A Way of Using a Small MT System in Industry

Yoji Fukumochi
Information Systems Product Development Labs., Sharp Corporation
492, Minosho-cho, Yamatokoriyama, Nara 639-11, JAPAN
e-mail: fuku@isl.nara.sharp.co.jp

1. Introduction

In 1982, Sharp launched research into machine translation, English to Japanese. Since then, the company has commercialized a series of machine translation systems, DUET E/J, E/J-II, J/E, and Qt, and most recently has ported the system onto a business-use Japanese word processor *"SHOIN"* adding a new feature called "phrase-to-phrase translation function." This is to overview the features of DUET Qt and the "English Text Translation Support Software" on the word processor as well as how the small MT systems are used in industry.

2. SYSTEM

DUET Qt is a rule-based, transfer-type MT system which runs on a workstation, providing a total support for translation in the office. Its major features of translation engine module, system dictionary, and translation support module are summarized below.

Tree matching-based prioritizing interpretation

Given a very simple sentence like *"Three and four is seven,"* a MT system might fail in parsing it, because the subject is a coordinate noun phrase consisting of two noun phrases and a conjunction *"and,"* which is normally regarded as plural, thus contradicts with the number feature of the predicate *"is."* Therefore, the system has to loosen such a basic constraint when it encounters such an irregular input. However, once the constraint is loosened, the system may not always be able to output a plausible answer to other cases. Namely, an exceptional case like the above should be treated differently from a normal one in terms of priority. For this reason, modification of the preference based on the features of a given structure is accumulated in heuristic rules for structural preference in interpretation as show below:

```
MATCH:    T:     S(NP,VP)
          COND: T.2.checkParallel
          IF:    T.2.syn = PLURAL
          IF:    T.3.syn = SINGULAR
          THEN:  (a negative score)
```

The procedure "checkParallel" in the COND part is evoked to check whether the NP is a coordinate noun phrase consisting of more than one noun phrase. The conditions defined in the IF parts represent that the node number 2 "NP" is "PLURAL" and that the node number 3 "VP" is "SINGULAR." If the conditions are satisfied, then the system gives the tree structure less priority, without discarding the tree.

Again, given a phrase like *"the volume of the box and weight of the content,"* an MT system outputs four parses caused by syntactically ambiguous nature. Obviously, an interpretation "((A of B), and, (C of D))" is the most plausible parse, in which an MT system uses semantic affinity of *"volume"* and *"weight."* Besides such semantic information, we can use structural preference as heuristics which human translators know from the structurally plausible interpretation. Such heuristic rules are also accumulated in the syntactic preference rules of the system as shown below:

```
MATCH:    T:    NP(NP(NP,PP),CONJ,NP(NP,PP))
          IF:   T.4.1ex = T.8.1ex
          THEN: (a positive score)
```

The above rule shows that an interpretation shown in the described tree structure must be strengthened if the lexical entries of the head of the prepositional phrases 4 and 8 are equal. In syntactic analysis, the system applies rules like the above to the structures built, when applicable. Such processing leads to a single syntactically best parse (which may not be necessarily semantically, contextually best) for an input sentence.

Lexical preference information in basic dictionary

The system uses two types of lexical preference, namely that of syntactic categories of given words and that of verb patterns. The former preference is defined under three types; MAJOR, NORMAL, MINOR, based on the frequencies of use. For example, *"high school"* is MAJOR against *"high"* plus *"school," "can"* as a noun or verb is MINOR against *"can"* as an auxiliary verb. Despite the system using a breadth-first parsing algorithm, the analysis is carried out in two-path mechanism. If the first path fails, stacked candidates defined as NORMAL/MINOR over MAJOR or as MINOR over MAJOR/NORMAL are used for the second path, in order to avoid combinatorial explosion. The latter preference is, on the other hand, defined as according to the frequencies of occurrence in such verb patterns. For example, occurrence of *"do"* as an intransitive verb is less frequent than that as a transitive verb. Such preference affects the accumulated parse score, and thus influences the determination of the best parse. The preferential information described above is concerned with determination of syntactic structure of the input. At the same time, the system uses semantic constraints in case frames to determine the readings of given words. Semantic constraints are usually difficult to describe consistently without risking getting no answer for the case frame. In order to avoid the problem, the system uses a fail-safe mechanism in which the system picks up a default reading, if none of the case frames accepts the input sentence. However, it is still a difficult issue whether we should count on syntactic preference more than semantic preference.

Translation Support Functions

To resolve lexical ambiguity in multiple-reading words, the system employs field codes encoded in the system dictionary, such as chemical engineering, mechanical engineering, economics, current events, information processing, electronics and medicine. Once the fields

of documents to be translated are attributed to a given field, the reading of a particular word is limited to a certain extent. The system outputs a translation matching users' selected fields. However, as above, it is still a problem how much we would prefer the reading which matches the fields even when the reading does not satisfy the governors' case frame semantic constraints, or whether we should prefer the reading to the translation of words defined by users.

