

Organizational Use of Machine Translation Systems

• ZENSHIRO KAWASAKI

1 Introduction

In recent years, international cooperation in both governmental and private sectors has been expanding rapidly in various fields, especially in industry, finance, and commerce. A key factor for business organizations to initiate and expand their business activities abroad is the ability to prepare documents written in the native languages of their customers for marketing, advertising, sales, and services. Without these documents, business success in foreign countries would by no means be certain. For these companies doing business in overseas markets, translation costs will represent a significant factor.

Researchers, engineers, businessmen, and government officials in various organizations need information from abroad in order to follow trends in their professional fields, but only a fraction of the materials worth distributing internationally are translated. The information barriers are much higher than most of us would imagine. They not only cause people to miss business opportunities, but also hinder international cooperation and mutual understanding. This is potentially a major cause of international friction in various fields.

In these circumstances, translation will be the keyword for the 1990s and beyond. To cope with this vital need for translation, we have to develop new technologies for multilingual communication. The introduction of computerized translation technology, i.e., machine translation, will be one of the most promising solutions to this challenging problem.

In the following, the term “machine translation” will be used in a narrow sense, i.e., automatic full-scale translation performed by a computer, with or without human involvement.

2 Two Uses of Machine Translation Systems and Human Involvement

As is well-known, the current versions of commercial machine translation systems do not provide us with fully automatic, high-quality general-purpose translations. The translation quality of the raw output is generally much lower than that of the average human translation, and peculiar errors are frequent. Therefore, constant human involvement is essential at various stages throughout the translation operation in order to ensure adequate translation quality.

There are basically two kinds of translations for which machine translation is used, regardless of the type of domain or document. The first kind is translations for “information dissemination.” Translations of this type are generally used for publication or mass distribution. The quality requirement is generally high in every respect: accuracy, readability, smoothness, and style. Therefore, substantial revisions or corrections of the machine’s raw output, called “post-editing,” is essential.

The second kind is translations for “information assimilation.” Translations of this type are mostly used for information scanning purposes, e.g., translated scientific and technical papers allow researchers and engineers to follow trends in their respective fields in foreign countries. For this kind of translation, the quality requirement for the finished product may be relaxed. The translations must, however, be at least understandable with a minimum amount of post-editing or preferably without any post-editing, depending on users' diverse quality requirements. Lower quality is acceptable mainly in the interest of speed and timely information acquisition. The translation for information assimilation will be most effectively used in filling potential demand. In other words, translating documents that previously would never have been translated at all. Sometimes this kind of translation is also used to select documents or articles for high-quality translation.

To recapitulate, there are two uses of machine translation systems that have different quality standards for both raw and finished translations. The first of these is translation for information dissemination. The quality standard of the finished translation should be high, which means that substantial post-editing is essential. Therefore, the evaluation of the raw machine translation output will be based on “post-editability,” in other words, manipulability of the output sentences.

The second of these is translation for information assimilation. The quality standard of finished translations may be lowered, which means that post-editing may be at a minimum level or preferably dispensed with altogether. Therefore, the evaluation of the raw machine translation output for this kind of translation will be based on accuracy and understandability at certain levels depending upon the user requirements. This means that the quality standard of the raw machine translation output for information assimilation may not necessarily be lower than that for information dissemination. In some applications, fairly high-quality raw output will be required, especially when no post-editing is expected.

Besides post-editing, human involvement is also needed for building or modifying user dictionaries. This is necessary because the source texts to be translated use a variety of expressions in terms of sentence pattern, style, and terminology for the field concerned, with which the general-purpose basic dictionary and technical term dictionaries provided by the system developer are unable to cope. Therefore, in order to improve the translation quality, the system must be provided with more information for each of the specific fields pursued.

On the other hand, since lexical ambiguities are a major cause of translation failures, the

dictionary data should be limited to a necessary minimum for the current domain of interest. In other words, lexical data irrelevant to the domain should be eliminated or trimmed off before full-scale translation by a computer in order to avoid unnecessary disambiguation operations. As a consequence, “add and delete” is the basic strategy for the dictionary adjustment.

