日本技術

M.I.T.-JAPAN SCIENCE AND TECHNOLOGY PROGRAM

A CONTEXT FREE RULE SYSTEM FOR PARSING JAPANESE

Carol L Tenny

Department of Linguistics and Philosophy and
Athena Language Learning Project,
Massachusetts Institute of Technology

A CONTEXT FREE RULE SYSTEM FOR PARSING JAPANESE

Carol L Tenny

Department of Linguistics and Philosophy and
Athena Language Learning Project,
Massachusetts Institute of Technology

1. Introduction

A project is underway at MIT to develop computer based language teaching materials that will combine highly advanced computer technology with modern language teaching methods. The Athena Language Project draws on MIT's considerable resources in computation, linguistics and artificial intelligence to produce language lab materials that will incorporate interactive video and audio components as well as computer intelligence capable of understanding commands in natural language. The final product is envisioned as a set of language exercises that will allow the student to interact with the computer in a creative way, in the language he is learning by giving commands or responding to fictional situations posed by the computer. The computer in turn understands and responds to the student in the language of instruction. The languages included in this project are those currently taught at MIT German, Spanish, French, Russian, English as a Second language, and Japanese.

Japanese is fast becoming a language of international importance in technical and scientific exchange as well as in the business world. It has only recently begun to be taught at MIT. The popularity of the Japanese language offering at MIT has outstripped the ability of the foreign languages department to accomodate all students who want to study Japanese. In view of the popularity and importance of the language, it was essential that Japanese be included in the project To this end, the MIT Japan Science and Technology Program has funded the development of a Japanese component of the software. A rule system and lexicon for parsing Japanese were designed by the author. The Athena Language Project is translating the rule design into

LISP code for implementation within the program that translates natural languages into abstract representations of meaning.

The parsing rules are the component of the software that is tailored to the particular language The system used by the Athena project has, built into it, some knowledge of the world and a capacity for manipulating certain fictional situations. This part of the system remains the same for each language to which it is applied, but the system must have in addition, a separate set of rules for understanding input sentences in each language. These rules work together with a lexicon for the language, to translate a sentence or a string of words in the natural language into meaning structures which the system can then manipulate. This invariably means taking a string of words organized linearly, in one dimension, and reorganizing them into groupings with hierarchichal as well as linear relationships between the parts. The precise way in which this is done by human intelligence is the topic of some research and speculation, but several working models for this process are in use in work on artificial intelligence. The system employed by the Athena Language Project is the Malone parser, a rule-based parser with a lookahead of one, operating on a system of context-free rules. The set of rules designed for Japanese will not be listed in this document, but in the remainder of the document, the general coverage of the rules, as well as the approach taken towards particular problems, will be outlined.

In order to design this rule system, it was necessary to first establish a satisfactory idealization of Japanese grammar. This meant deciding on two things: 1) a version of the grammar that could compromise between the rigorous order required by the parser, and the natural fluidity expected by

the human user, and 2) how much and which part of the language the rules should be able to handle. The parsing rules cannot be expected to model everything a human being can do, because the parser responds only to the strict grammar or syntax of a sentence, while the human handling of language is affected by pragmatic, semantic and discoursal as well as syntactic conditions. Decisions had to be made about how much of the syntax the rules would cover, and what kinds of linguistic information they should be prepared to handle. Linguistic, computational and pedagogical considerations entered into this problem. Linguistically, the goal was to incorporate as many insights as possible from current linguistic theory into the structure of Japanese grammar. The more the system is based on sound and simple principles, the easier it will be to expand, alter or otherwise develop it later. Computationally, it was necessary to keep the rules as simple as possible, and avoid building into them extra machinery that might be linguistically accurate but would greatly inhibit the efficient working of the rules. And pedagogically, the rules should be designed to handle linguistic constructions that a fourth semester student of Japanese might reasonably be expected to produce.

