A Prototype English-Japanese Machine Translation System for Translating IBM Computer Manuals

Taijiro Tsutsumi

Natural Language Processing
Science Institute, IBM Japan, Ltd.
5-19, Sanban-cho, Chiyoda-ku
Tokyo 102, Japan

ABSTRACT

This paper describes a prototype English-Japanese machine translation (MT) system developed at the Science Institute of IBM Japan, Ltd. This MT system currently aims at the translation of IBM computer manuals. It is based on a transfer approach in which the transfer phase is divided into two sub-phases: English transformation and English-Japanese conversion. An outline of the system and a detailed description of the English-Japanese transfer method are presented.

1. Introduction

The Science Institute of IBM Japan, Ltd. has been involved in English-Japanese machine translation for four years (1). We have developed a prototype capable of translating IBM computer manuals into Japanese. This system is based on a transfer approach in which the transfer process consists of English transformation and English-Japanese conversion. This MT system aims at 1) high-quality translation; 2) an easily maintained transfer component; and 3) a smaller English-Japanese terminology dictionary. The transformation rules and the conversion rules are presently being constructed through tests using the IBM manual "VM/SP General Information" (60P).

We are focusing on translation of IBM computer manuals for 3 reasons: 1) high-quality translation is expected in a limited area; 2) English IBM manuals are presumably written as clearly as possible according to an IBM internal standard; 3) we already had a practical English-Japanese terminology dictionary for human translators.

Most MT systems developed in Europe and the U.S. deal with language pairs in the Indo-European language group (2). In the case of English-Japanese translation, since both languages are categorized in different language groups, a more powerful linguistic mechanism must be implemented. For instance, word order and sentence style are different and moreover an English word sometimes corresponds to more than one Japanese equivalent. To overcome these difficulties, an English-Japanese or Japanese-English MT system might be based on a transfer or interlingua approach with a wide range of tree-transducing capabilities and a semantic processing mechanism.

2. Overview of the system

Fig. 1 shows the overall translation process. First of all, an English sentence is syntactically analysed in the English analysis phase. The output of this analysis is one or more English parse trees. Second, in the English-Japanese transfer phase, an English parse tree, or an English intermediate representation, is transferred to a corresponding Japanese tree, or a Japanese intermediate representation. During this transfer, an English parse tree is at first transformed by the transformation component to an English tree in Japanese-like style, and this result is converted to a Japanese tree by the English-Japanese conversion component.

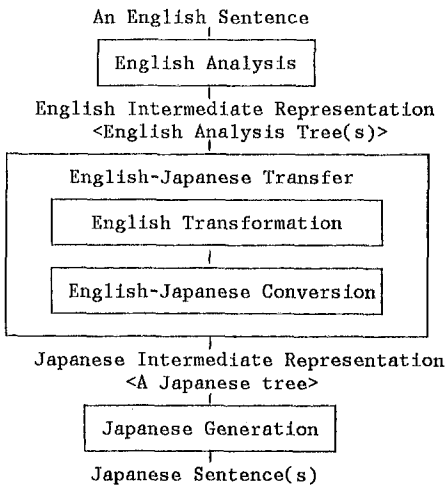

Fig. 1. Overall translation process

Finally, in the Japanese generation phase, one or more Japanese sentences are produced by operations such as generating Japanese auxiliary verbs, determining Japanese case particles, and rearranging word order. At present, the components shown in Fig. 1 are all implemented in LISP.

3. English Analysis

For analysing English, we are making use of the English parser, the English analysis grammar, and the English analysis dictionary developed by G. Heidorn et al. at the IBM T.J. Watson Research Center (3). The English analysis is based on an augmented phrase structure grammar and is syntactically performed in a bottom-up and parallel manner. This English analysis aims at area-independent, high-performance and fail-soft analysis. The area-independent feature means portability of this analysis component to other application areas. The fail-soft feature is important for a practical MT system which should provide some Japanese segments for a human translator even if the parser fails to analyze the input sentence as a complete sentence.

As the syntactic analysis of English sometimes produces more than one parse tree, the English parser computes metric values which indicate plausibility of the parse trees based on the characteristics of the modifications between phrases (4). When more than one parse tree is obtained by analysis, semantically incorrect parse trees are discarded during the English-Japanese transfer. If more than one Japanese tree remains after the transfer, the metric values copied from these English parse trees to corresponding Japanese trees are used to rank these Japanese trees in terms of plausibility. The Japanese tree which has the least value, namely the most plausible one, is chosen by the MT system.

4. English-Japanese Transfer

Generally, the transfer process of a transfer approach including semantic processing tends to become complicated and then difficult to maintain. But a transfer approach seems to be the most straightforward for implementing human translators' knowledge which includes various types of linguistic information such as specific words, syntactic structures, and semantic information.