Sometimes, even the grammar rule system, which is another key component of the translation system, must be modified and enhanced. However, direct involvement by the user to update grammar rules is a very sensitive issue. Grammar rules are generally written using a specially designed grammar description language. The coding and registering of grammar rules are not easy tasks for general users. The reason for this is that rule addition and tuning require a substantial knowledge of the total grammar system, the dictionary data, and the idiosyncrasies of the behavior of the machine translation system. Moreover, since interference among rules is generally unavoidable, consistency of the rule system must be maintained in a very sensitive and fragile balance. Therefore, grammar rule modification and enhancement have so far been performed either by system developers or by selected users in close collaboration with the developers.

The procedures mentioned above for adjusting dictionaries and grammar to a specific domain or document are called “customization.” By customizing the system to a specific application domain, we can get much better translations compared to those with only the basic dictionary and grammar rules. Since the post-editing cost is significantly affected by the quality of the raw translations, efficient and effective customization is essential to realize cost-effective machine translation.

3 Planning for Introducing Machine Translation Systems

As noted in the previous section, post-editing and customization are two of the most important types of human involvement for effective use of the current commercial machine translation systems. That is to say, machine translation systems are not ready-to-use systems like word processors; considerable effort and expense are required to bring these systems to a successful operational level. Therefore, organizations that plan to introduce these systems to solve their language problems must be ready to make a sizable commitment to them. For this reason, management staffs planning to introduce machine translation systems must give serious preliminary consideration to how the systems can be used effectively.

The planning phase should include not only the evaluation of translation quality, efficiency, and supporting facilities of the prospective translation systems, but also consideration of how to set up a translation environment and how to organize and manage a translation staff. Some of the issues which must be given a great deal of thought in the planning phase are outlined below.

3.1 Needs Analysis

First of all, the strategic position of machine translation, and the present and future translation needs in the organization should be identified in the early stage of planning. This is because cost-effective use of machine translation systems absolutely requires a continuous demand for a large volume of translation in specific fields and specific languages. The reason for this is that if we compare machine and human translation using a few pages of a random text with the idea of producing high-quality translations of the same quality level, it is almost certain that the machine will have no chance to beat the human in both speed and cost, if the time

and cost for the user dictionary customization and post-editing are taken into account.

However, this does not mean that machine translation is useless. Consider a situation where we have to process thousands of pages of foreign documents every month. Is it possible to hire or arrange for a sufficient number of competent translators with knowledge in the fields concerned? And what about the source texts provided in a machine-readable form which often contain nontextual symbols, i.e., tags for editing and printing operations? How can we process these tags in the input text before translation and then integrate them into the final output translations effectively? The task is simply beyond the capacity of human capabilities. Machine processing of the texts is obviously essential in these cases.

Even in less obvious cases, machine translation will still be effective in the long run. This is because a long-term accumulation of customization efforts will further improve the translation quality, which will eventually make the required post-editing time much shorter. Accordingly, the average operational cost for customization will decrease as the customized dictionaries and grammar rules are used for translating a large volume of documents for a long period of time. In this way, continuing high demand for translation in specific domains is the prime requirement for cost-effective use of machine translation.

3.2 Evaluation

The second issue to be considered in the planning phase is the evaluation of a machine translation system to be introduced. Unfortunately, there are no established criteria as to what constitutes acceptable levels of translation quality and system's capabilities.

With respect to translation quality, evaluation should be made in terms of the accuracy and readability of raw translations of sample texts, using both customized and uncustomized dictionaries. The sample texts should be typical examples of the intended applications. For information dissemination purposes, the post-editability of the output texts should also be evaluated.

The next item to be considered is the system's capabilities. Ease of use is of prime importance for cost-effective use of machine translation. Many means of achieving this have been developed. The most important of these are outlined below.

First of all, there is the facility of dictionary customization, which constitutes specifically designed user interfaces for frequent updating of dictionaries. One aspect of this facility is the possible extent of user involvement in updating dictionary data. In other words, what kind of information the user can put into the user dictionary: nouns and their translations alone, verbs with their syntactic and semantic information, or even co-occurrence relations and idiomatic expressions with semantic restrictions. In general, the more detailed the information is, the more effective the customization is, and consequently, the more the translations can be improved.