2. Approaches to Problems Posed by Japanese Grammar

2.1 The Topic Marker "Wa"

A particular problem is posed by the topic marker "wa". There are no precedents for parsing this type of construction in the other, Indo-European languages included in the project. The use of the topic marker "wa" by human speakers is governed not only by syntactic principles, but by semantic and discourse factors as well, some of which are more subtle than a system such

as this one can capture. The following set of idealizations were employed for the rule system:[1]

i) There is only one topic marker per sentence.
ii) The topic marker is only used in matrix clauses.
ii) The topic marker may only occur at the beginning of the clause.

This means that the system of rules will accept the following sentences:

Hanako wa gohan o tabemashita
gohan wa Hanako ga tabemashita
Tokyo wa Taroo ga ikimashita
Taroo wa Tokyo ni ikimashita
Hanako wa Taroo ga Tokyo ni iku to iimashita
Tokyo wa fuyu ga samui
Taroo wa otosan ga shinda

The system will reject the following sentences. At this point it does not handle constructions with the contrastive use of "wa", which are somewhat freer than the straightforward topical use of "wa".[2]

gohan o Hanako wa tabemashita
Tokyo ni Taroo wa ikimashita
Taroo wa Tokyo ni iku to Hanako wa iimashita

2.2 Free Word Order

The relatively free word order of Japanese, compared to English, French,

---

1. Professor Michio Tsutsui of the MIT-Wellesley Exchange Program kindly advised on this problem. Any inadequacies however, are the author's. These idealizations are still being used on a trial basis. They may be changed as pedagogical or computational considerations require.

2. For more discussion on the use of the topic marker "wa" see Kuno, S. (1973) The Structure of the Japanese Language MIT Press, Cambridge, pp. 35-79.

Spanish or German, presents a major problem for a rule-based parser such as the Malone parser. Each rule must express a possible fixed word order -- of which there are a great many in Japanese. This means that coverage of even a small number of basic constructions requires a large number of rules, reducing the overall efficiency of the parser. This problem was in part circumvented by labeling several different kinds of constructions with one name, so that fewer rules would be required to address all the constructions. However, this solution has reduced efficiency elsewhere in the grammar because some extra machinery is required to distinguish the various constructions at a later stage.

The problem of free word order necessitates another compromise between linguistic and computational demands. Japanese is treated by these rules as a non-configurational language, in spite of the growing evidence to the contrary.[3] A non-configurational language is one in which no word order is more basic than any other, yet some facts about how pronouns are construed with noun phrases within a sentence point to the idea that the word order, Subject-Object-Verb is most basic in Japanese. To build this into the rule system would increase the number of rules dramatically but make the system only slightly closer to human ability. These facts about pronoun construal would not be likely to be within the reach of second year students, or perhaps even any non-native speakers of Japanese. It was judged that the

---

3. See Saito, M. (1985) "Some Asymmetries in Japanese and their Theoretical Implications", Ph.D. dissertation, MIT Department of Linguistics and Philosophy; and Whitman, J. (1982) "Configurationality Parameters", unpublished ms. Harvard University Department of Linguistics. A revised version is to appear in T. Imai and M. Saito, eds., Issues in Japanese Linguistics, Foris Press, Dordrecht.

gains in linguistic coverage that would result from a configurational treatment of Japanese would not be worth the added complexity it would generate in the rule system.

2.3 Noun Phrase Drop

The meaning structures into which the natural language is translated are essentially annotated case frames. That is, they are the simple listing of a predicate (e.g., a verb) and its arguments (e.g., noun phrases), and the relation between them. For example, the following sentence would be represented with the case frame below it.

