Current Status and Future Trends in Machine Translation

Makota NAGAO
Dept. of Electrical Engineering, Kyoto University, Yoshida-houcho, Sakyo-ku, Kyoto-shi 606, Japan

A survey of the current machine translation systems is given, which includes not only activities in Japan, but also abroad, especially European, US and Canadian activities. Then the components of a machine translation system are explained from the standpoint of software, linguistic components, and users' demands. The importance of pre-editing and post-editing is stressed. The semantic and contextual processings are essential to obtain a better translation quality, which are the future problems to attack. Attention is given to the difficulty of contemplating a pivot method in machine translation instead of transfer methods, because the projection from a word or a phrase to a concept is very difficult if we want to have a very exact concept representation and translation. A new transfer method which accompanies the pe-transfer structural adjustment and post-transfer adjustment is explained. This method was adopted by the Japanese governmental project of machine translation which was directed by the author. Various mechanisms of structural transformations in the transfer and generation processes are explained, which are necessitated by the language translation between the two languages of different language families like Japanese and English.

Finally some comments are given from the standpoint of users of machine translation systems. Systems always are imperfect, and users must use them after recognizing the possibilities and the limitations of the system.

North-Holland
Future Generations Computer Systems 2 (1986) 77–82

Overview of Machine Translation Systems

Machine translation systems are complex, and hence not easy to understand in their entirety. They are overviewed below from several standpoints.

(1) Translation systems receive texts written in the "source" language and produce their equivalent in the "target" language. Ordinary systems translate a single source language into a single target language. By contrast, the systems being developed by the European Economic Community (EC) are multilingual and can translate multiple source languages into multiple target languages.

(2) Machine translation proceeds in three steps: an input sentence is analyzed, its content is then represented by an internal structure, which, in turn, is used to generate a target-language sentence. This is called "transfer approach". An ideal approach, called the pivot language method, assumes that a universal internal structure can be applied to all languages. However, the difficulties involved in constructing such ideal intermediate linguistic representations have caused most translation machines to adopt a transfer method. With the exception of morphological and syntactic an analysis, these processes are not yet sufficiently understood.

(3) Machine translation systems must perform not only morphological and syntactic analyses, but also semantic and contextual analyses, and even text understanding, to produce high quality translations. With the exception of morphological and syntactic an analysis, these processes are not yet sufficiently understood. Some systems perform syntactic and semantic analysis separately, in that order, while others perform the two processes almost simultaneously. In recent systems,the latter type outnumbers the former.

(4) It is almost impossible to completely elucidate the structure of any given natural language. As machine translation technology progresses, the following strategies seem viable for developing practical machine translation systems: (a) Machine translation systems are applied to

specific documentation, such as scientific and technical papers. For such documents, even incomplete translations are considered acceptable, in as much as they enable specialists in the field concerned to understand the informational content of the original text.

(b) Machine translation requires the active involvement of human translators to post-edit low-quality output texts. Machine translation systems must be built with emphasis on maximizing the total efficiency of the processes involved, ranging from text entry, pre-editing, translation, and post-editing, to output printing.

(c) Besides the post-editing, pre-editing is likely to be required in some cases to modify the style of input text. A so-called restricted language (sublanguage) constitutes a valuable tool for pre-editing.

2. Grammar Models and Analysis Models

2.1. Context-Free Phrase Structure Grammar

Many grammars for computer text analysis are based on the Phrase Structure Grammar (PSG) by Chomsky. He assumed that grammar was defined as a set of rewriting rules, and that sentences were generated through repetitive application of these rules.

2.2. Expansion of Context-Free Phrase Structure Grammar

2.2.1. Attribute grammar

Unlike programming languages, natural languages often allow multiple syntactic interpretations. To minimize the possibility of multiple interpretations, researchers are constructing more sophisticated grammars, and introducing semantic elements. A concept referred to as attribute grammar provides the basis for these efforts. An attribute grammar consists of rewriting rules, which differ from those of the Context-Free Grammar (CFG) in that each symbol in a rewriting rule has a parameter. The format for a rewriting rule of an attribute grammar is shown below:

$$A[a] \rightarrow B[b] \ldots C[c]$$

Parameters can represent anything like multiple pieces of information, tree-like structures, etc. In many cases, sub-classified parts of speech, information on grammatical gender/number/case, or semantic primitives are used as parameters. Checking grammatical gender/number/case, and semantic matching are performed between neighboring words and phrases.

Lexical Functional Grammar (LFG) is one of the frame-works for grammar description currently attracting the attention of linguistic researchers. While LFG is based on CFG, LFG's condition checks are separated from rewriting rules and defined as independent equations.

A grammar description format called Definite Clause Grammar (DCG) has become widely known since it was implemented in Prolog. DCG is also a variant of the CFG-based attribute grammar.

