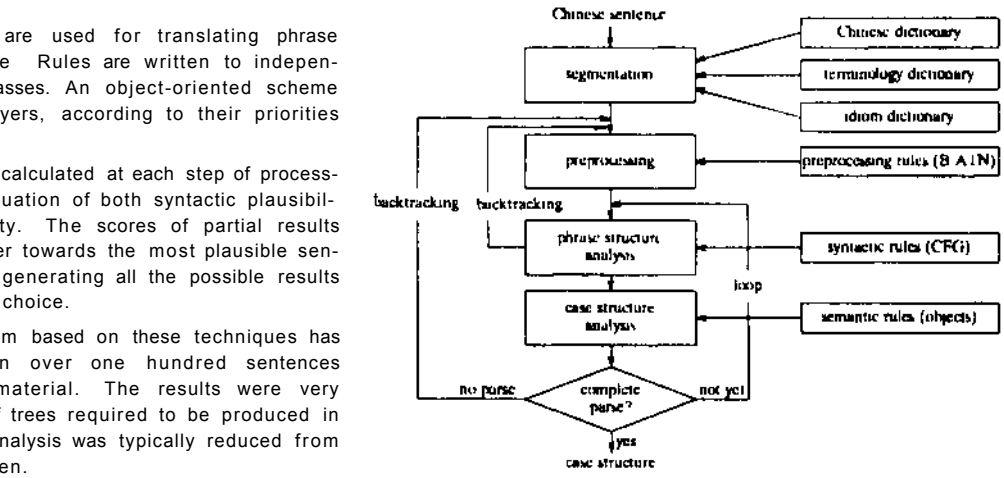


Figure 1 The basic configuration of the system

The segmentation component translates the input string of characters to a string of words (This is not trivial because it is customary to write Chinese words without spaces between them, but this is not the focus of this paper.)

The preprocessing makes a partial prediction of the phrase structure according to the appearance of characteristic words in the sentence. More than one prediction (of partial structures) may be obtained, a preference scheme is used to choose the best combination of compatible predictions,

The phrase structure is built by a context free (CF) parser, following the prediction from preprocessing. The parser first does local analysis on the parts where partial structures are predicted to be exist, then composes the partial results into a global structures. In case there is no structure is found as predicted, backtracking (to preprocessing) occurs for the next prediction. Since the context free analysis is restricted on the partial structures, the combinatorial growth of possible results is effectively reduced.

The case structure is derived from the phrase structure, according to the semantic constraints associated with words. The relationship between words or component structures is identified by case labels on the case structure, and the semantic consistency between the components is measured by preference scores.

The phrase structure analysis and the case structure analysis work in an interactive way. The phrase structure is built bottom-up. The semantic analysis is evoked each time when a partial syntactic structure (a phrase) is obtained, giving the corresponding case structure. The partial result in such an approach is a combination of a phrase structure (partial), its case structure, and a preference score which is a synthetic evaluation of both the syntactic and the semantic plausibility. For each cycle of the syntactic and semantic analysis in the bottom-up process, the parser looks at the set of partial results, chooses the one with highest preference score, and expands further analysis from this piece. The other partial results are saved, so the searching direction can be adjusted each time the preference of partial results is changed during the analysis. The above process continues until a complete parse is found. The priority-driven parsing arrives directly more plausible structures first, without trying all the possibilities and then comparing them

3. USAGE OF HEURISTIC KNOWLEDGE

1) Handling Prediction Rules in Preprocessing

The characteristic words used in preprocessing are a subset (about 200 words) of Chinese functional words, such as prepositions, locative particles, auxiliary words, modifier verbs, etc. They are frequently used and give hints of the sentence structure when they appear, but their use is often optional

Knowledge about the appearance of characteristic words is written to prediction rules of partial syntactic structures There are sets of rules for groups of characteristic words with similar functions. Approximately 40 ATNs were writr ten, each expressing the rules for a group of related characteristic words.

The preprocessing is done in the following steps.

