[From: *Japan Annual Reviews in Electronics, Compters & Telecommunications* 1982]

# 2 . 1 A Survey of Natural Language Processing and Machine Translation in Japan

**Makoto NAGAO***

*Abstract*

In this survey brief historical review of natural language processing and machine translation in Japan is given first. This is followed by an explanation of a few typical research results in the morphological and syntactic analyses of the Japanese language. Machine translation has been getting the spotlight once again in Japan. Two research results are explained in some detail. One is the machine translation of the titles of scientific and engineering papers from English into Japanese, and the other is the machine translation of the text of computer software manuals from Japanese into English. Also, other machine translation research in Japan is briefly illustrated.

## 2.1.1 Introduction

Research in machine translation started as early as 1956 in Japan, almost the same time as when the U.S. and the European countries started their projects. Active research was done at Electrotechnical Laboratory, Kyoto University, and Kyushu University as well as at other institutions. Electrotechnical Laboratory and Kyushu University developed the special digital computers which were fitted for machine translation at that time. A special interest group on machine translation was formed in the Information Processing Society of Japan, and they have continued their activity up to now. They have changed the name of machine translation to computational linguistics, and then to natural language processing. The Eighth International Conference on Computational Linguistics, COLING 80, was held in Tokyo in 1980, where many of the present-day activities in computational linguistics in Japan were presented to the world-wide audience.

In the early time of computational linguistics research, the main subject was machine translation, especially from English into Japanese. The Japanese output was only in Roman characters because the Chinese character input and output were not available at that time. Due to this difficulty of handling Chinese characters, machine translation from Japanese into English was almost non-existent.

The National Institute of Japanese Language introduced a computer a little more than ten years ago, and then started investigation into the state of the actual usage of the Japanese language. They installed a number of off-line Kanji (Chinese character) teletypes and made statistics of character and word usages of the Japanese language. They handled a huge amount

* **Department of Electrical Engineering, Kyoto University, Sakyo-ku, Kyoto 606**

of language data from newspaper articles. The institute's activities of statistical analysis of Japanese has been continuing up to the present, extending their language usage to journals, periodicals, school textbooks as well as other printed material. They produced a number of useful results on the Japanese language, many of which were published in the series "Denshi Keisanki niyoru Kokugo Kenkyu (Japanese Language Studies by Digital Computer)" Vol. 1-10 (Shuei Shuppan Pub. Comp.).

The activities in the U.S. had great influence on the Japanese researchers of natural language processing. Besides Chomsky's theory, C. Fillmore's idea of case grammar was adopted in much of the research of Japanese language analysis, because it fitted to the Japanese language. The results of T. Winograd on natural language understanding also had influence to many Japanese computational linguists.

### 2.1.2 Analysis of the Japanese Language[22)]

The Japanese language has lots of specific features which are not found in western languages. An example is that the sentences are written using a mixture of Chinese characters, Hiragana characters, and Katakana characters. Chinese characters are mainly used for nouns and stems of verbs and adjectives. These words, however, are also permitted to be written in Hiragana characters. Hiragana characters correspond to phonetic writings and are mainly used for inflectional parts of words, connectives, postpositions as well as a few others. Katakana is used mainly for the transliteration of foreign words.

Japanese sentences have no spacing between words, and are written continuously using these three kinds of letters. Therefore the morphological analysis of a Japanese sentence by computer is a kind of trial-and-error process, which includes the arbitrary word separation operation and the determination of the parts of speech and inflections. There exist several programs of morphological analysis of Japanese.[8,14)] These are roughly classified into two different approaches: the table look-up method and the program method. The former uses a large vocabulary of words, and simple grammatical rules. The latter does not use vocabularies but instead utilizes the full information of grammatical constraints. The former approach is generally considered better because improvement is easy simply by augmenting the vocabulary of words.

There is much research in statistical and word ordering analysis of Japanese, of which we cite only some examples here.[9,18,20)]

The syntactic analysis of Japanese is studied in many places by different approaches.[12)] A syntactic analysis system based on the dependency theory was developed by Ishiwata.[3)] The analysis goes in the following sequence: (1) consultation of the dictionary, (2) locating the predicates, (3) matching the sentential parts including the predicates with the sentential patterns prepared in the dictionary, (4) determining the dependency structure, (5) applying transformation rules to the parts grammatically undetermined by the previous processes, and determining the function of these unknown parts. The software system is completely separated from the linguistic data, so that any language can be accepted if the grammatical rules and a vocabulary are given.

