What Can MT Do for Multilingualism on the Net?

Toru Nishigaki
Professor, Institute of Social Science
University of Tokyo, JAPAN

Abstract

The recent rapid spread of the Internet suggests that soon everybody on earth will be able to communicate freely with each other across national borders. The 21st century will be the age of multilingualism and multiculturalism, when various languages and cultures are dynamically exchanged on a global scale. One may call it the New Great Age of Translation. In such an age, Machine Translation (MT) is obviously expected to play an essential role. What kind of technological efforts, then, are required in the 21st century? A new type of MT product will be sought, which is different from the application products for traditional translation or real-time interpretation. This paper first clarifies the coming need, and introduces an interactive and evolutionary MT technology. Secondly, it addresses Han characters (Kanji) which have distinct origin and functions from alphabets. As highly developed visual symbols, Han characters are expected to promote Internet communication.

1 Introduction: Languages on the Internet

The use of new global media, especially that of the Internet, will certainly have enormous impact on the world's language situation in the 21st century. Generally speaking, any new media can change languages, with the typical example found in Western Europe over the past few hundreds years since print media appeared. It is well known that print media played an important role in the establishment of European national languages. Had it not been for those national languages, there would be no modern nation-states. Thousands of people were tied together and integrated into a community, through sharing experiences every morning by reading the same newspapers printed in the national language. It was really an *imagined community,* as political scientist Benedict Anderson rightly indicated [1].

One of the major merits of the nation-state is dense and intimate communication among members of the state. The compilation of dictionaries, the establishment of a grammar and the elementary education system, together with print media, have made *standard national language* come into being, on which the high level activities of commerce and industry in modern society have been achieved. On the other hand, the demerit of the nation-state is the difficulty of cross-border communication. People find it difficult to communicate in any language other than their own standard national language. In usual nation-states it is not easy to publish books or magazines in foreign languages, because the cost of printing requires a large readership. Japan is typical of those nation-states, where even English texts have few chances to be published. So people in Japan tend to live in a small closed world of Japanese communication. Accordingly, the level of cross-border communication remains unsatisfactory, even though it is eagerly encouraged by the current trend of globalization.

Multilingualism on the Internet, however, is expected to break down this linguistic blockade and usher in various kinds of international communication. With just the pertinent character fonts based on the international common character code installed in a personal computer, foreign texts can flow into our office and home all the time from all over the world through the Internet. Our chances to read and/or write foreign texts increase drastically, which brings about high motivation among people to study foreign languages. Multilingualism on the Internet aims to encourage different peoples on earth to study foreign languages with each other, thus promoting true cross-cultural communication. It is expected to achieve deeper mutual understanding than English monolingualism which utilizes only English for cross-border communication.

It should be noted that until a few years ago the language for international communication on the Internet had been virtually limited to English. This is because the Windows system - the de facto operating system for persona] computers - of non-English speaking countries was basically a bilingual system where one could use only the official language and English. For example, one could use Japanese and English on the Japanese Windows system, French and English on the French Windows system, Chinese and English on the Chinese Windows system, etc. Therefore a user in Tokyo could not send Japanese texts to his Japanese friend living in Paris, nor could he read Chinese web pages even though he can understand Chinese. The character code systems were completely different between Japan and China, where different codes are allocated for the same Han character (Kanji), resulting in the failure of text exchange between the two countries.

The situation mentioned above has been changing rapidly these past few years. In 1993 the International Standardization Organization (ISO) officially recognized *Unicode* as the basic part of the Universal multiple-octet coded Character Set (UCS) which is called ISO-10646-1. This means that Unicode, although there remain some problems requiring further improvements, has in fact become an international common character code for electronic communication. Since there are already several Unicode-based software packages working on Windows system (web browsers, word processors, mailers, etc.), it is expected that soon we will be able to have a multilingual information processing environment on the Internet. That is, we will be able to exchange various texts in diverse languages freely via the Internet in the 21^{st} century. This naturally leads us to the next set of problems - how to assist users' understanding of foreign texts, which is directly related to Machine Translation (MT) technology.

In short, MT technology promotes multilingualism on the Internet. This paper first addresses the current practical needs for MT technology. Then it describes the feature of Han character (Kanji) that is completely different from alphabets, and attempts to give some insights into its role for the future Internet communication.

