Development of an Intranet MT System Adapting to Usage Domain

Nobutoshi HATANAKA

Information & Communication Systems Headquarters
CANON INC.

Abstract

The machine translation (MT) system came out brilliantly. Today, however, the development of the MT system is losing the vigor it once had. The cause of the system's infrequent use became clear as the survey of usage patterns in our company progressed.

The problem was caused by the fact that the results of the MT system was not as expected. This study analyzed the usage patterns and characteristics of the translated documents, including technical documents of the copier business division's service department and specifications or drawings prepared in overseas factories.

The conclusion drawn from the analysis was that any MT system should include an adequate dictionary and the ability to select an appropriate adverb and verb by applying the co-occurrence rule. Furthermore, an MT system should be able to translate fixed form sentences that are used repeatedly.

After the usability of the MT system was improved, the translation staff started using it more frequently at various sections in our company. Moreover, we developed an MT system with the above functions incorporated. Accordingly, the machine-translated documents turned out as expected.

In this paper, I will report on the circumstances of our MT development and discuss the requirements for an MT system.

1 Preface

The MT system made a glorious debut. Today, however, the speed of its diffusion has slowed down. As we continued to study the actual use of the system in our company, the reasons for the system's infrequent use became clear. When these causes were eliminated, various sections of our company started to use this system.

The cause of the infrequent use of the MT system can be assigned to the fact that not enough effort had been made to adapt the system so that the translated documents conform to the expectations of the users.

We began this research by first analyzing the usage patterns and characteristics of the documents to be translated. Then, on the basis of the analysis, we clarified the essential requirements for the MT system. By developing an MT system with the necessary factors added, we tried to adapt the system so that the translated documents conformed to the users' requirements for such documents. In this paper, I will report the whole process of the research and discuss the necessary elements that should be incorporated into an MT system.

2 Aim of introducing the MT system

In 1989, the number of overseas addressees of our business correspondence sent from the Service Department of Canon's Copier Business Division alone were 7,260, including dealers.

The correspondence included overseas mail and other documents transmitted mainly through the use of fax machines. The number of business correspondence was 6,259 letters per year, including documents concerning service information, notices of replacement parts delivered, and quick reports related to the quality of new products. The number of copies sent amounted to 90,217 per year, and the lead time for preparation was from 10 to 24 days. It was a great burden in terms of both the number of days it took and the number of documents sent. When the correspondence-related costs of all the sections were summed up, it became obvious what a considerable burden it was to the company.

When a serious complaint arose in the market, it became necessary to send the information to the development sites and factories in order to identify the cause of the complaint and to address it. However, there were language problems concerning the translation from Japanese to English as well as the major problem of how to curtail the lead time in sending the information.

The translation from the local language into a particular language was done at each site. At the sites located in Europe, the staff there translated documents written in English into German, French, or other languages using MT systems developed for these sites.

Therefore, an urgent task for Canon Headquarters was to introduce an MT system that translated documents from Japanese into English.

The following problems existed at the time of the system's introduction.

1) Ninety percent of the work was to translate materials written in Japanese into English, therefore a system for this type of translation was required.

2) Foreign correspondence had to be sent immediately. Correspondence by mail took 10 to 24 days

3) Correspondence overseas was frequent and the burden in terms of quantity was heavy.

3 Systematic problems of an MT system

In this section, I will begin with typical problems regarding the characteristics of the translations done by the system. Then, I will present the usability problems that surfaced when the translation staff used the MT system.

3.1 Quality of translation results produced by machine translators

On the basis of the issues stated in section 2, 1990 saw the start of the development and establishment of the MT system. At the time it was introduced, problems concerning the use of the MT system were notified by the person specializing in the use of the system. Next, the system was evaluated by panelists taken from the staff who did the translation job at the time when the system was introduced.

When the panelists rated the sentences translated by the MT system, the degree of correctness was rated at 60 %, which was not a bad figure. Of the remaining 40 %, which was rated as defective, 70 % contained grammatical mistakes and 30% had mistakes in wording.

Three years after the establishment of the MT system, the performance of the system was evaluated by native speakers of English. They rated 44 % of the translations done by the system as acceptable and the remaining 56 % as unacceptable for use. They judged that approximately 70 % of the unacceptable translations came from inappropriate words selected by the MT system and 30 % was due to unusual expressions or unintended modifications generated in the sentences.