At Kyushu University, Yoshida made an extensive research on the governor and dependent relation between two phrases in a sentence[21)] and established seven standard modifying

relations. All the phrases of an input sentence are analyzed from this standpoint and are categorized into these seven modifying relations. By this operation an input sentence is transformed into a standard sentential form, which is next converted into a deep structure, that is, a dependency tree. Another good result he had was the detailed analysis of the predicative part of Japanese and especially post-positional part in the predicate, which incorporates the information of conjugation, semantics of postpositions, and the elements of aspects and modality.[15)]

At Kyoto University a language processing program called PLATON was developed by the author's group.[4)] By using this rule-writing system, a very sophisticated language analysis system was developed.[5)] It accepts Japanese sentences from junior high school chemistry textbooks. It performs syntactic analysis, semantic analysis and contextual analysis all at the same time. The analysis result is obtained as a kind of semantic network. It has the ability of determining anaphora relations, and of recovering the omitted words, which often occurs in Japanese sentences.

At Musashino Electrical Communication Laboratory of Nippon Telegraph and Telephone Company, a question answering system was developed which accepted natural language input.[13)] The analysis of input questions mainly depends on the case grammar, and the analysis is done similar to shift-reduced-parsing by using the stack.

At Electrotechnical Laboratory (ETL), Ikeda generalized the concept of case not only for verbs, but also for others such as nouns.[1)] Information of this kind is incorporated into the vocabulary. He intends to write a kind of word expert parser. Tanaka (ETL) improved the programming language LINGOL, and gave his new system the name E-LINGOL (Extended LINGOL).[17)] By giving context-free rules and a vocabulary to E-LINGOL, syntactic analysis can be performed easily. Syntactic rules can be made more precise by attaching semantic check functions which are to be written in program form. The system has many utility programs and provides flexible conversational facility in the syntactic analysis. The analysis is performed from left to right fairly rapidly, and all the possible analysis results are obtained. It was used for the extraction of semantic structure of Japanese sentences.[16)]

### 2.1.3 Machine Translation

At Kyoto University a machine translation system was developed by the author's group, which translates the titles of scientific and engineering papers from English into Japanese.[6)] The fundamental idea of this system is that the sentential structures of titles are finite. About one thousand different sentential structures were extracted from ten thousand title sentences. After the merging of such local structures as "adjective + noun — noun", "noun+noun — noun", "noun + of + noun — noun", "noun + and + noun — noun", we fixed only 15 fundamental sentential structures, which we called "skeleton structures". To each of these skeleton structures we assigned the word ordering of the corresponding Japanese skeleton structure. For example, to the English skeleton structure "noun1 + preposition + $noun_2$", the Japanese skeleton structure "$noun_2$ + postpositions + noun1)" was assigned. The title sentence:

"New Apparatus for Inductance Measurements"

has the skeleton

noun + preposition + noun
(Apparatus) (for) (Measurements)
and the Japanese translation is

インダクタンス　測定　のための　新しい　装置
(inductance) (measurements) (for) (new) (apparatus)

There are many ambiguous structures involved with the translation. For example, "verb-ing + noun" can be interpreted both as "verb + object" and as "adjective + noun". Simple syntactic analysis can not distinguish "measuring temperature" and "measuring instrument" as having different structural interpretations. A minimum amount of semantic information is introduced to distinguish this ambiguity. Nouns are classified into one of the six semantic categories: tool, physical material, abstract, aspect, theory, and measuring unit. Then the following descriptions are given to each verb: (i) what categories of nouns can be subjects, (ii) what categories of nouns can be objects. By checking these conditions the proper structure is determined.

The system is now running experimentally with the titles and keywords from the INSPEC database. Translated titles and keywords are stored in a database, a Japanese version of INSPEC. People can obtain access to this database by using Japanese keywords, and can receive the answer in Japanese. The translation error is less than 5% by using only 18 skeleton patterns, which is really astonishing.

A machine translation system is also being developed at Kyoto University by the author's group. This system translates computer software manuals written in Japanese into English.[7] The syntactic and semantic analyses of Japanese are very sophisticated because the sentences to be translated are complex. Essentially it is a lexicon driven analysis procedure. The case frame of each verb of Japanese is included in the vocabulary of the system. An example is

修正する　amend
surface pattern : 1 が　2　で　3　を　修正する
(agent) (instrument) (object) (amend)
deep structure : (AMEND) (AGENT (1)) (INST (2)) (OBJ (3))

The matching operation between noun phrases governed by a verb and the case slots of the verb is performed, and is almost order-free matching. The keys for matching are postpositions used in the noun phrases and the meanings of the nouns compared with the case frame of the verb.

The embedded phrases in Japanese generally have the following structure:

(the noun has a certain case to the verb)

( )　noun phrase + verb + noun
---------- ----------
**(embedded sentence)**
**I**
**(modify)**

In this case the noun which is modified by the embedded sentence is normally either the subject or the object of the verb, but the modified noun has no postposition which existed when the noun was the subject or the object of the verb. The determination of the role of the noun whether it is the subject or the object is therefore very difficult. It can only be done by the combination of the postpositions which the verb can take, the semantic information of the noun, and the case slots of the verb to which the noun is to be placed.

