Outline of the Machine Translation Project of the Japanese Government

MAKOTO NAGAO

Department of Electrical Engineering.
Kyoto University.
Kyoto, Japan

The machine translation (MT) system under development is intended for translating abstracts of scientific and technical documents in both directions between English and Japanese. The well-known transfer approach was adopted as the basic model for MT. The system has many specific features, as well. The concept of subgrammar has been introduced, making it possible to change the analysis sequence in dependence on source language structure. All transfer processes are executed after the analysis phase has been completed, i. e. partial structures are not allowed to be transferred. Deep structures of input sentences are standardised before transfer, and immediately before generation there is a phase for converting the structure into internal expressions suited to the target language. Case grammar has been chosen as a model using 33 cases. Selection of words for assignment to respective case frames is accomplished by means of attributes combinations. Specific word usages are written in the word dictionary as local grammar rules applied before general ones. At present, some 15,000 words have been collected, and it is planned to cover about 1 mln terminological words from the main areas of science and technology.

PURPOSE AND ORGANISATION OF RESEARCH

This research project was initiated with a fund allocated out of the Government's science and technology promotion reserve budget. Its official R & D title is 'Fast Information Transfer System for Japanese & English Documents in Science & Technology', and the research objectives are outlined below.

In view of the necessity of promoting scientific and technological document exchange with other countries, development of the system as a means of efficient document translation includes the creation of the following:

(1) Japanese-English (to be construed as including English-Japanese) scientific & technological terminology data base dictionary based on collected terminology in the science and technology fields, compiled into a form suitable for computer use;

(2) A 'transfer' system type J-E machine translation software, and grammatical rules for use in the translation of scientific & technological documents, using the terminology data base dictionary; and

(3) An integrated translation system collectively incorporating the above.

In short, the objective is to develop a system for translating abstracts of S & T documents in both directions between English and Japanese, with emphasis placed on the information transfer of the contents of the abstract. In this sense, the selection of proper stylistic forms in the translation is not always the first priority. Unlike conventional university research, this project is characterised by the requirement to yield a system that is linked firmly to practical use, and the main focus is on combining all the latest available technologies into a practical system.

The project covers a period of three to four years from April 1982. Those organisations directly participating in the project are the Japan Information Center of Science and Technology (JICST), under the Science and Technology Agency, the Electrotechnical Laboratory (ETL), and the Tsukuba Information Center (RIPS), both under the Agency of Industrial Science and Technology, a branch of MITI. Kyoto University is involved in the project commissioned by ETL. JICST has been allotted the task of compiling the noun and terminology dictionaries of specialised lexical terms, etc., and their translation equivalents in the various scientific and technological fields; ETL and Kyoto University have been allotted the task of developing the machine translation software, grammar and verb dictionaries; and RIPS is responsible for developing a comprehensive machine translation system incorporating the results of all projects and of carrying out overall system testing.

The main researchers involved from the respective organisations are as follows: At JICST, Mr. Hiroshi Nakai was originally head of the research group, but Mr. G. Toriumi took over in April 1983; many researchers are involved in the programme under Mr. Toriumi. At Kyoto University, Mr. Makoto Nagao, Mr. Jun-ichi Tsujii, Mr. Jun-ichi Nakamura and Mr. Toyoaki Nishida are heads, and they are supported by Mr. Shinobu Takamatsu of Osaka Prefectural University, Mr. Hiroshi Kusanagi of Tsukuba University, and Mr. Makoto Hirai of Toyohashi College of Technology & Science. In addition, for the development of software, grammar rules, etc., software companies, electronics companies, translation companies, etc., are voluntarily involved in sharing the huge work load and are making a great contribution to the progress of the project. In ETL, the computer depart-

ment manager, Mr. Kashiwagi, Messrs. Yoshiyuki Sakamoto and Hozumi Tanaka are heads, and at RIPS, Mr. Mitsuji Yada, Director of the Center, is directly promoting the systems application development. These organisations have held many meetings for the coordination of mutual research and development activities, and have been working in close cooperation. For details of the separate research conducted by the respective organisations, reference should be made to the relevant reports.

