Outline of the JICST Machine Translation System

Tatsuo Ashizaki

Machine Translation, JICST
Tokyo, Japan

1. Introduction

The Japan Information Center of Science and Technology (JICST) has been developing a practical machine translation system since 1986, making use of the results from "Research on Fast Information Services between Japanese and English for Scientific and Engineering Literature" [The Japanese Government MT System; Mu-Project](1982 -1986) produced by the Special Coordination Fund of the Science and Technology Agency.

2. Purpose of Development

(1) Construction of JICST English database

Demands for information on science and technology in Japan have been increasing year by year. JICST has been provided information in Japan by JOIS (JICST On-line Information Service) and STN(The Scientific & Technical Information Network). To understand the information completely and refer to it properly, it is necessary to translate the abstracts as well as the literature titles into English. In addition, it is also necessary to develop a machine translation system, because there is a limit to the manpower which can be used in the prompt translation of a English database, including many of the abstracts.

(2) Translation service business

Demands for translation services provided by JICST are increasing yearly. With an increase in Japanese information on science and technology which is supplied overseas by online database service, greater demands for translation of such information can be predicted. However, there is a limit in the current manpower system in dealing with such a large quantity of translation promptly, and therefore, efficient translation service requires the development of a machine translation system.

3. Development

(1) Development of the Japanese-English translation system

The translation system is primarily intended to generalize grammatical

rules and improve the accuracy and speed of the translation by making use of the study results of the Mu-Project. Functions of the batch process, conversational process, pre-edit of the input sentences, post-edit of the output sentences, etc. are to be developed to comply with various uses of the translation system.

(2) Development of Japanese-English translation dictionaries

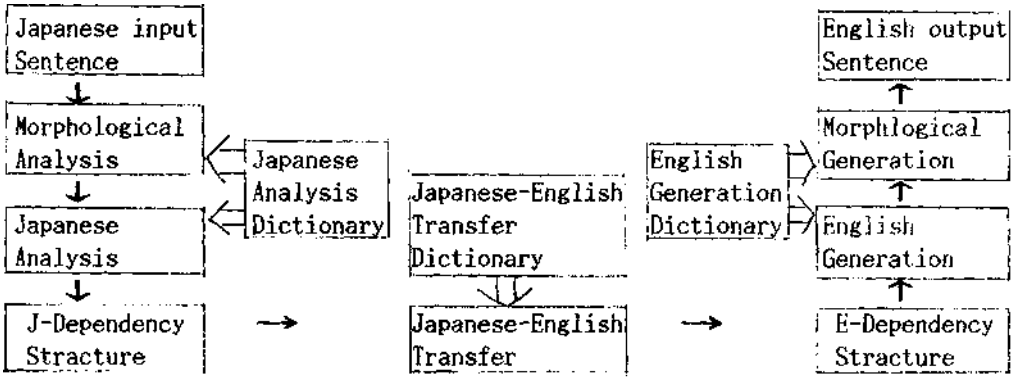
Japanese-English translation dictionaries covering about 300,000 words are to be produced by using the Japanese-English terminology list. Translation dictionaries covering about 200,000 words are also scheduled to be produced to make a medical terms.

4. Development Progress and Plan

Development of the practical machine translation system was started in 1986 with the production of the JICST terminology list and development of the Japanese-English translation system. Data collection for Japanese-English translation dictionaries was started in 1987. In 1988, the basic Japanese-English translation system was completed as well as the database construction system of English literature and abstracts in which the Japanese-English translation system was incorporated. A test operation of this system is scheduled to be carried out in late 1989, and it will be put into practical operation during 1990.

5. Translation Process

The translation system is primarily composed of translation dictionaries, grammatical rules and software.
Below is the flow chart of the Japanese - English translation process.



The translation process from Japanese into English, entering of Japanese sentences, Japanese morphological analysis, Japanese syntactic analysis, Japanese-English transfer, English syntactic generation, and English morphological generation. Then, English sentences are outputed. These translation processes are called the phrase structure transfer. These grammatical rules describe the linguistic phenomena in each process by using the grammar description language(JGRADE).

5.1 Japanese Syntactic Analysis

Japanese syntactic analysis is the process of making a tree structure of the input sentences and to make the Japanese dependency structure that shows the relationship between the words. This tree structure of the input sentences is made by breaking the sentences down into individual word in the process of Japanese morphological analysis, then referring to the Japanese analysis dictionary. Various analyses, such as word analysis of a minimal unit for the process, phrase analysis (i.e. noun phrases, verb phrases, adverbs, adjectives, etc.), and single/compound/complex sentence analysis are performed in the process of Japanese syntactic analysis. The Japanese dependency structure, that is the final form of the Japanese syntactic analysis, is the intermediate structure when translated from Japanese into English and shows the relationship between the elements of a sentence.