Example:

(a) 誤り を 修正する プログラム : Program which amends errors
(error) (object) (amend) (program)

(b) 修正する プログラム : Program which we amend
(amend) (program) Program which amends ( )
(Program for amendment)

(c) 修正すべきプログラム : Program which must be amended
must

In case (a), an object exists in the embedded sentence, so that the noun "program" can not be the object of the verb, and the structure is uniquely determined. In case (b), the noun "program" can be either the subject or the object of the verb, but in case (c) where there is a small change in postposition, the solution is unique. Some translation examples are shown in the following.

ジョブ制御文をソースプログラムライブラリに保存しておくことができる。
Job control statement can be retained in source program library.

制御データセットのユーザ見出しラベルを処理するため、利用者が用意するルーチン名を指定する。
Routine name which is provided by user is to be specified in order that user header label of control data set may be processed.

ユーザラベルをSYSUT 1からSYSUT 2に複写したい場合、LABELパラメータで標準ユーザラベルを指定しなければならない。
When user label is to be copied from SYSUT 1 to SYSUT 2, standard user label must be specified by LABEL parameter.

At the University of Osaka Prefecture, machine translation from English into Japanese is done based on the case structures which are derived from Hornby's verb patterns.[10] Though the complete case structures are still to be worked out, what has been achieved up to now provides good results in the analysis of a wide variety of English sentences. After the determination of schematic dependency relations, the problem of choosing appropriate equivalents is

dissolved using subcategories of terms and cases. The case structure of an English sentence which is transformed into the case structure of a Japanese sentence, depends on the choice of the Japanese predicate, and so a Japanese sentence is generated using a proper Japanese generative grammar.

A machine translation system is being designed at Fujitsu Laboratory.[19] This system is based on a conceptual structure which is the extension of the case grammar from a practical point of view. The conceptual structure is composed of concepts and the relations between them. A given Japanese sentence is transformed into a conceptual structure, and a corresponding English sentence is generated from it.

At the Department of Information Science, Kyoto University, a machine translation between English and Japanese was tried via Montague's logical expression as a pivot language.[11] The analytical part of the system is applicable as the input data handling part to the knowledge database.

At Kyushu University, Tamachi's group are involved in varieties of projects in language processing.[2] Their recent achievement in English-Japanese translation depended upon the dependency structure theory, which they call the D-tree method. An input English sentence is checked for vocabulary and a local parsing which evidently has no ambiguity is performed first. Then the dependency relation is established among the locally parsed terms. This process is repeated several times, and gradually a structure of a wider scope is established. The order of the Japanese word groups is next determined and the synthesis is performed. All the possible syntactic structures are produced for the final output.

At Musashino Electrical Communication Laboratory of Nippon Telegraph and Telephone Company, a new project of machine translation has been started. Their approach is greatly influenced by the recent results in artificial intelligence, especially by knowledge-based natural language understanding.

## *References*


1) Ikeda, S.: IECE Japan, *60-D* (1977) [in Japanese].
2) Ishihara, T., et al.: IECE Japan, *57-D* (1974) [in Japanese].
3) Ishiwata, T.: Denshi Keisanki niyoru Kokugo Kenkyu *VIII* (1976) [in Japanese].
4) Nagao, M. and Tsujii, J.: Information Processing in Japan, 15 (1975).
5) Nagao, M. and Tsujii, J.: Microfiche, *41* (1976).
6) Nagao, M. and Tsujii, J.: Int. Forum, 5 (1980).
7) Nagao, M., et al.: Proc. COLING 80 (1980).
8) Nakano, H., Nomura, M.: J. Inf. Process. Soc. Japan, 20(1979) [in Japanese].
9) Nakano, H., Tsuchiya, S. and Tsuruoka, A.: Proc. COLING 80 (1980) 338.
10) Nishida, F., et al.: Proc. COLING 80 (1980).
11) Nishida, T. and Doshita, S.: Proc. COLING 80 (1980).
12) Sato, T. and Tanaka, H.: J. Inf. Process. Soc. Japan, 20(1979) 865 [in Japanese].
13) Shimazu, A.: IECE Japan, Special Interest Group, *AL 78-94* (1978) [in Japanese].
14) Shudo, K., Narahara, T. and Yoshida, S.: Proc. COLING 80 (1980) 1.
15) Sudo, K., et al.: Trans. IECE Japan, *60-D* (1977) [in Japanese].
16) Tanaka, H.: IECE Japan, *61-D* (1978) [in Japanese].
17) Tanaka, H., et al.: IECE Japan, *60-D* (1977) [in Japanese].

18) Tanaka, T.: Proc. COLING 80(1980) 315.
19) Uchida, H. and Sugiyama, K.: Proc. COLING 80 (1980).
20) Uemura, S., Sugawara, Y., Hashimoto, M. J. and Furuya, A.: Proc. COLING 80 (1980) 323.
21) Yoshida, S.: Trans. IECE Japan, *55-D* (1972) [in Japanese].
22) Special Issue on Japanese Language Information Processing, J. Inf. Process. Soc. Japan, *20* (1979) [in Japanese].