BASIC CONCEPT OF THE TRANSLATION SYSTEM

So far, several machine translation systems have been investigated by various research groups around the world, The key consideration in our system in comparison with the other systems is: (1) how best to incorporate current linguistic theories, and (2) how best to handle phenomena that do not conform to linguistic theories. Language contains more exceptions than expressions that can be treated by generally defined grammatical rules, and is full of phenomena which are logically unexplainable. Therefore, in constructing a system, it is considered important to have an open-system design to maintain the adaptability to incorporate any future developments.

The main guidelines we adopted in starting the research programme are:

(1) To adopt, as the basic mechanism for the translation process, a system of 'transfer' from a list of trees to another list of trees. This allows incorporation of more and more complex linguistic theories as they emerge.

(2) To develop easily understandable grammar and dictionary systems to enable linguists and other people who are not familiar with computers to write grammar rules, dictionaries, etc. easily.

(3) To adopt LISP as the basic programming language. This is because LISP is the best language available at present for processing tree transformations and other complex symbolic representations.

(4) To adopt the 'transfer' system approach to machine translation. When possible future requirements for multilingual translation are considered, the concepts of a pivot language and symbolic logic seem to be inadequate as models for use in machine translation.

(5) In the analysis phase, emphasis is to be given to the treatment of meaning, based on case grammar. To deal with Japanese, the concept of 'deep case' is currently the best approach.

(6) To use a dictionary-based approach. This is important to cope with the numerous linguistic exceptions.

SOFTWARE SYSTEM

The selection of the software structure for language processing is of vital importance for the system performance. Representative systems so far are Montreal University's System Q and Grenoble University's ARIANE. After extensive study, we adopted a system structure which has considerable expressive power for rewriting rules; that is, the processing of a list of tree structures and the addition of various information to tree nodes are possible. This system allows the same partial structure In a rewriting rule to be repeated any desired number of times, the arbitrariness of the order of grammatical elements in a rule to be specified, and the variables to be used freely. The system has especially powerful pattern-matching capabilities for the activation of rules. Furthermore, the system can group a set of rewriting rules as a subgrammar, and, as will be described later, can deal with subgrammar networks. In this way, the system is capable of handling very complex linguistic structures, and of incorporating more advanced linguistic theories as they arise. It has the capability of representing the linguistic structures of any language, including, of course, Japanese and English.

Since the rewriting rules are very powerful, the same software system covers all the machine translation phases, the parsing process, inter-language transfer at the level of deep structure, and the generation both of Japanese and English. This is an important feature of the system. However, a separate program is used for morphological analysis and morphological generation in view of the efficiency of Japanese morphological analysis. To generate systems of this order of performance and complexity, LISP is by far the most powerful programming language available at present, because it is a well-tried and tested language with many powerful functions. Recently, several LISP machines have been built, and a Kanji version of LISP has become available. We have nearly 20 years of experience with LISP, and furthermore, we have found several times that, to write the same function as a program, LISP is much faster and more compact than PL/1. This is another factor in the present project where the objectives must be achieved in only 3 years.

The greatest problem with using LISP has been the potential problems of running the system on other types of computer. But, because this software will find a very wide application range in future intelligent information processing, and because future personal computers will become powerful enough to run LISP, we are confident that LISP is a proper programming language for machine translation.

In a machine translation system, system software and linguistic data such as grammars and the dictionaries should be separated, and a grammar and dictionary writing system should be provided to make the writing and rewriting of grammar rules and dictionaries easy for people with no knowledge of software. For this purpose, GRADE was developed, which is a system composed of a metalanguage for specifying the writing of grammatical rules, a program for compiling these grammatical rules into an internal form which the computer can directly process, and the execution of these rules on the input linguistic data. GRADE is a system that can be used at all stages of analysis, transfer and synthesis as explained in detail elsewhere [2]. GRADE is written in LISP.