2 Interactive and Evolutionary MT Technology

There has always been great enthusiasm for MT technology among people in general, since the invention of the computer in the middle of the 20^{th} century. The ability of automatic translation of foreign texts was thought to be an essential skill of an almighty robot who could think and act as an human being. Automatic translation based on natural language understanding has been a dream of the general public, although the dream has also been related to practical needs of modern world where too many different languages are spoken. It is well known that enormous amount of effort has been made on MT by engineers and researchers over the past few decades. However, I dare say that people in general tend to see the results as insufficient, even a failure. They can buy and test many MT products on their personal computers, but the translated texts are often far from complete, including numerous errors which they must spend much time to fix by themselves. Nobody believes that current MT products could translate novels and replace human professional translators. Although a difference in translation quality can be expected if one considers how difficult it is for computers to grasp contexts and situations, quite a few MT product users do not hesitate to express their disappointment.

In this paper, however, I would like to emphasize that now it is high time for engineers as well as users to drastically change their perspective. It may be ideal to dream of a computer capable of *understanding* human language, but there is the risk that the difficulty to realize the dream may prevent the sound development of related technologies. What is important now is, to clarify and meet the concrete needs of MT in the Internet era. We must scrutinize whether or not MT technology so far has neglected the efforts to provide users with truly practical functions, in search of the blue bird of making a computer understand human language. The output of MT is very often only one sentence, which can hardly be edited by users. This brings about the reputation that MT often produces too many mistakes. The reputation is strange, however, if one regards MT as a tool to help users, with which they approach to final solution. A beginner of foreign language study tries to understand a text using a dictionary and a grammar book. It is very reasonable, if one considers a modern personal computer with large memory, that a beginner would make good use of a computer - because it can have an enormous amount of knowledge of vocabulary and grammar stored in its memory. A beginner could refer to the knowledge on demand. Such a tool is expected to promote significantly international communication.

Generally speaking, there have been two kinds of translation. The first one is *traditional translation* in which one translates a large quantity of foreign texts, often taking as long as several months or more. A typical example of this is the translation of novels and academic documents, where subtle nuance and/or strict logic must not be ignored. The translation results are usually preserved for a long time. The second one is the *real-time interpretation* of spoken language. A typical example is simultaneous interpretation in an international conference. The gist must be transmitted accurately in quite a short time, although the translations are usually not preserved. To date MT has been applied to both of these, each with its own specific difficulty. Most novels and/or academic documents contain long sentences with special technical terms, which are naturally hard to translate. Real-time interpretation, on the other hand, must translate in a few seconds many incomplete spoken sentences that contain colloquial expressions.

In the Internet era, however, there will appear quite a new type of translation need. The translation required for cross-border Internet communication is in the midway between traditional translation and real-time interpretation in many aspects. The sentences in web pages and e-mails are usually shorter than those in traditional translation but longer than those in real-time interpretation. Most of the sentences are grammatically complete but their styles are less formal than those in traditional translation. The time limit of translation is probably around a few days or a week - which is also between the two. Moreover the required level of accuracy is lower than that for traditional translation but higher than that for real-time interpretation. Sometimes, however, the accuracy level may be almost the same or even lower than that for real-time interpretation. This is due to, for example, the case where only the translation of keywords is necessary to browse foreign web pages. These features of Internet communication all make the application of MT very promising. Another important point is that the users of Internet communication are ordinary people - not professional translators or international conference attendees - and therefore there is likely to be a significant market for MT products.

Two properties are desirable in the MT products for Internet communication. The first is *interactivity.* The MT products must allow users to freely edit and/or modify their translation output. It would be better to display on a screen several translation examples, and allow a user to select the best one. In addition, such an interactive MT should have the function to pick up and translate only keywords of inputted texts, thus making it easy for a user to decide whether or not the text is worth full translation. Usually it is likely that only a small part of the texts flowing in via the Internet is worth reading in detail.

