# LANGUAGE CONTROL FOR EFFECTIVE UTILIZATION OF HICATS/JE

Hiroyuki KAJI
Systems Development Laboratory, Hitachi,Ltd.

## 1. INTRODUCTION

The quality of machine translation results depends heavily on input texts. Language control will enable the maximum benefit to be drawn out from a machine translation system. Language control or pre-editing approaches are practicable in many circumstances a machine translation system is used to produce foreign language texts from native language texts, although cost-effectiveness should be of course the criterion for actually being adopted.

In this paper, we present the language control approach in HICATS/JE (Hitachi Computer Aided Translation System/ Japanese to English). For the last few years, we have made efforts to establish a pre-editing method and develop a pre-editing aid as well as to improve the grammar and dictionary of the system.

First, a brief outline of HICATS/JE is given in Section 2. Then, Section 3 describes guidelines for writing Japanese sentences fitted to machine translation, and pre-editing conventions. Section 4 describes a pre-editing aid diagnosing input sentences and pointing out critical expressions. Lastly, Section 5 describes an experiment evaluating the cost-effectiveness of pre-editing.

## 2. BRIEF OUTLINE OF HICATS/JE

HICATS/JE, of which the technical issues were reported at Hakone MT Summit[l], is summarized as follows.

(1) Language pair: Japanese to English.

(2) Subject area: Science and technology.

(3) Document type: Manuals, technical reports, abstracts, etc.

(4) Hardware configuration: Two configurations are available. In a distributed environment, machine translation and pre/post-editing are

performed on a host computer and a workstation respectively. In a stand-alone environment, both are performed on a workstation.

(5) Translation method: A case-grammar based sentence analyzer transforms an input Japanese sentence into semantic representation consisting of semantic relations among concepts. A phrase-structure-grammar based sentence generator generates an English sentence from the semantic representation. Gaps occasionally remaining in semantic representation are coped with by transformation rules of semantic representation.

(6) Size of grammar: Approximately 5,000 rules are included which cover the principal linguistic phenomena observed in science and technology texts.

(7) Dictionary configuration and vocabulary size: The dictionary consists of a basic term dictionary (containing 50,000 words), a technical term dictionary for science and technology (containing 250,000 words) and a user dictionary. They are successively accessed according to fixed priority.

(8) Execution mode: Both batch and on-line (Translation itself is executed noninteractively) are available.

(9) Throughput: 48,000 to 120,000 words per hour when the fastest cpu (HITAC M680) is used.

(10) Support software: Bilingual text editor for pre/post-editing, dictionary editor, Japanese sentence diagnosis software, etc.

## 3. PRE-EDITING METHOD FOR HICATS/JE

### 3.1 Guidelines for Writing Japanese Sentences

HICATS/JE users are recommended to prepare source texts according to the following guidelines.

(1) Make simple sentences as much as possible.
Structural ambiguities are often included in a complex sentence. Moreover, it is difficult to analyze the semantic relationship between sentences connected by *'ren'yo chushi'* form. Accordingly, making simple sentences is essential for obtaining good results from machine translation.

(2) Put a word close by its governor.
The ability of the system to disambiguate based on semantic and contextual knowledge is not sufficient. Final decision in sentence analysis often relies on statistical knowledge that a word tends to actually depend

on the nearest one among its grammatically possible governors. Accordingly, it is desirable that a word should be put as close as possible to its governor.

(3) Avoid unbalanced coordinate structure.
Disambiguation of coordinate structure is difficult, as it often requires domain or contextual knowledge. Therefore, it is desirable to avoid structurally unbalanced coordination. Nested coordinate structure also should be avoided.

(4) Avoid elliptical sentences.
Elliptical sentences, which are common in Japanese language, often cannot be translated into English without restoration. However, it is difficult to fill in ellipses, as it requires contextual reasoning. Particularly, a sentence of which both the subject and the object are omitted should be avoided.

(5) Do not overproduce compound words.
It is difficult to analyze semantic relationship between words constituting a compound word, as there is not a function word between them. Compound words, especially compound verbs, should not be overproduced.

(6) Write content words in Kanji or Katakana.
Word segmentation of a Japanese sentence is not a completely solved problem. The segmentation algorithm of the system relies on the convention that most content words are written in Kanji or Katakana while most function words are written in Hirakana. A fairly strict restriction should be put on writing content words in Hirakana, although it is allowable in texts for humans.

### 3.2 Pre-editing Conventions

The guidelines shown above is not sufficient for the purpose of making a sentence unambiguous to a machine. In addition to the guidelines, some conventions for pre-editing were introduced.

(1) Structural disambiguation by parenthesizing a phrase.
Pre-editors can restrict sentence structure by parenthesizing a phrase, as a parenthesized phrase implies that the words in the phrase but the head word neither depend on words outside the phrase nor govern words outside the phrase.

e.g. 1 (Disambiguation of dependency structure)

学会で〔得られた結果〕を発表する。

e.g. 2 (Disambiguation of coordinate structure)

〔日本語の特性〕と機械翻訳

e.g. 3 (Disambiguation of word boundary)

高〔価格〕

(2) Phrase not to be translated.
Unnecessary analysis of a proper noun or a formula in a sentence often causes an erroneous translation. This can be avoided by bracketing off a portion which should be dealt with as an unknown word.