In our machine translation system, we want to place emphasis on the man-machine interaction parts, such as post-editing, improvement of grammatical rules, and dictionaries. In order that the system may be gradually modified and improved according to the actual results of translation tasks, the system constructors would like to monitor examples and change the linguistic informa.-tion held in the computer during the machine translation process. This is to be achieved by utilising the powerful

interactive functions of a LISP machine terminal. In the course of the translation process, human intervention is required at various stages, particularly for pre-editing and post-editing, and ways of achieving this smoothly are under consideration. However, since they depend on the actual translation work mode, various different systems are being considered.

THE TRANSLATION PROCESS

The machine translation system that we are developing belongs to the 'transfer' type system, but it incorporates many new functions. Figure 1 shows the basic concept. In the analysis phase, especially of Japanese input, the concept of the subgrammar network can be introduced to make the analysis as precise as possible, and to raise the analysis efficiency at the same time. This is because in the analysis of a sentence, there is an analysis sequence which is suited to the structure of each respective language.

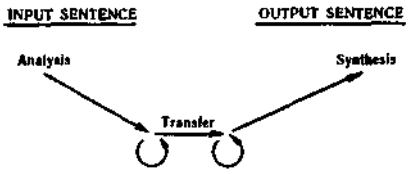

Fig. 1

Handling all the rewriting rules as equal for all the analysis situations, and examining all the possibilities for all situations are not always necessary and not always correct. For example, if an article, an adjective and a noun appear in this order in an English sentence, this group should be uniquely parsed as a phrase without examining other possibilities. Several rewriting rules corresponding to a certain linguistic construction are grouped (in what is called a subgrammar), so that the recognition and parsing of a certain construction will be done at a certain stage of analysis. These subgrammars will be joined in a network, and analysis will progress in the sequence defined by the network. Such networks are, needless to say, recursive in structure. If the use of the subgrammar concept is not desirable, all the rewriting rules may be written within a single subgrammar. In the process of applying all the rewriting rules within a single subgrammar, selective control of the order of application is possible.

Another important new concept in the analysis phase is that a rewriting rule need not always execute a parsing. For example, in the analysis of parallel noun phrases, the last words (usually the principal nouns) of parallel phrases are usually identical or end in similar suffixes, so that parallel phrases can be detected by the recognition of these features. Special rewriting rules are prepared for the detection of the extent or range of the parallel phrases. In this type of process, only the range of parallel phrases is determined, and no parsing is done. Then, in the subsequent step, the noun phrases within the ranges of the individual parallel phrases are analysed, parsed, and trees arc generated. This method of first examining the range, and then analysing the respective sections is necessary not only in dealing with parallel noun phrases, but also with various linguistic phenomena such as declinable interruptions in Japanese. These analysis steps are, it is believed, what the human brain does for the determination of the structure of complex sentences.

There is a loop between analysis and transfer in Figure 1. This loop represents a process of changing the analysed sentential structure into a more deep and universal one before inter-language transfer. This process will be essential to translate between more than two languages in the future. Deep structures of input sentences are standardised as far as possible before transfer into the other language. After the transfer phase, again immediately before generation, there is a phase for converting the structure into internal expressions suited to the target language, represented by the loop immediately before synthesis in Figure 1.

Because our system contains various new analysis stages not found in the other systems, such as a wide range situational check during the analysis, and the adjustment of deep structures immediately after the analysis, the transfer processes are all executed after the analysis phase has been completed. In this way, our system is fundamentally different from conventional systems where partial structures are converted into other languages in the course of analysis as they are determined, and is considered to yield much better results.