The second property is the capability of *evolution.* No MT products should be offered to users in a final form. When a user buys the product, it should look rather like a kind of translation-support kit, namely a simple assembly of basic parts - dictionaries, sentence structure analyzers, grammatical rules, etc. Users must make their MT products evolve and develop in such a way that it becomes more and more useful for them through their own utilization efforts. Therefore the function to add user-defined words and idioms to the dictionary of MT is indispensable. Such a function is useful not only for reading but also for writing foreign texts. If users register frequently used sentence examples, the efficiency of their writing foreign texts cannot fail to improve.

One feature that which plays an important role in such interactive and evolutionary MT products is the development of various kinds of thesauruses describing synonyms, antonyms, super-concepts and sub-concepts. Users refer to the thesauruses in addition to the dictionaries of MT products to aid their understanding and composition of foreign texts. The use of MT products mentioned above is expected to contribute greatly to cross-border communication on the Internet.

3 The Role of Han Characters in Internet Communication

The theory of MT has mostly been influenced by Western linguistics, and often it has taken Occidental (Indo-European) languages as a model. Therefore many MT products try to analyze the structure of a given sentence by paying great attention to its grammatical elements. In Occidental languages the syntax of a sentence can be comparatively well analyzed, as it has explicit grammatical elements of declension, conjugation, etc. The famous Case Grammar of C. Fillmore [2], which is widely used in many MT products, is typical of this syntax-dependent approach. Nobody could deny the importance of syntax analysis in MT. Nevertheless, in order to develop an interactive and evolutionary MT product, it may also be important to take an Oriental language - especially Chinese - as a model. There are few explicit grammatical elements in a Chinese text, which is a string of invariable Han characters (Kanji) each conveying meaning as an ideogram. Such a string of ideograms is considered a less syntax-dependent way of expressing human ideas. There is a possibility that Han characters give a useful insight into the development of future MT.

Historically, the characters on earth can be divided into two groups. The first group is the alphabet group, including most of major character sets like Latin, Cyrillic, Greek, Arabic, Devanagari, etc. They are all phonograms, and the descendants of Hieroglyphics, which originated about 5,000 years ago in Sumer and Egypt. The second group is the Han group, which originated about 3.400 years ago in ancient China. Only Han and Japanese Kana belong to this group, and all the rests have disappeared. Because of its different origin, quantity, and difficulty in writing, Han characters have always been criticized and attacked in the name of an *obsolete writing system* which prevents modernization. Famous Chinese leaders like Lu Xun (Ro Jin) and Mao Zedong (Mo Takuto) insisted on the abolishment of Han characters. They thought that China would not be able to be successfully modernized with Han characters, which need a long time to memorize and thus possibly hinder democratic education. Although total abolishment was impossible, Chinese leaders succeeded in creating the simplified Han characters and having people use them.

The situation has been almost the same in Japan. Since the Meiji era, many people insisted on the abolishment of Han characters in Japanese writing. In the 1880s, for example, Kana no Kai (The Association of Kana writing) and Roma-ji Kai (The Association of Latin alphabet writing) were organized, which advocated the writing of Japanese with only Kana and the Latin alphabet, respectively. Every effort to improve Han characters has aimed at the gradual simplification and decrease of Han characters in ordinary use. In short, one may say that during the process of modernization the goal in both China and Japan has been the replacement of Han characters with phonograms.

It is important to point out that this was neither a problem of efficiency nor modern ideology. There has been an academic and linguistic view which asserts the superiority of phonograms to ideograms. The Western linguistics founded by Ferdinand de Saussure basically concerns *spoken* languages. A text sentence is simply a written form of speech, and a character is nothing but a symbol to express the sound of a spoken word. Therefore the desirable character should be one which can represent the pronounced sound in a manner as simple and rational as possible. Naturally this brings about the evaluation that Latin alphabet is the best because it represents a phoneme, as followed by Kana representing a syllable, and then by Hebrew representing a consonant.

From this linguistic view, Han characters are too irrational as a written form of Chinese (Japanese or Korean) speech. It is sometimes hard to reproduce the sound by Han characters, where there is no simple rule and even the same character has several readings. This leads to the idea that Han characters are obsolete for writing and are doomed to disappear in the near future. Many Japanese intellectuals thought this way, among whom is famous rationalist Tadao Umesao who published a Japanese journal in Roma-ji (Latin alphabet), and favored a Kana typewriter.