Pre-editing Functions

The system supports a variety of pre-editing which are roughly divided into two groups; one that is placed right in front of a particular word to specify its feature, and the other that is placed in between words to specify the range of phrases. For example, the former one is:

- Specifying the part of speech of a given word in a sentence

The latter ones are:

- Marking a given part of sentence to be translated
- Marking parts of sentence to be omitted from translation input
- Inserting a mark for breaks of sentence
- Marking the range of a given phrase
- Marking a given phrase which needs to be output in source language
- Specifying the part of speech and the range of a given phrase

Range-specific Translation Function

Most MT systems hold more than one translation for a single input sentence, as alternatives in parsing, transfer, or generation. However, it is sometimes difficult to get an alternative, especially syntactic one, since just displaying a list of alternative translation output as a surface output sentence as a combination of possible alternative translations is too much for a user. Therefore, the system adopts a function to be able to get an alternative translation for a particular part of the input sentence. After getting the first translation to the input sentence, a user checks the output. If he is not satisfied with the output, he can point to any word in source language or in target language, according to where he wants to change the output. The system automatically searches the minimal packed nodes including the pointed word, and show the range of the input string covered by the packed node. If he is not still satisfied with the range, the system goes on showing a wider range of the input string. Once the range is determined, that is, the packed node where the system gets an alternative, the system searches it and outputs an alternative translation.

We have recently ported the English-Japanese MT software onto a Japanese word processor "SHOIN," and added a new feature which is described below.

English Text Translation Support Software on SHOIN

As to English-to-Japanese MT systems, vast amount of information in English documents available needs to be translated in a rough manner, so that users can skim through the contents very rapidly. The feature comes from a review of current MT systems in use and users' reaction

to its daily use. The former translation system uses input/output text display in a left-right correspondence on the display, which sometimes hinders users from judging quickly whether the output is reasonably good enough or not. If the output is misleading because of any one or combination of errors in analysis, transfer, or generation, users have to revise the output. However, the left-right correspondence is not suitable for this purpose, especially when the output is rather long or complicated because of source and target structure difference as shown below. In order to help users to skim through the contents with easy check of the input/output, we have devised "Phrase-to-phrase Generation" as shown below:

Normal Translation

<<source>>	<<target>>
In later chapters you will find out how you can easily change the program to calculate accurate results when working with calendar date.	後の章において、あなたは、いかにカレンダー日付けを使って作業をしているとき、正確な結果を予測するためにあなたがプログラムを容易に変え得るかを見い出すであろう。

Because the nesting of clauses in source language results in many cases where subjects of the clauses and the predicate are placed far from each other, the results in the above target language are rather complicated. The complex structure leads to

Phrase-to-phrase Translation

In later chapters	you will find out	how you can easily change the program to
後の章において	あなたは見い出すであろう	いかに正確な結果を予測するためにあ
calculate accurate results		when working with calendar date.
なたがプログラムを容易に変え得るかを		カレンダー日付を使って作業をしているとき

The techniques of which features are summarized have made such implementation possible.

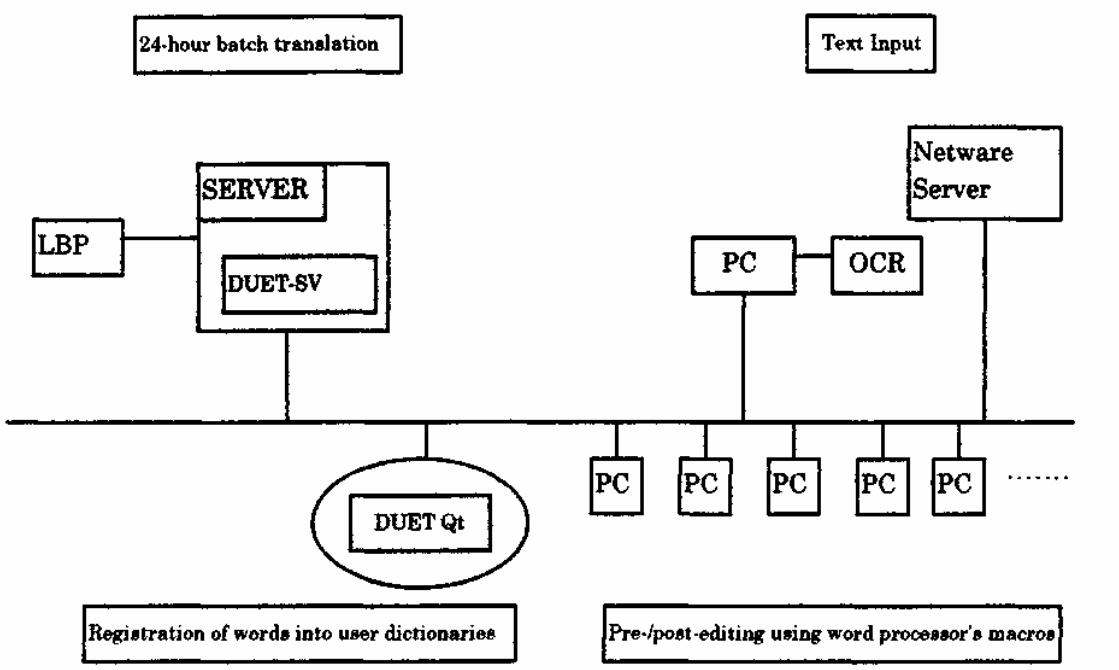
- Executing ordinary syntactic, semantic analysis and transfer in the same way as in normal translation.
- Using rules for finding major breaks in the input sentence, referring to the source analysis structure, and deciding the position of breaks in the sentence referring to the source analyzed structure.
- Referring to the break positions, create correspondences between phrases in SL and phrases in TL.
- Displaying source phrases and target phrases in a manner as shown above, so that users can identify which words of source language phrases is translated into which phrases in target language.