GRAMMAR MODELS IN MACHINE TRANSLATION

The selection of grammars, or models, for machine translation is a very important problem. Although many models have been experimented with, the majority of Japanese researchers seem to have now reached agreement that as far as Japanese is concerned, the concept of case grammar is basically a good choice*. In Europe, some researchers have recently come to recognise the advantages of case grammar over the conventional phrase structure grammar for translation between European languages. Basically we have also adopted case grammar. The most difficult problems in the use of case grammar in our system are (1) to determine what and how many cases are to be used, and (2) how to determine the case frames for the verbs. At present, no study seems to exist which gives stable structures for these two problems. Researchers differ in subtle ways in the method of defining these two categories and in the estimation of the varieties of linguistic phenomena that can be clarified by these categories. Naturally, the case structure can differ with the purpose of the analysis system, and accordingly many different ones have been proposed.

Although details must be left to a separate paper [3], our case grammar has 33 cases. Some 50 meaning attributes are established to select words for entry into the respective case frame by the proper combinations of the attributes. This means that words are screened with respect to meaning attributes for assignment to case frames. This in turn means that each verb is described in detail both syntactically and semantically. To describe the case pattern of a standard expression is not at all

* This idea was present in Japan even before it was proposed by Fillmore.

difficult, but to determine case patterns for large numbers of special expressions is very difficult and is subject to variation between lexicographers. In our project, since one person cannot handle the task of determining case patterns for several thousands of words, how to standardise the criteria for determining case patterns among all lexicographers is a serious problem. In an attempt to solve this problem, a working manual was compiled for verb case pattern description. With this, it is expected that nearly identical results will be obtained irrespective of personal differences. The manual gives many examples for reference in difficult cases.

The linguistic components to which most attention was paid in compiling the Japanese syntactic analysis rules were the treatment of compound words, parallel noun phrases, inflections, tense, aspect and modality.

DICTIONARY CONSTRUCTION

Among the many important elements in machine translation system, the dictionaries are one of the foremost. Because software systems, grammars, etc. cannot be easily modified once they have been established, the first thing attempted when improving the system and quality of translation is the revision and improvement of the dictionary contents. The work of dictionary revision and maintenance must be a patient and continuous process over time, and the dictionaries must be structured to allow the acceptance of an infinite number of revisions.

On the other hand, great discretion is also required in revising dictionaries. A dictionary which was compiled with an overall balance of contents in the initial stage may become extremely unbalanced if revisions are made frequently to incorporate specialised data, and later restoration of balance may be nearly impossible.

With all these factors taken into consideration, at present, the following three hierarchical levels are under consideration for dictionary structure:

(1) Basic dictionary (verbs, general lexical items, etc.).

(2) Terminology dictionaries by field.

(3) Private dictionaries (for individual system users).

Dictionaries (1) and (2) are to be standardised, to be kept free from frequent revision, and be maintained by dictionary system specialists.

As explained in detail in separate papers [5] and [6], verb dictionaries, noun dictionaries, etc. also contain features to cope with special features of the languages. The standard usages of verbs and nouns are expressed by their case frames and by the meaning attributes of the noun, but there are many idiomatic usages that do not fall within these frameworks. Therefore, for special usages, provisions are being made so that these can be written into the dictionary entry for that particular word. For example, when special English expressions are to be used to cover special verb expressions or special Japanese contexts, provisions are made so that these can be written in the word dictionary as local grammar rules. They may be written both under the noun or verb entry to be used, but the majority of idiomatic phrases are more often accessed from the noun side.

When rewriting rules are applied to analysis, transfer, and synthesis, the local grammar rules given in the dictionary corresponding to the respective words in the sentence are first applied, and if not successful, general grammar rules are applied next. Such specification can be given in the GRADE expression for each subgrammar in the subgrammar network. That is, our grammar system can describe in this way both the rewriting rules specific to a word and the general rewriting rules at the same level, and can specify the priority sequence for them.