There are many English-proper expressions, such as 'It-that', 'too-to', and 'there-be'. Their styles are very different from Japanese ones and have no simple contrast expressions in Japanese. The English-Japanese transfer component of our system is divided into two separate components: an English transformation component and an English-Japanese conversion component. We call our approach a two-pass transfer method. By using English transformation rules, the English transformation component rewrites an English parse tree and produces a new style of English tree which is close to Japanese syntax. This can then easily be converted to a corresponding Japanese tree. When we expect different English expressions to be translated to the same Japanese expression, we only have to write English transformation rules instead of E-J transfer rules of a conventional transfer approach. Moreover, when we have a MT system change a Japanese expression, we are required only to modify some E-J conversion rules instead of modifying a larger number of relating E-J transfer rules. Consequently, the two-pass transfer method provides us with modularity and maintainability of the transfer component.

4.1 English Transformation

English transformation is performed by using English transformation rules and a transformation dictionary. The transformation sometimes requires a derivative form of an English word, such as a verbal form of a noun and an adverbial form of an adjective. The transformation dictionary contains this sort of derivational data. The transformation rules are categorized into groups according to syntactic categories of nodes of parse trees. Each group is also classified into several sub-groups. For example, the rule group for a sentence consists of 22 sub-groups, such as an inversion-rule sub-group, an insertion-rule sub-group, and an ellipsis-rule sub-group. The following are examples of applications of the rules to sentences.

It is required that you specify the assignment.
-> That you specify the assignment is required.

There are several records in the file.
-> Several records exist in the file.

System operation is so impaired that the IPL procedure has to be repeated.
-> Because system operation is very impaired, the IPL procedure has to be repeated.

The routine has a relatively low usage rate.
-> Usage rate of the routine is relatively low.

The following are examples of applications of the rules to noun phrases.

execution of the program
-> executing the program

a disk available with ...
-> a disk which is available with ...

one type of ...
-> one-type-of ...

The transformation is performed in a top-down manner along an English parse tree. At each node of a tree, a corresponding rule group is retrieved according to the syntactic type of the node and this rule group is applied to the sub-tree only once. In this application of the rule group, each sub-group is sequentially applied to the sub-tree only once. If a matching pattern of a transformation rule matches the sub-tree and a target pattern produces a new tree, the rest of the rules in the sub-group are no longer used and processing of the next sub-group begins. We have designed the rule groups and their sub-groups to avoid backtracking and repetitive application of the same rule.

4.2 English-Japanese Conversion

A transformed English tree is converted to a corresponding Japanese tree by using conversion rules and a conversion dictionary. The functions of this process are 1) determining appropriate Japanese syntax, equivalents, and additional linguistic data such as tense, aspect, modality, and voice; and 2) disambiguating modifications of English phrases.

4.2.1. Semantic Markers

One of the basic approaches to semantic processing in MT is to use semantic markers of nouns. We have defined 24 semantic markers specific to computer manuals, which will be effective in translating IBM computer manuals. Table 1 lists all of the semantic markers and their meanings.

Semantic Markers	Meanings	Semantic Markers	Meanings
LC	Logical Container	WK	Work/Action
LE	Logical Entry	PS	Predicate
LP	Logical Path	AP	Attribute of PS
DM	Document	SL	Supply
ST	State	PT	Part
TH	Technique/Theory	DT	Term of documents
FA	Feature/Ability	ML	Material
IF	Information	TM	Time
AT	Attribute	PL	Place
VA	Value of AT	PN	Person's Name
HM	Human	PO	Point
UD	Unit/Device	OG	Organization

Table 1. Semantic markers

Nouns in computer manuals have one or more semantic markers. For example, "file" has "LC" and "LE", "program" has "LE", and "operator" has "LE" and "HM". This set of markers is so simple that maintenance is easy.

4.2.2. E-J Conversion Dictionary

In the English-Japanese conversion dictionary, conditions for conversion are described by a combination of English syntax, semantic markers and sometimes specific Japanese words. The conversion dictionary is divided into sub-dictionaries, such as a verb-dictionary, a noun-dictionary, and a prepositional-dictionary. Fig. 2 shows an example of an entry of the verb-dictionary in the case of "provide".

```
("provide"
  ((SB   (S ((LE UD) Y1 "ga")))
   (DO   (S ((FA AT) Y1 "wo")))
   (P    "sonae" PY1 (V SHIMO1 NIL JYOOTAI TRANS)))
  ((SB   (S ((DM LE UD) Y1 "ga")))
   (DO   (S ((HM)       Y1 "ni")))
   ("with" (S ((IF FA AT) Y1 "wo")))
   (P    "teikyo" PY1 (V SAHEN NIL KEIZOKU TRANS))) )
```

Fig. 2. An example of an English-Japanese conversion dictionary entry

The upper half of the description in Fig. 2 specifies that if the subject of a sentence has semantic marker "LE" or "UD" and the first object has marker "FA" or "AT", then choose the Japanese case particle "ga" for the first Japanese noun phrase, the Japanese case particle "wo" for the second one, and the Japanese verb "sonae" as the proper equivalent for the English verb "provide". "Y1" and "PY1" in Fig. 2 specify types of corresponding Japanese sub-trees to be generated. The lower half of the description gives a similar rule to the previous one except for an additional condition on a prepositional phrase. This part specifies that if the conditions are met, then use Japanese case particles "ga", "ni", and "wo" in this order and select "teikyo" as the appropriate Japanese verb.