Our efforts to improve the system by adding more terms to the dictionary and correcting syntax and semantics rules produced positive results during those three years. However, the score given by native English speakers who participated in the experiment was lower than that given by general users three years ago because the native speakers used higher or stricter criteria to evaluate the quality of the translations. The undesirable result made us think that although the system had an extensive vocabulary and syntax, it did not select appropriate English terms for the corresponding Japanese words according to the context of the original Japanese sentences.

It became clear that the sentences were not just the collection of words but words that relate to each other. We realized that these collocations were an important constituent of English sentences, which brought about the development of the co-occurrence rule.

Based on the evaluation results shown above, I came up with the conclusion that the following problems should be solved to improve the MT system.

1) The word order of the translated sentences was not appropriate.

When Japanese sentences with modifiers for declinable words and substantives were translated by the machine, the word order was rearranged and the structure altered. This is called the destruction of modification relationship.

2) Adverbial and adjectival phrases modified inappropriate parts of the translated sentences.

(1) Words or phrases were modified when they should not have been.

-Modification occurred at the beginning or end of sentences or before or after modifiers when it should not have been.

(2) Independent phrases were not generated clearly.

- Independent phrases modifying whole sentences incorrectly
- Independent parentheses inserted between clauses

(3) When there were several clauses or phrases, the relationship between the modifier and the declinable or indeclinable words were altered in the translated sentences.

3) With regard to phraseology, idioms, and idiomatic phrases, the translations done by the system included awkward expressions peculiar to the system, and in some cases the meaning was entirely different.

4) Although the system included a dictionary of words, proper words were not chosen by the system.

In particular, many complaints regarding 3) and 4) above were submitted by professional translators.

3.2 Usability problems with the MT system

When the MT system was introduced in 1989, translation engines could only be run on a host computer or on a UNIX machine, such as the SUN Sparc-LT Therefore, the maintenance of the machine itself was necessary, and specialists were needed to boot up and shut down the system.

Moreover, even when the system was started successfully, the characters displayed on the screen were small. Due to the low operability of the system from the menu-driven screen, users had to operate it by entering appropriate commands when the system was down or affected by such problems as computer bugs. Because the translation staff comprised clerical workers who did not have knowledge of computers, usability problems surfaced frequently. At the site where the MT system was used, panic ensued and intensified users' dislike toward the MT system.

4 Establishment of a knowledge base comprising dictionaries and rules

From the sentences we analyzed, as described in subsection 3.1, and the 8,425 sentences taken from service manuals for copiers, we selected technical terms for the copier business and analyzed its syntax and semantics rules.

As a result, 6,000 technical terms used in the copier business, approximately 2,000 rules of syntax, semantics, and common phrases, and approximately 3,000 co-occurrence rules of verbs were compiled into a knowledge base. We developed the knowledge base so as to be utilized by the fundamental translation engine of the MT system as well as the "turbo engine" installed outside the system.

Of these rules, the syntax and semantics rules were stored as an internal component of the MT system. A function that enables users to register co-occurrences of verbs from external devices was added to the system. Furthermore, we enriched our MT system by preparing a basic dictionary containing 600,000 words and a technical dictionary containing 560,000 words from various fields. (Areas included business, chemistry, metals, information processing, computers, machinery, plants, construction, medical sciences, physics, mathematics, electricity, and electronic machines.)

5 Usage domain and document characteristics of the MT system

As shown in Figure 1, *usage domain* is defined as "the domain where the user utilizes the translation." In other words, it refers to the usage pattern or the purpose of the translated document. Through the analysis introduced in section 4, the usage domains of the documents translated by machines are categorized as follows.

1) Browsing domain

Information is necessary to make a decision. Also, in order to remove the uncertainty that arises during the course of performing a task, you need to distinguish useful, relevant information from useless ones.

At this stage, the need to easily understand the outline of the information arises. When one uses the MT system to browse the outline, the usage pattern is referred to as the browsing domain. Searching the Web to get technical information, patent information, etc., is an example of the usage pattern categorized as the browsing domain.

2) "On the spot correspondence" domain

This refers to the usage pattern in which translated materials are used as a means of two-way communication between the location where the information emerges and the location where the action takes place based on the information. Documents categorized into this domain include letters, facsimiles, and e-mails which demand immediate delivery and whose speed is considered more important than the quality of translation.

3) "Translation for products" domain

This refers to the usage pattern in which the translated documents are included in a product or incorporated as a part of the product. In this domain, the value of the product can be increased by improving the translation. As shown in Figure 1, documents falling into this domain include manuals, operating instructions, and written contracts.