5.2 Japanese-English Transfer

Using the Japanese dependency structure as an input tree, the Japanese-English transfer process translates the information from Japanese into English. Main translation in this process are: 1) Japanese lexicon corresponding to English ones, 2) Japanese parts of speech and subdivision corresponding to English ones, 3) Japanese deep cases corresponding to English ones, etc.

5.3 English Syntactic Generation

The English syntactic generation uses the English dependency structure, output from the Japanese-English transfer, as an input tree referring to the English generation dictionary. In the English syntactic generation, word order is determined by obtaining information about prepositions, conjunctions and translation structure from the generation dictionary.

English syntactic generation that show the relationship between the English phrases, clauses and sentences are also produced from the English dependency structure in this process.

6. Construction of Translation Dictionaries

6.1 JICST Terminology List

Literatures written in English are extracted from the JICST's abstracts (JICST Current Bibliography on Science and Technology). The Japanese titles of the literatures are divided into individual words. Then, classified Japanese words are compared with the JICST terminology list masters (JICST Thesaurus, JIS Terminology, Scientific and Technical Dictionary, etc.) as retrieval keywords. If any of words have turned out to be inconsistent, Japanese titles and original English titles containing such disagreeable words are removed and earmarked as candidates in compiling a new list.

Then, English words corresponding to the Japanese words are selected from the output sentence and further classification codes are assigned in accordance with the JICST Integrated Classification Table. Newly selected words in such a master are merged with conventional lists. By effectively utilizing these masters, a JICST terminology list covering about 300,000 words has been completed.

Items incorporated in the JICST terminology list are as follows;

1) Japanese lexicon
2) Japanese reading
3) English lexicons
4) Field of category codes

6.2 Translation Dictionary for Nouns

Words for the noun dictionary are selected based on JICST terminology list for scientific and technical terms. In the JICST terminology list, it is predicted that a number of words will have multiple equivalents in translation, and it may become difficult to determine the proper term from among them. To overcome this problem, the most frequently used scientific and technical terms are regarded as typical equivalent terms in a translation. Words varying with respect to context and meaning are then classified. In addition, in order to meet the requirements for executing a translation, parts of speech are further subdivided and semantic markers are assigned.

Applying these procedures, new words are added to the noun dictionary based on the JICST terminology list. Dictionaries for proper nouns and unknown words are also to be added. The number of words contained in the noun dictionary was approximately 300,000 words in early 1989. It is planned that an additional 200,000 words will be added during 1990, and thus totaling to about 500,000 words.

Items to be incorporated in the translation dictionary for noun are as follows;

1) Japanese lexicon
2) Japanese reading
3) Parts of speech subdivision
4) Semantic marker
5) English lexicon
6) Field of category codes

6.3 Translation Dictionary for Verb

Words for the verb dictionary are extracted from 200,000 Japanese sentences in the JICST's abstracts. This stock of words has also been enhanced by addition from such sources as Japanese-English dictionaries, "Sa Hen" (irregular conjugation in Sa column of the Kana syllabary) nouns, Katakana words and so forth. As a result, approximately 10,000 words were produced at. early 1989. The verb dictionary structure is basically identical to the noun dictionary. It has, however, come to contain much more indigenous items. That is, the dictionary structure is composed of verbs effectively usable for the following aspects of the translation process; morphological information; syntactic information of sentence pattern; semantic information of verbs; transfer information (Japanese surface case, Japanese deep case, Japanese semantic marker and English surface case, English deep case).

Items to be incorporated in the translation dictionary for verb are as follows;

(1) Morphological information
(2) Syntactic information
(3) Semantic information
(4) Transfer information

6.4 Evaluation and Improvements

About 300,000 nouns and 10,000 verbs are being utilized for the development of translation system in 1989. In order to evaluate dictionary data, it is planned for collecting unknown words from the input sentence at the morphological analysis and unregistered words from the output sentence.

At present, the noun dictionary is primarily being produced from the JICST terminology list. However, JICST also plans to either introduce medical terms developed in other organizations or to further extend proper nouns in the future.

Regarding the verb dictionary, improvements resulting from translated sentences and enhancement provided by the use of "Sa Hen" nouns and so forth, are expected.

7. Conclusion

The machine translation system will be used to a construction of the JICST English database from the JICST's abstracts. A test operation of this system is scheduled to start in 1989 and its practical operation will be started during 1990. By developing the proposed practical machine translation system, following the benefits can be expected.

1) Construction of the JICST English database become quickly and accurately.
2) Large quantities of translation become quickly and easily.

References

1. J.Tsujii: The Current Stage of The Mu-Project, Machine Translation Summit p.122-127 (1987)
2. J.Tsujii, T.Ashizaki: The Dictionaries of MU-2 Project, International Symposium on Electronic Dictionaries p.85-87 (1988)