The verb-dictionary is used to convert an English surface case structure into a Japanese one directly by depending upon the semantic markers. This conversion must be more efficient than in the case where deep cases are introduced so as to pursue similar semantic processing. This conversion determines an appropriate Japanese verb, Japanese case particles, and Japanese syntax of a simple sentence at the same time. In some cases, an appropriate Japanese equivalent for an English noun phrase is successfully selected based on these conditions when the English noun phrase has more than one Japanese equivalent. Moreover, application of these entries also means a semantic check of the input from the computer area's point of view. Consequently, if there is no entry applicable to the input simple sentence, it is deemed inappropriate for computer manuals and is rejected by the system. This contributes to disambiguation of English analysis trees.

Additional linguistic data of an English simple sentence concerning tense, aspect, modality, and voice, are also converted to corresponding data of a Japanese tree by using a contrast conversion table and the conversion dictionary. For example, voice and aspect of an English sentence are changed in a Japanese sentence according to the characteristic of the verb.

4.2.3. E-J Translation of Simple Noun Phrases

One of the issues in MT is how to create and maintain a large terminology dictionary. Generally, a technical document includes a number of technical noun groups. We call a noun phrase which basically has no post modifier a simple noun phrase (SNP), such as "a procedure library", "system-to-operator communication", "IBM supplied licensed and nonlicensed programs" and "page 34".

Our MT system facilitates a component for translating SNPs. Even if the terminology dictionary does not have the entry in whole, a long SNP which is composed of many words is successfully translated by appropriately assembling the dictionary data of all elements of the SNP. This is mainly due to the similarity of syntax of SNPs between English and Japanese.

The functions of the SNP translation component are to choose appropriate Japanese equivalents for various parts-of-speech(e.g. noun, adjective, adverb); to insert "no" between noun phrases; to reorder Japanese equivalents; to process conjunctions within a simple noun phrase; and to handle hyphenated words. These are achieved by using a special dictionary for translating SNPs and co-occurrence frequency data of words or semantic markers in IBM computer manuals.

4.3 E-J Conversion Process

The English-Japanese conversion component subsequently converts a transformed English tree to a Japanese tree in a bottom-up and parallel manner along the English tree.

First of all, the English-Japanese conversion dictionary is searched for all English words which are terminal symbols of the English parse tree. This is part of English-Japanese conversion of the lowest level sub-trees of the English tree. An upper level English sub-tree is converted to a corresponding Japanese sub-tree by using the English-Japanese conversion rules and by using the English-Japanese conversion results of the current level English sub-trees. The category of the top node of the upper sub-tree determines which set of English-Japanese conversion rules is to be applied. During the conversion of sub-trees, semantic processing is performed according to the data in the English-Japanese conversion dictionary as mentioned earlier.

5. Japanese Generation

The Japanese generation component produces one or more Japanese sentences from a Japanese tree which conveys Japanese syntax, Japanese equivalents, and other information.

The functions of this component are to generate Japanese auxiliary verbs; to determine appropriate Japanese equivalents of adverbs, negation, determiners and conjunctions including subordinate conjunctions; to position Japanese adverbial phrases in a Japanese sentence; to modify Japanese case particles; to reorder Japanese noun phrases; to insert punctuations; and to erase a duplicate Japanese subject. Japanese auxiliary verbs are generated based on Japanese verb information, such as the original form of the verb, the conjugation type of the verb, tense, aspect, voice, and modality.

6. Conclusion

We have described a prototype English-Japanese machine translation system based on a two-pass transfer approach. Introduction of separate English transformation in the E-J transfer makes the transfer component easy to maintain.

We have proposed a set of semantic markers specific to computer manuals and the English-Japanese conversion dictionary so as to perform high-quality translation. The mechanism of selecting appropriate Japanese equivalents and syntax is simple and effective. We will continue to enhance our MT system to translate many kinds of IBM computer manuals into high-quality Japanese.

7. References

1. Tsutsumi, T. "On the Machine Translation from English to Japanese" in Tokyo Scientific Center Report N:G318-1571 (1982)
2. Slocum, J. "A Survey of Machine Translation: its History, Current Status, and Future Prospects" in AJCL 11-1 (1985)
3. Heidorn, G.E., K. Jensen, L.A. Miller, R.J. Byrd, and M.S. Chodorow. "The EPISTLE Text-Critiquing System" in IBM Sys. J. 21.3 (1982), 305-326.
4. Heidorn, G.E. "Experience with an Easily Computed Metric for Ranking Alternative Parses" in Proc. 20th Annual Meeting of the ACL. Tronto, Canada (1982), 82-84.