At this point, let me introduce the definition of *document characteristics.* This term refers to "the characteristics or the style of the sentences in the original document to be translated."

In the analysis performed as described in section 4, we studied service manuals for the NP6650 copier. As shown in Table 1, the document characteristics of the manuals can be categorized as (1) operating or maintenance instructions, (2) explanations of the machine, and (3) lists. On the other hand, a characteristic of documents prepared at overseas factories for manufacturing products is that they have no subject or predicate, as in the drawings and specifications.

(1) Operating or maintenance instructions
(2) Explanations of the machine
(3) Lists
(4) Titles and headings
(5) Notes for drawings

The above categories were defined as a result of analyzing a limited range of sentences or phrases, including only those in the drawings and specifications needed at overseas factories and in the technical documents for the copier business that had been analyzed to establish the knowledge base.

Naturally, we cannot apply the results to patent documents, written contracts, or invoices for overseas shipping because thorough analysis of these documents has yet to be performed.

Table 1 Categories of document characteristics

Category	Characteristics	Examples	Solutions
Operating or maintenance instruction	Although the original sentence is an assertive sentence, the translated sentence is an imperative sentence. The length of sentences are short.	原稿台カバーを上げ、 原稿面を下にして サイズ指標に合わせる Open the copyboard cover, and set the document face down aligned with the size index.	Translate it as an imperative sentence. Register the terms in the word dictionary. Enhance the base for the co-occurrence rule of the verb.
Explanations of the machine	The document are composed of logical sentences and the same sentence structure is used repeatedly.	コピー用紙の裏面から プラスコロナをかけ、 用紙の裏面にプラス 電荷をのせ、ドラム表面 のトナーをコピー用紙に 転写する。 In this step, a positive corona is applied to the back of the copy paper so as to attract the negatively charged toner to the paper	Translate the sentence into the passive voice . Register the terms in the word dictionary. Enhance the base for the co-occurrence rule. Provide a function to discriminate between the different ways of interpretation. (Expressions)
Lists	The documents contain tables, and the length of sentences are short. Items are listed in the form of compound nouns,among others. Also, the documents contain phrases and compound nouns.	個所Part | 備考Remarks 防塵ガラス | 濡れ雑巾 Dust-proofing glass | Use a moist cloth 標準白色板 | : Standard white plate | :	Translate using phrases such as "use a moist cloth" or a compound noun, such as "dust-proofing glass". Register compound nouns as adjectives, noun phrases, and adverbial phrases. (In certain cases, register compound noun as fixed form sentences)
Titles and headings	Although the length of the sentence is short, the sentence style is vague.	Original sentence: トナー補給 supplying a toner bottle	Example Base : fixed form of sentences are used. ①Nominalize the expression by use of gerunds.(ex.for titles and headings) supplying a toner bottle. ②Imperative sentences (operating manuals, procedure for repair) supply a toner bottle. ③ Assertive sentences (ex.sentences in tables : expressed in third person, singular form and itemized.) supplies a toner bottle ④The Passive Voice (sentences explaining the copying machine) A toner bottle was supplied.
Notes for drawings	Although the document can be understood in Japanese, 1) There is no part that functions as the nominative case. 2)It is not clear which word is the main verb.	Original sentence: 絞り深さ 100 Pulling from drawing allowable within 100 mm of specified shape.	Short sentences do not provide enough information for the parser to sufficiently analyze and translate them. It is necessary to add more information to the original sentence. -Pulling from drawing allowable within 100 mm of specified shape. Ordinary MT system failed to translate sentence accurately.(The system's parser would translate it as a compound noun as follows.) Iris-diaphragm depth 100