Hanako ga Taroo ni tegami o okurimashita.
"Hanako sent a letter to Taroo."

predicate: okuru
  subject argument: Hanako
  object argument: tegami
  dative argument: Taroo

In Japanese the relation between predicates and arguments is expressed via case markers ("ga" for subject, "o" for object, etc.) rather than word order, as in English. Translating a Japanese sentence into its case frame is relatively straightforward as long as the roles of the case markers are unambiguous, and all the noun phrases are mentioned overtly in the sentence. However, Japanese allows noun phrases to be dropped from sentences, on a greater scale than in the other languages included in the project. The Athena system operates by building the case frame required by the particular verb and inserting arguments into the case frame as they are encountered in the sentence. Then the case frame is checked, and if some argument slots are empty, the sentence is rejected. This poses a particular problem for

Japanese, because in Japanese the arguments may not be overtly in the sentence, yet be understood by the speaker and hearer as being there. Incompletely filled case frames must be permitted in Japanese, and dummy arguments inserted into the empty argument slots. (The dummy arguments can then be assigned meaning in the discourse component of the grammar.) Some general code for the handling of case frames had to be rewritten to accomodate noun phrase drop in Japanese.

2.4 Morphology and the Lexicon

A fourth problem presented by Japanese is the rich morphology of the language. The morphological analyzer used by the Athena project is sophisticated, but not as sophisticated as Japanese morphology. Adjustments had to be made in the Athena system to allow morphology to change the case frame of a verb. For instance, the verb "nageru" ('throw') would have a case frame requiring two arguments -- the subject that throws, and the object that is thrown. But when the morpheme "sase" is applied to make the verb "nagesaseru" ('make someone throw'), the new case frame requires three arguments -- the thing thrown, the person throwing, and the person who causes the thrower to engage in the action of throwing. Morphological processes such as this are common in the languages of the world, and taking measures to streamline handling of the phenomenon will be beneficial to the system in the long run.

Japanese, as well as the other languages in the project, will be more effectively handled when a word-formation component is added to the Athena system -- a change planned for the summer of 1986. This will allow the rules to recognize morphological operations that change the category of a word, and

will reduce the size of the lexicon. An example of this is suffixation that changes a noun into an adjective. e.g.:

```
"otoko" + "rashii" --> "otokorashii"
 man        -like       like a man
```

The lexicon is the repository of the information about a language that is irregular and cannot be generated by rules. This is the information that must simply be listed. The Japanese lexicon includes the following:

words or word stems:
- -- their meanings
- -- their case frames, if they are predicates
- -- what endings may attach to them

word endings:
- -- how they change the meaning of what they attach to
- -- what words or word endings they may attach to
- -- what other word endings may attach to them

Some of the tenses that have been accomodated by word endings listed in the lexicon are illustrated below, with the verb "ik-", ('to go'):

iku	present-future informal
ikimasu	present-future formal
itta	past perfect informal
ikimashita	past perfect formal
ikanai	present-future informal negative
ikimasen	present-future formal negative
ikanakatta	past perfect informal negative
ikitai	present-future volitional
ikitakunai	present-future volitional negative
ikitakatta	past perfect volitional
ikitakunakatta	past perfect volitional negative
itte kudasai	present-future formal imperative
ikinasai	present-future informal imperative
ikanai de kudasai	present-future formal imperative negative
ikeba	provisional
ikanakereba	provisional negative
ittara	conditional
ikanakattara	conditional negative
itte	gerund
ikanai de	gerund negative

itte iru	present-future imperfect informal
itte imasu	present-future imperfect formal
itte ita	past-imperfect informal
itte imashita	past-imperfect formal
itte inai	present-future imperfect informal negative
itte inakatta	past-imperfect informal negative

The lexicon will continue to be expanded as more words and word endings are entered into it. It is expected that idioms will also eventually be included in the lexicons. The goal of the project is to enter the equivalent of a small dictionary into the lexicon of each language.

3. The Coverage of the Rule System

In order to illustrate the coverage of the system a few sentences are included below. These sentences and constructions are a sample of those that can be understood by the rule system. The symbol __ represents dummy arguments. The English glosses are listed for convenience, and do not reflect the way the meanings of the Japanese sentences are represented in the system.

tabemasu.
'__ eats __.'