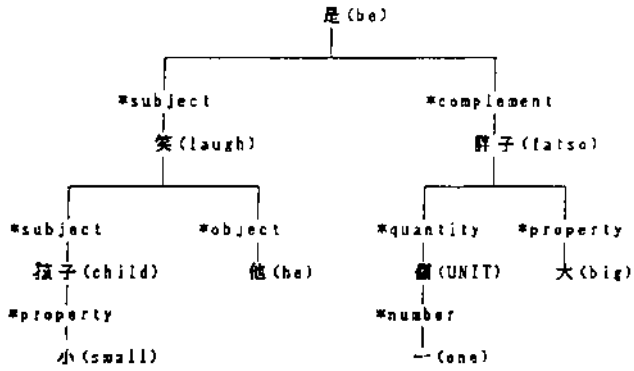
- extract fragments (partial predictions) by applying the rules of the characteristic words which appear in the sentence, and calculate the plausibility score for each fragment, according to its type, length, or the number of the recognizable characteristic words it contains,
- detect the conflict between fragments by checking their types and locations (to see if they overlap),
- construct the "most likely" combination of compatible fragments, according to the scores of fragments,
- backtrack for the next combination when a prediction is rejected in the subsequent analysis.

2) Semantic Processing

Semantic processing is used to analyze the roles of component phrases, and represent the relationship on a case structure. More than one case structure may be derived from a phrase structure. The semantic analysis gives each of the case structures a preference score by checking the semantic consistency between components.

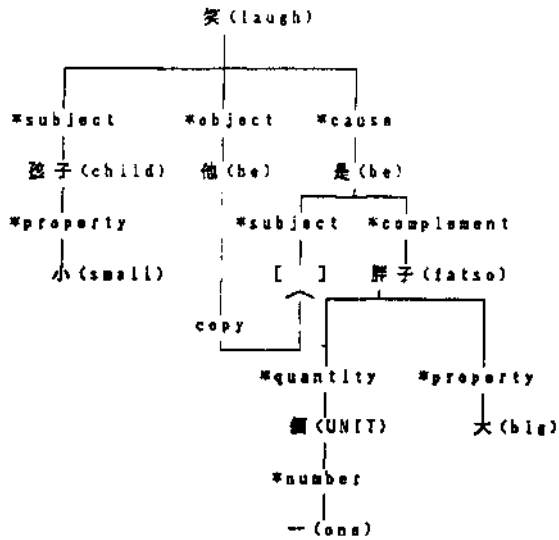
Consider, as an example, a sentence "小孩子笑他是一個大胖子。" ("small child laugh he be one big fatso") which means that "Small children laugh at him *for being* a big fatso." This sentence shows a sequence NP1-V-NP2-V-NP3 which is a compound sentence pattern called "Serial Verb Construction" (SVC) in Chinese grammar. Syntactic analysis can not recognize which verb in the SVC string is the predicate of the sentence, the "laugh" or the "be"; or where the embedded clause should be located on, the NP1-V-NP2, the NP2-V-NP3, or the V-NP3? That is because there is no morphological variation, like the *"for"* or the "-tny" in English, to indicate the structure of such a sentence. This ambiguity can be resolved only when the meaning of words is considered

This system does not distinguish SVC sequences by phrase structures, but maps them to different case structures (seven types) according to the semantic consistency Figure 2 shows two possible case structures for the above sentence



translation

The fact that small children laugh at him is a big fatso



translation: Small children laugh at him for being a big fatso.

"" : case label, [ ] : absent component*

Figure 2 The two possible case structures for a Chinese sentence

In calculating the semantic plausibility of the two case structures, constraints between verbs and their subject, object and/or the subordinate clause are checked. In the first structure, the predicate "SHI4" (be) has a embedded clause ("small children laugh at him") as its subject. This structure is semantic unlikely because the subject, a clause, does not match the complement "fatso", a HUMAN. The preference for this matching is a negative score of "-1". In the second structure, the main clause "small children laugh at him" is modified by subordinate clause "he is a big fatso" which *causes* the previous event to happen. The semantic constraint for such a structure is that the predicate of the caused event should be an emotional verb, "laugh" is such a verb, and so a positive score of "4-1" is added for this evaluation. On the other hand, the semantic checking for the subordinate clause brings another "-hi" because both the subject "he" and the complement "fatso" belong to the semantic category HUMAN. The partial scores are added together and the correct structure (the second one) has higher score

Semantic constraints used are written into separate modules of rules for different word groups. Different sets of semantic groups are chosen according to the linguistic phenomena where the ambiguities occur. Seven verb classes, for example, are used for analyzing SVC sentences. They are named CAUSATIVE, EMOTIONAL, POSSESSIVE, NARRATIVE, SPEC1AL-1, SPECIAL-II and NORMAL, and each of them corresponds to one type of case structures About sixty classes are used for nouns.