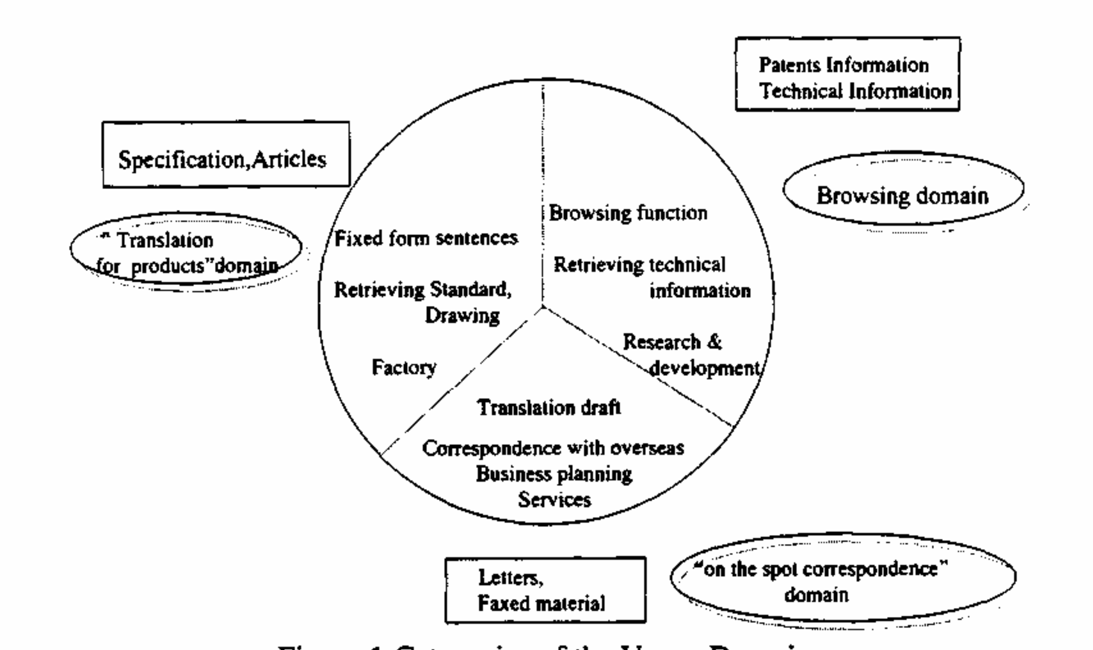

Figure 1 Categories of the Usage Domain

6 Adaptability of the MT system for use

In order to improve the existing MT system or to develop a new MT system, we should consider whether the system could be developed to users' satisfaction.

As I have mentioned in subsection 3.1, our MT system received a score of 60 % for its performance. However, for the quality of the translation to be raised even higher, a tremendous amount of time and investment will be required.

To avoid such investments, we need to find the usage domain that users can be content with regarding the quality of translation performed by the MT system without any further improvement. When such domain is found, the MT system can be adapted for translations done in that domain.

To simplify the discussion of adaptability for use, I set up the following evaluation model.

Domain:

R_i: Usage domain i

Independent variables:

d_i: Element vector for the usage domain i of translations done by machines

c: Vector variable that indicates the document characteristics

Function:

F, G

Dependent variables:

s: Degree of users' adaptability to the system

s is divided into s_1, and s_2 where

s_1 : Adaptability of users to the quality of translation

S_2: Adaptability of users to the usability of the MT system

$$s_1 = F(c,d_i),\ d_i \in R_i \qquad (1)$$

$$s_2 = G(c,d_i),d_i \in R_i \qquad (2)$$

$$s = s_1 * s_2 \qquad (3)$$

The above equation implies that in some usage domains, even if the quality of translation is not good, the user is content with the quality and the translation is as expected.

Furthermore, it implies that if the document to be translated consists of short sentences with structural elements of a sentence intact, the syntax and semantics rules accumulated in the system will easily be utilized. As a result, we can obtain a translation in a form we determine.

The reason why s is expressed as *s_1 multiplied by s_2* is that, as I have mentioned in section 3, if either the quality of the translation or the usability declines, the user will not use the system. In other words, users' contentment with both the translation quality and the system's usability is important in the development and introduction stages of an MT system.

7 Developing an Intranet MT system

This system was developed to improve the translation quality (s_1) and usability (s_2) that are mentioned in section 6. For the improvement of the system's usability, we incorporated a Web browser into the system so that the user will be able to operate the MT system easily. If the user can operate PCs, he or she can easily use the MT system on the server through the use of the Web browser. As I will state in section 8, with the development of the system, the number of people who used the system per day increased to approximately 4 times than that before.

As pointed out in Figure 2, the system was configured as follows. The fundamental translation engines were installed in two SUN Sparc-IIs, one of which was used in development while the other was used in operations. The one for development can be used as a backup machine.

In addition, the parser for the fixed form documents was placed as an auxiliary function of the fundamental engine, and its example base was housed in the Oracle D/B system.

A turbo engine for the system was also developed and installed in the Oracle D/B system so that the user could access the parser via intranet by using the Web browser.