2.2.2. Generalized Phrase Structure Grammar

Chomsky contended that natural languages could not be explained using CFG. In fact, grammars for machine translation has to be at least context-dependent. However, Gazdar recently claimed that enhanced versions of CFG would produce good results, because even context-dependent grammars cannot adequately explain the structure of a language. Gazdar devised a grammar for grammatical rules by generalizing regularities found in grammatical rules; he called these regularities metarules. Gazdar proposed a model to demonstrate that sentence structure could be dealt with by a set of context-free rewriting rules generated by these meta-rules.

2.2.3. Tree-to-tree transformation grammar

Large-scale machine translation systems implemented to date use a grammar description system, a powerful framework which converts one tree structure into another. GRADE (GRAmmar DEscriber), a typical grammar description system which was developed at Kyoto University, can treat words arranged in arbitrary order as well as trees consisting of any number of elements to permit free manipulation of Japanese and other languages. GRADE allows nodes of a tree to be annotated with additional information, and permits condition checking on additional data between various nodes to be easily described. In addition, user-defined functions can be invoked in GRADE.

Machine translation requires a description language to express the complex gramatical phenomena inherent in natural languages in a form understandable to humans. Such description languages must be equipped with a powerful capability for tree structure conversion.

3. Machine Translation Control Techniques

3.1. Multiple Analysis Results

The problem of control concerns methodologies for implementing syntactic analysis (performed by applying formal rewriting rules to input sentences) and text generation.

One of the problems involved in rewriting-rule-based sentence analysis is whether the analysis outputs a single result or all possible solutions. The latter type is implemented by either a "depth-first" search, which produces one solution at a time, or a "breadth-first" search, which tries to find all solutions simultaneously. The method of analysis most promising for machine translation seems to be one that provides a solution which seems most appropriate and, if that solution turns out to be unsatisfactory, offers the next most likely candidate. This process continues until the human operator is satisfied with the result.

3.2. Subgrammar Network

Generally, hundreds or even thousands of grammatical rules are necessary for language analysis. It is impractical to check whether all these rules can be applied to all intermediate analysis. Instead, rules for analyzing similar linguistic structures are grouped into a "subgrammar" [1]. Appropriate subgrammars are selected to analyze specific corresponding linguistic phenomena as they occur in the input text. In this approach, all subgrammars are linked together to perform sequentially analysis processes ranging from text entry, analysis and structure generation.

A link uni-directionally connects one subgrammar with another. Actually, the subgrammars are networked, because a link may form a loop, or may branch. This subgrammar network is to be programmed so that individual subgrammars can work as if they were co-routines. Such implementations have yet to be performed.

[1] The subgrammar discussed in this paper differs from that mentioned by Kittrage et al. Kittrage's subgrammar refers to domain-specific grammars; for example, a grammar suitable for sentences used in weather forecasting.

4. Analysis-Related Problems

4.1. Syntactic Analysis

Analysis grammars based on case grammar are used with increasing frequency. Most analyses of Japanese sentences are carried out using case grammar, because the word order in Japanese sentences is less strictly determined than in other languages, such as English, and therefore the analysis must make use of the meanings of given sentences.

Problems involved in case-grammar-based analysis include the following:

(1) What cases should be established? Although it involves considerable difficulty, a single case system should be shared by the source and target languages.
(2) Case structures must be determined for individual verbs. Also to be determined are parameters to express what words can satisfy individual slots in case structures.
(3) In general, semantic origins are likely to be used as parameters. Here, the number of semantic origins necessary to resolve ambiguity must be pre-determined.

4.2. Analysis of Phrases Linked by Conjunction

Conjunction-linked phrases often appear in sentences. Generally, these are noun phrases, or appear in sentences linked by conjunctions (including compound and embedded sentences). Analysis of such phrases tends to produce a large number of results: it is often very difficult to select the correct structure.

The structure of compound words cannot easily be determined. Particularly difficult is analysis of compound words containing parallel structures. Ideas which are expressed in Japanese using words simply linked by conjunctions may be represented in other languages with complex structure including prepositions and other parts of speech.

4.3. Anaphora

Determination of anaphoric reference also requires semantic processing. Words indicated by pronouns and demonstrative pronouns, such as "(kono)" (this), "(sono)" (its), and "(kare)" (he), must be referenced to a noun during analysis. Otherwise, Japanese sentences cannot be translated into German, French, or other languages with grammatical gender.

4.4. Estimation of Ellipses

In languages like Japanese and Russian, the subject and other words and phrases are often omitted. Therefore, complete sentences cannot be produced without a capability for estimating ellipses. This problem requires context processing; in Japanese, polite expressions may provide a key to ellipsis estimation. So far no translation system has been equipped with this function. It is essential to devise some approach to translation which can treat ellipsis-related problems.

4.5. Tense, Aspect, and Modality

The framework of a sentence can be almost completely described by the processes discussed above. Besides sentential structure, other problems related to tense, aspect, and modality must be resolved.

Analysis of tenses and aspects expressed by Japanese suffix-like words, such as auxiliary verbs and postpositional particles, involves complex problems that have been studied in various research activities. Recent efforts have been made to review and organize such words so that they can be dealt with by computers. Work is also currently underway on organizing model expressions consisting of a combination of suffix-like words, and on comparing tense, aspect, and modality in Japanese and other languages.