(3) Sentence type.
Titles and items must be translated in appropriate phrasal forms other than the ordinary sentential form. This is usually done automatically. However, some sentences are difficult to identify their sentence types. For example, a sentence expressing an instruction, which should be translated into an imperative sentence in English, is confused with an ordinary sentence. In such cases, pre-editors can tell sentence types to the system.

## 4. JAPANESE SENTENCE DIAGNOSIS TOOL FOR HICATS/JE

### 4.1 Purpose and Functions

Pre-editing is a job requiring rare skill. It is difficult for most of the users to select every expression to be pre-edited and decide how to pre-edit. Both oversight and overdoing cannot be avoided. Moreover, the results of translation do not necessarily come up to pre-editor's expectations. For these reasons, most of the users were unwilling to do pre-editing.

In order to fly out of this situation, we developed a tool diagnosing input Japanese sentences and pointing out ambiguities the system cannot resolve. The following are considerations for designing it.

(1) Batch processing or interactive processing?
We adopted a batch processing tool outputting a list of warnings. It has the advantage that a human operator need not be present all the time. However, it has the disadvantage that redundant warnings are inevitably included in the list.

(2) Analysis level.
A diagnosis tool performing only morphological analysis can be economically executed on a small computer or a word processor. However, it does not have much effect on pre-editing, as it is not able to indicate directly parts to be pre-edited. We adopted a tool performing the same syntactic and semantic analysis as the machine translation system itself does.

(3) Dealing with needless warnings.
A diagnosis tool of this kind inherently has the defect that the majority of warnings are actually needless, since the machine translation system chooses the most probable solution even if it cannot resolve an ambiguity.

This defect will be compensated by strategy to point out specifically the solution the system chooses when not pre-edited.

The diagnosis items by the tool are given below.
(a) Sentence length (the number of *bunsetsu's)*
(b) Unknown word.
(c) Ambiguity in word boundaries.
(d) Multiple parts of speech.
(e) Ambiguity in dependency structure.
(f) Ambiguity in coordinate structure.
(g) Ellipsis.

### 4.2 Evaluation

The tool can be evaluated by the ratio of elimination of needless warnings and the ratio of inclusion of necessary warnings, between which a trade-off exists. The following is an evaluation result of structural ambiguity detection. The evaluation was done using 210 sentences from JAPIO(Japan Patent Information Organization)'s patent abstracts.
(i) Ratio of needless warning elimination: $1 - |B| / |A| = 0.50$
(ii) Ratio of necessary warning inclusion: $|B \cap C| / |C| = 0.92$
(iii) Ratio of effective warning: $|B \cap C| / |B| = 0.25$
where A: set of ambiguities a purely syntactic parser would detect,
B: set of ambiguities the tool detects, and
C: set of ambiguities HICATS/JE analyzes incorrectly.
In short, the tool has the ability to detect 92 percent of ambiguities to be pre-edited, while reducing the total number of warnings to half of that of potential ambiguities.

It has been proved that the diagnosis tool improves the pre-editing efficiency of unexperienced users. For instance, a beginner completed pre-editing of ten pages from Japanese Industrial Standard documents in five hours. But more important findings is its usefulness as a training tool. A lot of users have reported that they could acquire pre-editing know-how in a rather short period due to the tool.

## 5. COST-EFFECTIVENESS OF PRE-EDITING: AN EXPERIMENTAL RESULT

An experiment was carried out to evaluate the cost-effectiveness of pre-editing. That is, post-editing costs (time) in the following two cases were measured and compared.
(a) Case-A: Translation by HICATS/JE and Post-editing.
(b) Case-B: Pre-editing, translation by HICATS/JE, and post-editing.

The source texts used were 180 JAPIO's patent abstracts (150 for Case-A and another 30 for Case-B). The average length of an abstract was 361 characters (173 words). The average sentence length excluding titles was 79 characters (39 words).

The pre-editors in Case-B were university graduates having pre-editing practice of one month. They were given 20 minutes an abstract. This is for pre-editing on sheet, and does not include word processing operation. Within this period of time, long sentences were split, the wordings were altered, and pre-editing parentheses were inserted.

The post-editors in both cases were professional translators, but they had little experience in post-editing of machine translated texts. The average times required for post-editing one abstract were as follow;
(a) Case-A: 25.8 minutes.
(b) Case-B: 13.6 minutes.
These are times for post-editing on sheet, and do not include word processing operation. Human translation is guessed to require 30 minutes and more an abstract, although it was not carried out in the experiment.

The experimental result shows that pre-editing is justifiable from the viewpoint of cost-effectiveness. The unit cost of post-editor, who must be a competent translator, is larger than that of pre-editor. Accordingly, pre-editing costs can be sufficiently compensated by reduction of post-editing costs.

## 6. CONCLUSION

The experience reported here proves pre-editing to be cost-effective, and encourages us to further pursue controlled language. If authors originate texts in a controlled language, the cost-effectiveness of machine translation will drastically improve. The development of controlled language is the key to practical use of machine translation, especially machine translation of Japanese language.

Acknowledgments: I am indebted to Mr. Akihiro Hirai for implementing the Japanese sentence diagnosis tool, and to Messrs. Setsuo Shibahara and Akimichi Hanaoka for conducting the translation experiment with patent abstracts.

Reference

[1] H. Kaji: "HICATS/JE: A Japanese-to-English Machine Translation System Based on Semantics", Machine Translation Summit (Hakone, Sept 1987).