Moreover, the dictionary of technical terms for the copier business, the co-occurrence rules, and the fixed form sentences were accumulated and stored in the knowledge base, which can be utilized through the fundamental engine and the turbo engine. The adaptability to the quality requirement s_1 was increased by this approach.

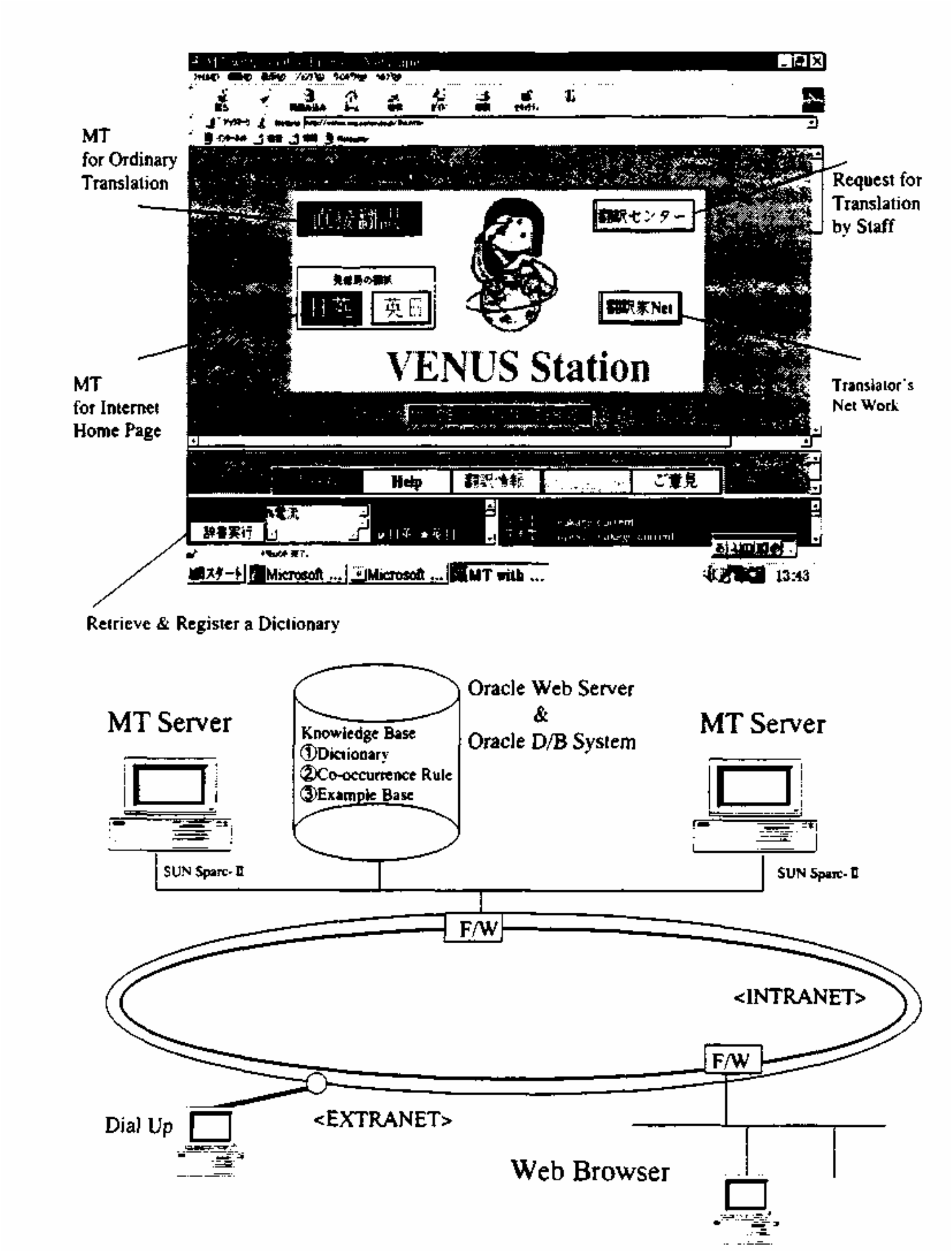

Fig. 2 The configuration of an Intranet MT System

8 Effects / Conclusion

The development of an intranet MT system improved the system’s usability dramatically. As Figure 3 indicates, the frequency of the system's use reached 200 per day on average. There was a rush of requests from various divisions to link their home pages to their intranet. As shown in these examples, the development brought fair success.

The merit of using the MT system via intranet is that users scattered in a huge organization who wish to use the MT system can access it without the barriers of distance and time.