5. Intermediate Representation

Sentence analysis results in an intermediate representation. Machine translation methods depend significantly on the form used for the intermediate representation.

Currently popular forms include phrase structure representation and dependency structure representation (tree structure), in which each node of a tree structure is annotated with syntactic, semantic, and other information. In case grammar, case-related information is further added to the branches of a tree structure. To express embedded sentences, a tree structure must have some means of indicating words corresponding to so-called "gaps".

Semantic network, conceptual dependency, and symbolic logic representations seem suitable for question-and-answer systems and other applications in restricted fields. They cannot easily be used in machine translation, however, because they have drawbacks similar to those of the Montague grammar.

6. Transfer and Generation

6.1. Syntactic Transfer

The transfer stage transforms the results of sentence analysis, the internal structure of source-language sentences, into the internal structures of the target-language sentences. The analysis results are likely to have characteristics of the source language and therefore cannot directly be converted into target language internal structures. When necessary, conversion must modify the source-language internal structures in addition to conducting word-for-word replacement.

For technical terms and other domain-specific words, word-for-word conversion can uniquely determine their equivalent in the target language. This is unlikely to common words. Even if the meaning of a common word is uniquely determined as the result of text analysis, it is not necessarily represented by a single word in the target language; in many cases, common words in the source language corresponds to multiple options in the target language.

Transfer processes to be considered are:

(1) *Single-word-to-single-word conversion*. This applies to technical terms and other domain-specific words without ambiguity.

$$W_s \rightarrow W_T$$
$$\Delta$$

(2) *Single-word-to-single-phrase conversion*. This applies to common words such as verbs. Such a word is expressed with a single substructure in the

target language.

$$Ws \rightarrow W_T$$
$$\Delta$$

(3) *Single-word-to-multiple-phrase conversion.* Processing which checks the local structure of source-language sentences before transfer may produce multiple synonymous expressions.

$$\triangle \rightarrow W_{T_1}, \ldots, W_{T_n}$$
$$\Delta \qquad \Delta$$

(4) *Single-phrase-to-single-phrase conversion.* In most cases, the local structure of a source-language sentence is transferred into a single phrase, rather than multiple phrases, of the target language.

$$\Delta \rightarrow \Delta$$

Other conversion methods, such as multiple-word-to-single-word and multiple-word-to-multiple-word, are theoretically feasible. However, these make the conversion stage extremely complicated. To eliminate the necessity of these methods, portions of the source-language internal structure that are highly dependent on the style of the original sentences are converted into more natural representations (or representations closely resembling the structure of the target language).

Any conversion must appropriately and uniquely determine target-language words (or phrases) by examining the internal structure of an input sentence as extensively as possible. The requirements for this operation must be described in great detail, which is an extremely difficult task. Because this operation definitely affects the quality of translation, detailed comparison of various languages must be made both at the general structural level and at the individual word level.

6.2. Sentence Generation

Sentence generation starts with the internal structures resulting from transfer. Each node of a tree representing an internal structure includes multiple words or subtrees. Sentence generation can be performed either top-down or bottom-up. However, when external factors, such as context and focus, determine the style of output sentences the generation must proceed top-down. In this case, properties are passed from the top node to lower nodes, and appropriate partial phrases are generated (or the appropriate expression is selected from multiple expressions placed at a node). Therefore, embedded sentences in the internal representation must sometimes be represented by infinitives or noun phrases in the output.

Syntactic conversion performed in the sentence generation stage can cause the part of speech of a word to be changed. To cope with such modification, word dictionaries must store derivative relations for each word (including reverse derivations). If a word has no variant for a particular part of speech, an alternative word of that part of speech (with an equivalent meaning) must be placed in the dictionary entry for that word.

6.3. Style Conversion

Japanese differs greatly from English in its form of representation. As described by the terms "HAVE-language" or "DO-language", English places emphasis on individuals such as agents and possessors. Japanese, on the other hand, is a "BE-language" or BECOME-language", because it aims to entirely express a state or action. Therefore, in English-to-Japanese translation, HAVE- and DO-type sentences must be converted into BE-and BECOME-type sentences.

Eliminating or pronomializing duplication of the subject in generated sentences is essential to produce easy-to-read translations. Due to the difficulties involved, however, these operations have seldom been implemented in machine translation. It is also tremendously difficult to add or remove language-specific words such as articles, depending on the target language: no promising algorithms is gradually shifting from syntactic parsing to semantic and context analysis.

7. Future Directions for Machine Translation Research

Efforts made over the past decade have resulted in considerable progress in machine translation research. Syntactic analysis was most eagerly pursued in this period. Recently the focus of research is gradually shifting from syntactic parsing to semantic and context analyses.

Key research subjects include:

(1) Research on the structure of languages

(2) Research on semantics

(3) Research on usage
(4) Research on translation from a knowledge engineering standpoint
(5) Development of machine-translation-oriented software
(6) Development of man-machine interfaces for machine translation
(7) Development of dictionaries
(8) Evaluation of translation quality
(9) Research on restricted languages (sublanguage).