The modules of semantic rules are organized into a hierarchy according to their operational properties. In semantic processing, an objectroriented scheme applies these modules, called objects, by searching along the hierarchy. Word specific rules are tried first if they are given; otherwise, their parent objects are applied, and so on.

Object-oriented systems usually work by sending a message to an object which returns one message as its result. In a natural language analysis system, we need to deal with multiple possibilities during parsing. This system supports a mechanism to send a message to more than one object and then return all of the possible results of objects. Also, more than one parent object is allowed for an object, for nondeterministic search

4. EXPERIMENTAL RESULTS

A experimental system (written in Lisp) was built to test the above methods. The system contained an 800 word dictionary for the syntactic analysis, 37 ATNs, for preprocessing, about 210 CFG rules of the phrase structure grammar, and about 1,100 "objects" (including word specific objects) for the semantic analysis. (The priority-driven mechanism is designed but not coded in yet, because it is not absolutely necessary in the stage of testing the main methods, and there should be no technical difficulty in implementing such a search "when necessary.)

Over one hundred sentences taken from a Chinese physics textbook, scientific papers, grammar books, etc., were chosen for testing the components of the system. The preprocessing gave the correct result as its first choice 94% of the time. The syntactic analysis subsystem was separately tested with the same set of sentences (the preprocessing was not used to restrict the syntactic possibilities), and the correct phrase structure was obtained for 83% of the time. A subset of these sentences was used to test the whole system, the correct result was the first choice for 14 out of 20 sentences and the second choice for the remaining 6.

A typical example shows the parsing efficiency of this system. In analyzing the following sentence:
"在没有摩擦的理想情况下物体将以恒定的速度运动下去。"
(*"In the perfect situation with out friction the object will keep moving at a constant speed.'),* preprocessing reduced the syntactic ambiguities from 552 possibilities (trees) into 8. These 8 syntactic structures were interpreted into 18 case structures. Using their preference scores, the correct result was obtained as the first choice.

5. SUMMARY

This paper described the technique for buiding a Chinese analysis system which combines segmentation, preprocessing, phrase structure analysis and case structure analysis.

This system applies two heuristic approaches for ambiguity resolution: predicting partial syntactic structure by characteristic words and introducing semantic constraints in translating the phrase structure to the case structure. Knowledge needed for this analysis is formalized in independent modules for each approach and managed as an integrated process. Interactions between the preprocessing, the syntactic analysis and the semantic analysis, make it easy to introduce constraints as early as possible, and thus efficiently reduce the growth of ambiguities The preference scheme makes it possible to use incomplete knowledge in partial analyses, and make a synthetic evaluation for the global structure by summing the partial scores.

In conclusion, these methods should be useful for natural language analysis, where ambiguity raises a serious problem, and complex linguistic information must be managed efficiently.

ACKNOWLEDGEMENTS

I would like to thank the members of Doshita's Laboratory at Kyoto University and the Computing Research Laboratory at New Mexico State University for their help and fruitful discussions.

REFERENCES

1. Yang, Y., Nishida, T. and Doshita, S. Use of Heuristic Knowledge in Chinese Language Analysis, COLING 84.

2. Yang, Y., Doshita, S. & Nishida, T. (1985), Partial Constraints in Chinese Analysis, IJCAI 85.

3. Feng, Z. Multiple-labeled and Multiple-branched Tree Graph Analysis Method of Chinese Sentence, in Proceedings of 1983 International Conference of Chinese Information Processing (ICCIP), Beijing, 1983.

4. Fan, J. and Xu, Z. The Application of Augmented Transition Network Theory to Chinese Understanding_an experimental model of man-machine dialogue by Chinese language, Language Study, Vol. 1, China, 1981.

5. Huang, X. A Bidirectional Chinese Grammar in A Machine Translation System, New Mexico State University CRL memo, 1986.