Another advantage is that, because it became possible for those users to use the Oracle D/B system via intranet, the ability to refer to a knowledge base of dictionaries and fixed form sentences (Example Base) was enhanced. The configuration also complemented the system by increasing the usability level, which had been a weak point of the MT system.

The development of an intranet MT system not only increased the frequency of the system's use but also changed the flow of translation work. As a result, the amount of work to be subcontracted to outside workers was reduced, which brought about the effect of curtailing relevant cost.

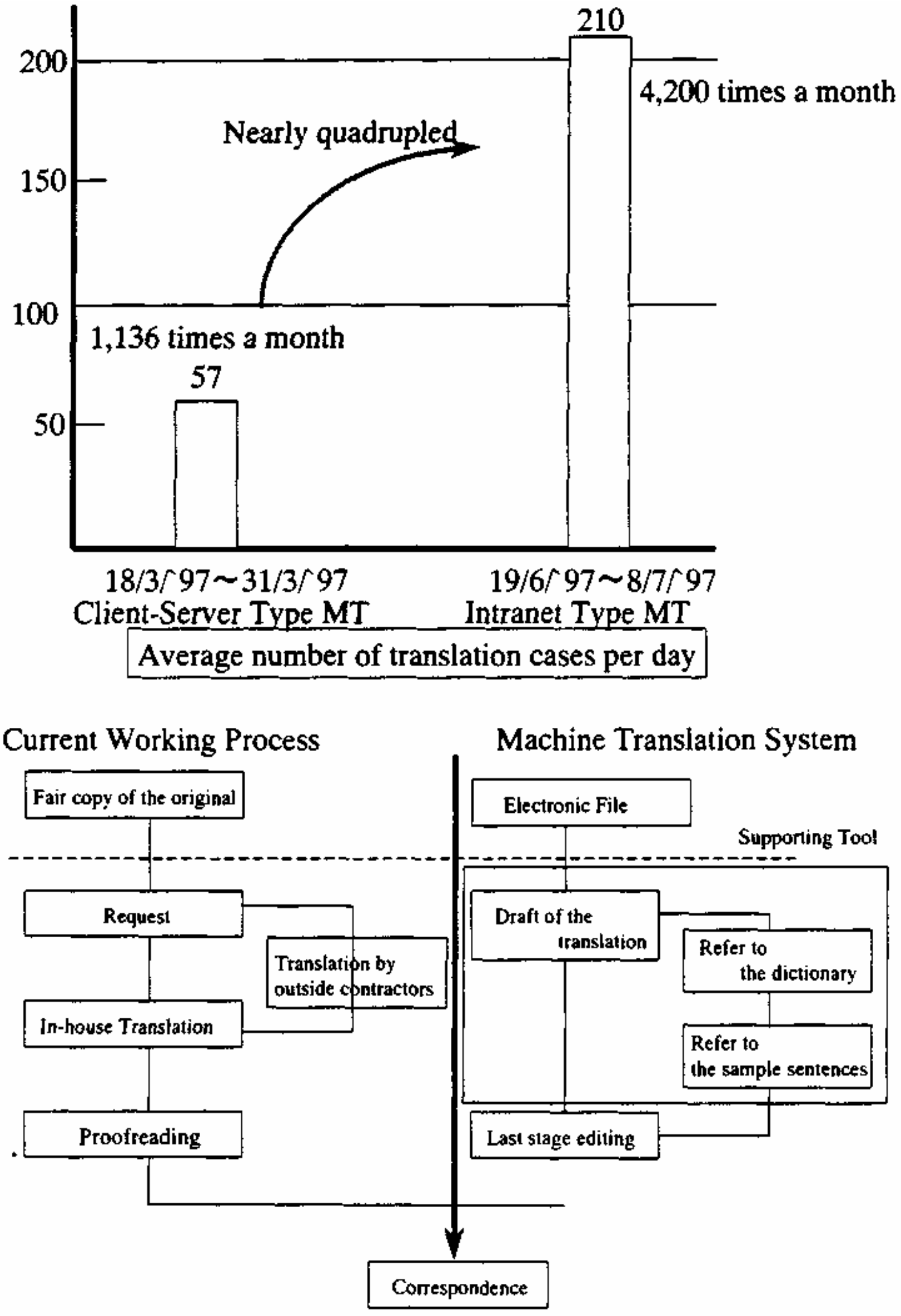

Figure 3 Change in the flow of translation work and the rise in the frequency of MT system use