The reason we do not cut a sentence into phrases before parsing is that, if we do so, we cannot utilize the information of relation between phrases for parsing, lexical selection, etc. This translation display mode is effective, when a user skims through the contents of a document in

translated Japanese. They can easily refer to the English original phrase displayed right above, whenever the translation sounds unclear.

3. Case Study

The figure below shows a system employed by Bio Information Center (a Tokyo-based translation company specialized at the translation of medical reports and database abstracts). PCs, OCR, DUET Qt (interactive-type MT system), DUET-SV (server-type MT system) are connected on a TCP/IP and Netware network.



Their translation procedures are briefly described below:

- Text input through an OCR connected to a DOS/V PC.
- Pre-editing the input text using a word processor's macros on PCs, giving the texts limited pre-editing such as a part-of-speech specification inherent to medical documents, etc.
- Pre-edited texts are sent to the Server in the form of mail in which input texts are in the mail body, and specification of translation modes and dictionary selections are specified in the subject field of the mail.
- Monitoring the receipt of the mail and the current spooled translation jobs, the Server translates the text into Japanese, on a 24-hour-working basis. After the translation is finished, the Server automatically sends the translation output back to the PC clients.
- Translators find the output mail from the Server delivered to their mail on PCs. Again they execute post-editing on the output using the macros. After the automatic post-editing, they add some manual post-editing, if necessary.

• Separate from the mail-based input and output, a translator working on the client MT system registers words into user dictionaries and demands dictionary maintenance such as merging the dictionaries on the Server.

The MT at Bio Chemical Center is characterized as follows:

Volume of Translation	50,000 to 100,000 words per day
Size of User Dictionary	approximately 30,000 phrases
Technical Term Dictionary in Use	medicine
Size of Macros for pre-/post-editing	ranging from a macro consisting of several hundred steps to short ones for each client or field
Improved Efficiency	about 50% in terms of time consumed

4. Future Perspective

The Recent development of PC hardware is so remarkable that many PC manufacturers have penetrated into offices and homes with a desk-top, notebook-size personal computers almost comparable to workstations. Owing to the development, there are now around 10 versions of MT software running on personal computers in Japan, and investment in PC hardwares and software costs much less than several years ago. Currently such software costs from less than 100 dollars to around 2,000 dollars. It is obvious that user profiles have changed greatly since even users at home can afford to buy one which used to be limited to professional users several years ago. In Japan, most of the MT systems are limited to E-J or J-E. The E-J MT softwares are getting especially popular, since the Internet connections are beginning to spread into homes. The demand for ability to read documents available on the network in Japanese is getting bigger and bigger. There is no doubt that MT is splitting into two categories in Japanese market:

- PC software at affordable price, with compact memory size and good user-interface which is a requisite for tool software just like front-end processors.
- a professional-use MT system, with an enormous amount of dictionary data and full support of pre- and post-editing functions, etc.

In order to expand the market size, many manufacturers are headed for the former type, in which user-friendliness is the key point, not performance in accuracy. The problem stems from the fact that the current MT systems have an accuracy somewhere between 60 - 70% in raw translation whether it is a big or small-scale system, and it takes many years to improve the accuracy. We have to think about how we incorporate the techniques which we have cultivated in the latter type for many years into the former type and have to wait for another breakthrough for achieving high-performance MT.