Another aspect of evaluating the dictionary customization is the data restriction or data trimming capacity explained earlier. For this purpose, an interactive, semi-automatic, dictionary data adjustment capacity is provided in some systems. This enables only user-selected or user-defined domain-specific usage of multiple-meaning words to be activated for the current document or domain of interest. This is a convenient and effective way to disambiguate word usage in a specific application.

The customization operations generally require substantial knowledge about the specifications of the dictionary data structure and the coding system. Therefore, the trade-off between ease of operation and the extent of the intended improvement reflected in the output translation should also be evaluated.

Editing capacity should also be considered carefully. Besides conventional bilingual word-processing capabilities, easy access to up-to-date lexical and technical information is crucial for post-editing work. Particularly, the availability of electronic bilingual dictionaries for language pairs of current interest is important. Effective treatment of tags, figures, and tables is also important especially for information dissemination purposes.

Since the usability of these customization and editing support features is a key factor in effective use of machine translation, these features should be carefully evaluated by the language staff. The current commercial machine translation systems may not necessarily satisfy all of these conditions we have discussed. Therefore, it is important to evaluate prospective systems carefully to find out whether they meet the organization's needs and can be integrated with the specific translation environment of the organization.

3.3 Design of Total Translation Operations

The next issue, to which the most serious consideration should be given, is how to design the overall translation operations consisting of machine translation systems, various peripheral systems, and translation staff. First of all, let us consider a typical example of both customization and translation processes.

Customization is, as we have seen before, an indispensable preceding step in full-scale translations for a given application. The customization process may be summarized in the following steps. First, a large corpus of sentences from the application domain is prepared. The corpus should include most of the typical expressions as well as word usage in the domain. The size of the corpus is preferably several thousand sentences or more. Then the corpus is analyzed using various tools, such as a word frequency analyzer and an extractor of possible idioms. Following the examination of these statistical data, domain-specific words and idioms with their syntactic and semantic usage, as well as the target language counterparts, are determined. Then, based on the obtained information, the dictionaries are updated so that they will be able to correctly deal with the frequent and typical expressions in the domain.

These dictionary customization steps should be completed prior to full-scale translation of texts in the domain concerned. The registered data in the user dictionary should be continuously modified and expanded in the course of succeeding actual translation operations.

The translation process consists of a series of steps, each of which requires substantial expertise. The first step is the preparation of machine-readable texts. Although the number of source texts available in machine-readable form has recently been increasing consistently, there is still a large number of documents which have to be manually keyed in for machine translation. Therefore, an optical character reader (OCR), which converts printed text into machine-readable form, is desirable, although further refinements in OCR technology are required. Moreover, future OCR technology will have to deal not only with characters but also with graphs, tables, and charts in the text. The prepared machine-readable text is then scanned and analyzed to find any missing entries in the dictionaries and to extract specific expressions and word usages for the domain of current interest. The resulting information is then used to update the user dictionary.

Before the source text is machine-translated, two kinds of operations are generally conducted on the text. One is the automatic preprocessing of nontextual symbols for editing or printing operations. This is necessary especially for already machine-readable texts. These symbols or tags must eventually be restored automatically in the translated target language output. This tag processing combined with machine translation is another advantage over

conventional human translation.

The other possible operation preceding machine translation is "pre-editing." The source text is revised to eliminate lexical or structural ambiguities, or to normalize irregular expressions before being translated by a computer. Portions in the text not needing translation, such as equations, are also marked in this step.

Sometimes pre-editing is a very effective means to obtain successful translation output, particularly when the pre-editor has good empirical knowledge of the peculiarities of the system's performance. However, it would quickly become very time-consuming and costly if structural modifications of sentences or intricate wording changes were required. Therefore, pre-editing should be limited to simple and formal procedures such as notational changes, particularly when the source language is not the native language of the pre-editor.

The translation is then executed by a computer, followed by post-editing. The machine's output is revised by post-editors, in most cases directly on the screen. As noted earlier, post-editing is an indispensable step of machine translation, particularly for information dissemination purposes.

The amount of the post-editor's involvement will depend on the degree of accuracy and stylistic refinement required for the final translation. Since the cost-effectiveness of machine translation is largely affected by the post-editor's ability and efficiency, and post-editing is quite different from conventional revisions, the post editor should be trained specifically for the work. There are already several post-editor training schools in Tokyo. One such school program costs about \$1,600 for two two-hour lessons a week over a six-month period.