In this way, our system is extensively dependent on dictionary information, and in this sense, may be called a lexicon-driven machine translation system.

SEMANTIC DESCRIPTION AND TRANSLATION WORD SELECTION

Language translation is both the translation of meaning and form of expression. In many cases, translating the words in a sentence literally word for word will not result in a good translation. There may be a limit to the accuracy of comprehension of the real meanings from only syntactic considerations. And it is widely recognised that semantic translation is indispensable. However, on the other hand, meaning is not correctly translatable without information in syntactic form. Furthermore, it is certainly doubtful that the 50 meaning attributes that we have adopted can express all existing concepts. However, they should be useful in distinguishing different semantic usages for a single word. Although the exact meaning of a noun may be projected to a concept by the combinations of several meaning attributes, it is quite difficult to do the same for the meaning specification of verbs at the present research stage. Verb usages are diverse, and a description of the behavior of individual verbs in detail is considered to be more important than conceptual classification.

The following principles are adopted in the selection of language-equivalents for each word. For each word, as many different semantic concepts are classified as represented by the combination of meaning attributes of that word, and to each of these semantic concepts, a corresponding word in the target language is assigned. However, we are not simply assigning words to words. Instead, we identify contextual usage meanings for each word of the source language, and then the words in the target language with the corresponding contextual meaning (and the required structure) are assigned to them. The basic concept of this operation is the conservation of meaning (concept). The utilisation of a thesaurus is a research topic for the future. While syntactic and semantic correspondence and transfer between two languages are the main concepts in our system, a means is provided for special linguistic expressions that cannot be dealt with at this general rule level; that is, a means for syntactic correspondence and transfer between linguistic expressions. This is an important feature of our system. Since languages have a vast range of expressions, they are not all treated by general grammatical rules. To deal with those linguistic phenomena which cannot be covered by these general grammatical rules, processing at the individual word and expression level is considered imperative.

COLLECTION OF TERMINOLOGY

To translate scientific and technological documents, not only common words, but also specialised words in scientific and technological fields must be collected with their translation equivalents. In this process, the following problems are involved:

(1) To classify specialised fields, to assign field codes to each word, and to give translation equivalents in the target language in the respective fields.

(2) To clarify the criteria for accepting a consecutive word group as a compound word which should be included in the dictionary entry.

(3) To collect newly-coined words, and to establish systems for creating translation equivalents for these words in other languages (dictionary maintenance and control).

(4) Procedure for generating dictionary systems for a large terminology data base.

Among the above, (1) is relatively free from difficulties, but (2) and (3) are formidable tasks, and no clear criteria have yet been laid down. However, at present, some 15 thousand words have been collected, and work on dictionary compilation has started. To cover the main areas of science and technology, around one million terminological words will be required. Just how far we can develop our dictionary within the framework of the present project depends primarily on the available budget and time. We are planning to start from electrical engineering and expand to medical and other fields. The EC, the Canadian Government, Siemens, etc, have data bases of over one million terminological words in their projects on multilingual translation, and these data bases are accessible by computer terminals. We are very much behind in this, and considerable efforts are required to catch up quickly. The results of these efforts must bring significant benefits to Japan as a whole.

RELATIONSHIP BETWEEN J-E AND E-J MACHINE TRANSLATION SYSTEMS

The target of the present project is to develop both Japanese-English and English-Japanese translation systems. However, this does not necessarily mean that one system will be required to cover bi-directional translation. At present, we are aiming at developing two independent systems. The two systems are compared in Table 1. To achieve high-quality translation, there is the possibility that special and different information and mechanisms are required for each translation direction. Japanese analysis grammar, for example, may not necessarily be the same as Japanese generation grammar. Hopefully, at an advanced research stage, the J-E and E-J translation systems will be integrated into one with a common grammar for analysis and synthesis, and with the same dictionary for both translation directions. We have to wait for this possibility until the completion of both systems to a satisfactory level.