All these arguments come from Western linguistics, the basic premise of which is that the world can be analyzed and described by spoken words. Namely, it is audio symbols that make the difference in the world. However, one must know that not only audio symbols but other symbols also can make the difference. By observing the communication of auditorily handicapped people, we see that gesture symbols can analyze and describe the world. Then why not visual symbols? As a matter of fact, we are familiar with many visual symbols for transportation control on highways.

We can assume the following two processes in order to invent successful visual symbols: First establish visual symbols that can describe the world. One communicates with others by writing or showing those visual symbols. Even if they speak totally different languages, they can communicate with each other if the visual symbols are well defined. They may read the symbol in different ways, according to their familiar spoken languages. Secondly, define *common way of reading* for each visual symbol. Surely it will make the communication much more efficient, because now one need not write or show the symbol any more, and one need just say its common way of reading. It looks as if they are speaking in common language, but actually they are exchanging visual symbols by saying their *names* - common ways of reading. It should be noted that in this case, the sounds of characters are used for recalling the already-defined visual symbols.

In order for these processes to happen, centralization of power is indispensable in a society. Spoken language can be born more or less naturally. But the creation of a set of complex visual symbols, together with their uniform ways of reading, inevitably requires centralized governmental force supported by high level bureaucracy.

A Japanese historian named Hidehiro Okada, who specializes in East Asian history, asserts that these processes actually happened in ancient China when the first emperor established the Chinese Empire in the 3rd century B.C. [3]. Until then there had been diverse spoken languages as well as great many different characters in the vast land. One may suppose that the way of reading characters had not been common from one area to another, either. The emperor ordered his ministers to select 3,300 characters, and acknowledged them as *official Han characters* to be used in formal communication within the Chinese Empire. Moreover, *official ways of reading them,* all in one syllable, were also determined for each of those official characters. This is the argument by Okada about how the basis of Chinese communication was founded.

There should naturally be some difference, according to this argument, between spoken language and written language. At least, the sentence composed of Han characters is not necessarily the written form of spoken Chinese. This is completely opposed to the theory of Western linguistics.

Some people believe that almost everybody speaks uniform Chinese in China, but its language situation is not that simple. Of course, there is *standard Chinese,* which is called *Hanyu* or *Putong hua* - an official spoken language acknowledged by the Chinese government. It was established as late as in the 20th century based on the spoken language in Beijing. But historically, there have been so many different languages in China, and even now the people in Shanghai and Canton each speak quite different languages from Putong hua.

Okada argues that those southern languages are not the dialects of standard Chinese but languages to be classified into other linguistic categories. According to him, the reason that those southern languages are regarded as the dialects of standard Chinese is that they have large common vocabulary with standard Chinese. The communication in China is based on this sharing of vocabulary, and a large part of which has been inherited from old Chinese classic texts. Therefore one may say that it is the authorized texts in Han characters and their official ways of reading, rather than common spoken language, that is essential to Chinese communication.

We must note here the special social system that enabled the sharing of vocabulary throughout China: the well-known examination system for the appointment of governmental officials termed *Keju (Kakyo),* which continued more than one thousand years. In order to pass the examination, a long preparation of memorizing and writing of Chinese classic texts was indispensable. Many young intelligent students throughout China attempted this central examination, thus resulting in the transmission of the authorized vocabulary to many local languages with sufficient accuracy.

There was no Keju system in Japan, but nobody could deny that Japanese language absorbed a vast amount of Chinese concepts by importing from the Chinese classics. The situation was almost the same in Korea too. In Japanese and Korean texts, one can find many concepts which are written in Han characters and inherited from the Chinese classics, although they are read in different ways. In short, all the East Asian languages have been strongly influenced by the Chinese classic texts. Okada asserts. “If you look on Cantonese as a dialect of standard Chinese, then Japanese and Korean can be looked on as dialects of it, too” [4].

The argument of Okada may not necessarily be widely accepted in the academic circle of historians. Nevertheless his argument is considered to include some enlightening points. What is most important is that a Han character is itself a kind of *translation* among different languages from the beginning. It was organized and authorized by the first Chinese emperor, aiming at the establishment of communication over large empire. That is, the original purpose of official Han characters was the connection among many different languages. As Saussure indicated, each spoken language describes the world in a different way. Therefore the connection among different languages must be a kind of translation.