Sometimes, particularly when publication quality is required, the post-edited texts are further inspected and modified by technical writers. They check mostly style and substance aspects such as naturalness and preciseness of the expressions. The technical writer should be familiar with the subject matter and preferably be a professional translator native in the target language.

The final step of the translation process is the preparation of the text for printout, including the tag restoration explained earlier. Feedback to the user dictionaries, i.e., dictionary updating based on the current post-editing knowledge for later translations, completes the translation cycle. The accumulation of this feedback provided by the post-editor is the most effective source of information for upgrading the translation quality.

An information network connecting the collaborating language staff and information resources in both in-house and outside facilities is important, and it should be designed in accordance with the existing corporate information system.

The design of the whole translation operation described above is a key issue which must be planned and evaluated carefully, long before the actual operation starts. Since there are few really successful cases of practical use of machine translation at this point, it is important for us to think creatively in designing and developing a translation framework most suitable for our own needs and environment.

3.4 Human Factors

The last but probably the most delicate issue to be considered in the planning phase is the human factor. Like other computer applications introduced into an existing organization environment, machine translation will have a considerable impact on the daily work of the people involved. If management fails to obtain cooperation from the translation staff in restructuring existing language operations, introducing machine translation is not likely to succeed. Failure of past machine translation trials can be attributed in part to the inadequate

consideration of this human factor. Therefore, it is critical for the language staff to be closely involved in the entire planning phase.

Another point worth commenting on in this respect is the role assigned to the language staff. Although the type of human involvement in each translation step varies and different expertise is required for each type, excessive specialization on the part of language staff, such as pre-editor, post-editor, terminologist, lexicographer, and linguist, is not desirable, especially in the early stage of machine translation use. This is because every language staff member should be well acquainted with the characteristics of the system behavior, so that he or she can deal with various linguistic aspects of a lexical, grammatical, and stylistic nature taking into account the system's performance in its entirety.

4 Future Prospects

If the above-mentioned planning phase is properly implemented, introduction of machine translation technology will bring about a noticeable improvement in the productivity of the language staff, particularly those involved in scientific, technical, and administrative translation.

We can also anticipate that machine translation combined with office and document processing operations will fully integrate language services, and that these integrated services will become indispensable organization activities. In some organizations, the language services will play a vital role in their global operations by the beginning of the 21st century.

We must even be prepared for the emergence of a new industry, which might be called the language industry. In other words, an integrated industry of language processing and multimedia document production and distribution. Some translation companies are already moving in this direction, and there can be little doubt that machine translation will be a key component of this industry's infrastructure.

The primary topic of this paper is with the organizational use of machine translation systems. This does not mean that I have neglected the impact of machine translation on potential private users, such as free-lance translators, secretaries, and researchers, although the system requirements and their uses may be quite different. Therefore, we need a variety of translation systems which can be used for diverse purposes. To cope with this future trend, let us take a brief look at a few issues deserving attention.

The first point is that machine translation research and development should be pursued not only from the point of views of linguistics or computer science, but also from a broader perspective, such as the fields of cognitive science and psychology for instance. I can even see that a focused effort on machine translation would have an impact on the emergence of a new integrated research field, let me call it "language technology," which aims to find ways to solve various language-related problems in developing new communication technologies for a more advanced information-based society. This term, language technology, was first used by Professor M. Nagao of Kyoto University as the title of his book published in Japan in 1983.

Another issue which I want to refer to is the research and development of a new generation machine translation system based on artificial intelligence technologies. In addition, a combination of machine translation and speech recognition and synthesis will also be an exciting and challenging technology for developing future multilingual interpretation systems.

5 Concluding Observations

As we have seen, machine translation is not a fully developed technology. Although a considerable number of machine translation systems are now in operation translating technical and administrative documents worldwide, machine translation technology is still in its infancy viewed from a long-term perspective. It is a risky but challenging field requiring active support and participation from various sectors including academic, industrial, and governmental circles.

Since the development of a machine translation system demands a considerable amount of resources to build and maintain software and large-scale knowledge bases including dictionary data and grammar rule sets, inter-institutional or international cooperation is essential. These cooperative efforts on machine translation should also be pursued as part of a worldwide project to establish a technological infrastructure for future global multilingual communications networks.