Let us consider, for example, the inter-language transfer phase. While, with the J-E transfer grammar, a certain internal structure in Japanese is transformed into an English structure as specified by a transfer rule, the same English structure can simply be re-transferred into the original Japanese internal structure with the E-J translation. Ostensibly, there seems to be a one-to-one correspondence between the J-E and E-J transfer processes, and one common set of transfer rules for both directions may be feasible. However, to play it safe, we are generating two independent sets of transfer rules at present, because we have still little confidence in this reversibility of the rules. The same approach is being taken for other stages of translation.

Table 1

	Japanese-English	English-Japanese
Software	Common	
Morphological analysis	Japanese analysis & English generation	Japanese generation & English analysis
Syntactic and semantic analysis	Japanese analysis grammar	Japanese generation grammar
Inter-language transfer	J-E transfer rules	E-J transfer rules
Generation grammars	English generation grammar	Japanese generation grammar
Dictionaries	J-E dictionary	E-J dictionary

INTEGRATED SYSTEM

The target of the present project is to develop J-E and E-J machine translation systems for scientific and technological abstracts, thus realising two-way information transfer between Japanese and English. The prime objective is to bring the system to the stage where it can be put into practical use, leaving the question of the literary quality of the output to one side for the moment.

Typical applications of the eventual system are in the batch translation of abstracts at JICST, etc. for collective post-editing, and an on-line conversational mode to be used for writing research papers in English via machine translation [7],

In any case, as we have no practical experience of machine translation systems in Japan, the selection of the overall system configuration for use in the various applications must be left to a future study. We are developing a core system that can be used as a comprehensive system base and that can be adapted to any particular applications.

CONCLUSIONS

The diverse features of the present R & D project have been described. The main parts of the software and the Japanese-English translation grammar were completed in the first year of the project (March, 1982). During 1983, the system was to be operational on a small scale to produce real translation results so that the shortcomings of the system could be pinpointed and im-

proved. At the same time, up to the end of the fiscal year, the main sections of the English-Japanese translation system were to be designed. At the end of the 4th year, all the sections will be completed and operational, and the quality of the output will be evaluated.

Any machine translation system is bound to reach an impasse sooner or later. This seems to be unavoidable, because language is intrinsically free, and we are trying to confine it within an artificial framework. However, the stage at which the system reaches this impasse depends on the structure of the system. While only simple sentences are being translated, the true potential of the system is not revealed, but when more and more complicated sentences are treated, systems based on simple models soon reach this impasse, and further improvement becomes very difficult.

The system we are developing is aimed at incorporating relatively sophisticated linguistic theories, so that its advantages over other systems may not be apparent in the beginning, but as the subject sentences become more complex, the differences will become clearer. Even so, our system is also bound to reach an impasse, in the very nature of special expressions. To overcome these difficult problems, various special mechanisms are incorporated in our system, including dictionary entries of specific solutions for specific cases, and the provision of additional processes (loops in Figure 1) for standardising special internal sentence constructions as far as possible. Just how far these mechanisms will be effective in overcoming the difficulties of dealing with complex sentences can only be found through actual use, but we are hoping that the inevitable impasse in machine translation can be banished as far into the future as possible. Your cooperation and support in this national R & D project are sincerely appreciated.

REFERENCES

1. Sakamoto, Y. Morphological analysis. *Johoshorigakkai. Natural Language Process Study Group data*, July 1983.
2. Nakamura, J. Software for grammar rules, GRADE. *Ibid.*
3. Tsujii, J. Syntactic analysis of Japanese. *Ibid.*
4. Nishida, T. Basic design for a transfer process. *Ibid.*
5. Nakai, H. Compilation of dictionaries for non-inflective words. *Ibid.*
6. Sakamoto, Y. A dictionary for inflective words and their case frames. *Ibid.*
7. Jada, M.; Nagao, M, Basic design for an integrated machine translation system. *Ibid.*