As mentioned earlier, a Han character can be read in different ways. In Chinese. Korean and Japanese, the same Han character is pronounced differently but its meaning is *almost* the same in many cases. So one may say that there is a possibility to extend this strategy to other languages - not only East Asian languages but also Western languages. For example, the Han character representing “light” is read “hikari” in Japanese, and can also be read “light” in English and “lumière” in French, etc. If some English keywords are changed to the corresponding Han characters in an English text, it will greatly help Chinese or Japanese people (anybody who can understand Han characters) grasp the general meaning of the original text. Naturally it is far from accurate translation, but one will find it very useful when one needs to read a lot of English documents on the Internet.

Generally speaking, a Han character is hard to write but not so hard to read. And one can use a computer to write a Han character by keying-in a string of alphabetic characters corresponding to its reading. Such an input system is already satisfactory for Japanese and Chinese writing.

An important point is that a Han character is a kind of icon - a highly developed visual symbol. One sees various icons on a screen of a computer with graphical user interface, but most of them are not easy to recognize in terms of their functions. For example, there is an icon for unnecessary files, which is usually displayed using a picture of a trash can. But sometimes it is too small to distinguish what type of can it is. It would be much better to replace it with an icon which shows the Han character representing “trash”. No visual symbols on earth have higher ability of representing abstract concepts than Han characters. A Japanese businessman can process documents very efficiently without special fast-reading training, because in Japanese documents crucial points are usually expressed in Han characters and inserted into Kana strings. Namely Han characters work as icons of keywords in Japanese documents. And if this icon system is applied to many other languages on earth, we can expect it to promote communication on the Internet.

4 Conclusion : For the New Great Age of Translation

Communication in the 21st century is predicted to become a global one. which is quite different from the current nation-based communication. It is the Internet that opens the way for new global communication. If only English is used on the Internet, the people in non-English speaking countries will find it difficult to participate in international communication. The goal of multilingualism on the Internet is perfect global communication where everybody on earth can utilize the Internet on equal footing. That is, it aims at interconnecting various peoples with each other, while preserving the plurality of culture and language.

Recent technological development has almost realized the information processing environment for the exchange of various characters on the Internet. But foreign language study in general requires a lot of time and effort, and the language barrier is not easy to overcome. Unless this Babel-situation is improved, most people will be locked up in the world of their mother tongue, with only the elites being able to manipulate English freely and monopolize Internet resources. The economical and cultural gap between the elites and ordinary people will cause social instability.

Therefore everybody expects of MT to solve this tough problem. In order for MT to meet this expectation, all of us - not only researchers and engineers but also general users - must change our attitudes. We must have a clear realistic vision, rather than seeking an omnipotent robot who understands human language. An interactive and evolutionary MT is expected to shed some light on what direction we must go. In the 21st century, it is said that more than a billion of people will come to use the Internet. And if they begin understanding foreign texts with the aid of MT products, it will change history.

In short, t*he New Great Age of Translation* is just around the corner. There have already been several such ages - in Renaissance era for example, a lot of Greek and Arabic documents were translated into Latin in Italy. This translation is considered to have set the base of the modern civilization. Then the modernization of Asian countries was brought about by another great age of translation - a lot of Western documents were translated into Asian languages in the 20th century. The development of MT technology in Japan is also a part of this cultural metamorphosis. Accordingly its research was done mostly based on Western linguistics and Western information processing theory. From now on, however, Oriental civilization is also required to contribute to the development of MT technology.

The Han character is very suggestive in this respect. It has a completely distinct origin from Western characters, and has long been used as a communication tool among various peoples speaking diverse tongues. Hence it is expected to play an important role in international communication on the Internet. As it is a highly developed visual symbol system, each Han character can be used as an icon, and works as a keyword when inserted into a text of phonograms. In the 21st century one may expect a new type of MT which is founded on Oriental civilization.

References

[1] Anderson. B., *Imagined communities.* London. Verso. 1983.
[2] Fillmore, C., The case for case, in E.Bach and R.Harms (eds.). *Universals in linguistic theory.* New York. Holt Rinehart and Winston. 1968.
[3] Okada, H., Kanji culture and McLuhan. *Dai-Kokai,* vol.17, 1997, pp.108-115.
[4] ibid. p.114.