Hanako ga tabemasu.
'Hanako eats __.'

pan o tabemasu.
'__ eats bread.'

Hanako ga pan o tabemasu.
'Hanako eats bread.'

Hanako ga Taroo ni hon o agemasu.
'Hanako gives a book to Taroo.'

Taroo ni Hanako ga hon o agemasu.
'Hanako gives a book to Taroo.'

Hanako ga Taroo ni tegami o yuubin de okurimashita.
'Hanako sent a letter to Taroo by mail.'

kesa soto e detara ame ga furimashita.
'This morning, when __ went outside it rained.'

ashita hayaku ikimasu.
'Tomorrow __ will go early.'

gakusei ga sannin kimasu.
'Three students will come.'

Osaka kara kimasu.
'__ will come from Osaka.'

Taroo ga Osaka ni itta to iimashita.
'__ said that Taroo went to Osaka.'

Hanako ga Taroo ni watashita hon o yomimashita.
'__ read the book that Hanako handed to Taroo.'

Hanako ga uchi e kaetta koto...
'The fact that Hanako went home...'

Taroo ga Tokyo ni itta toki, yuki ga futte imashita.
'When Taroo went to Tokyo, it was snowing.'

Taroo ga Tokyo ni ikeba, Asakusa ni ikimasu.
'If Taroo goes to Tokyo, __ will go to Asakusa.'

Taroo ga kita ka wakarimasu ka?
'Do you know whether Taroo came?'

Taroo ga doko ni ikimashita ka?
'Where did Taroo go?'

nani o tabemasu ka?
'What are/is __ eating?'

imoto wa ikutsu desu ka?
'How old is __ sister?'

naze kaerimasu ka?
'Why are/is __ going home?'

Hanako ga dare ni nani o agemasu ka?
'What will Hanako give to whom?'

Taroo wa gakusei desu.
'Taroo is a student.'

Kyoto no natsu ga attakai desu.
'The Kyoto summers are hot.'

hon desu.
'It's (a) book(s).'

ii desu.
'It's alright.'

yasui hon...
'(an) inexpensive book(s)...'

hon to zasshi...
'(a) book(s) and (a) magazine(s)...'

kuruma ya jitensha...
'automobile(s), bicycle(s) and so on...'

gakusei igai...
'other than students...'

futsuka inai...
'within two days...'

sanjyuu meetoru hodo...
'about thirty meters...'

binboo na gakusei to kanemochi-no roojin igai ni wa...
'as for those besides poor students and rich old folks...''

The completed system will have the ability to correct some errors made by students. In most of the languages error recovery will be applied to subject--verb agreement in number, gender and person, but in Japanese, error recovery will be performed for number--counter agreement. The rules require Chinese-root numbers to be followed by counters, and also check that the counters agree in class with their associated nouns. This will allow the system to recognize that the following are mistakes, and to correct the student:

*hon sannin
'books -- three people'

*otoko sansatsu
'men -- three volumes'

4. Conclusion

Some problems particular to writing parsing rules for Japanese have been discussed above. With the addition of the non-Indo-European language of Japanese to the Athena Language Project, the language processing systems used by the project have been stretched at times. The solution of problems posed by Japanese has resulted in a more flexible system with wider applicability to a greater range of languages. And finally, students of Japanese at MIT will be provided with the state-of-the-art language learning aids projected for students of the other languages traditionally taught at the Institute.

ACKNOWLEDGEMENTS

I am grateful to Professor Michio Tsutsui for his helpful comments and discussion on issues of Japanese grammar; to Janet Hirata for discussion about linguistic and computational issues regarding the rule system; and to Sue Felshin and Dr. Janet Murray of the Athena Project for their continued general support. In addition I would like to thank Dr. Richard Samuels and the MIT-Japan Science and Technology Program for funding this work.