4 Multi-Lingual Translation Technology

4-1 Special-Purpose System for Multi-Lingual High-Quality Translation

SUMITA Eiichiro

NICT is conducting research for realizing high-quality automatic translation system while restricting the domain of translation. We've been concentrated on travel conversation in speech translation and explanation of products in text translation, and recently we put our technology on a commercial basis. In this paper, we outline the technology.

Keywords
Automatic translation, Speech translation, Text translation, Corpus-based translation

1 High-quality automatic translation

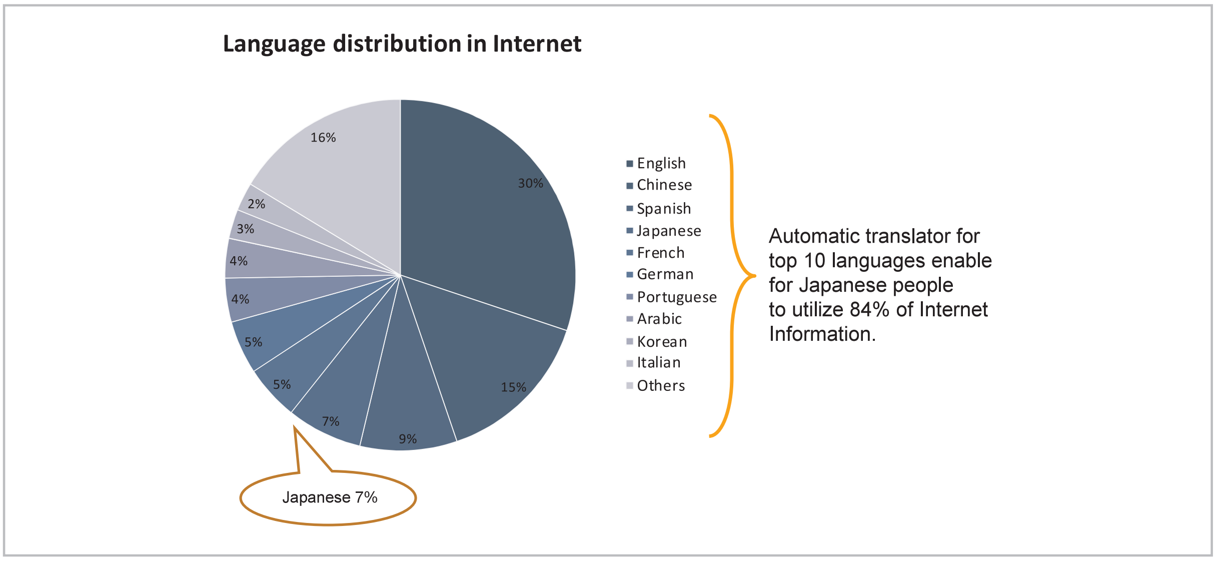
NICT has conducted research and development into translation technologies to achieve a high-quality automatic translation system, with a focus on special-purpose translation. On the other hand, others have conducted research and development with an aim to develop a general-purpose translation system. The former development corresponds to making a fish knife and the latter to making a versatile knife. A versatile knife can cut fish, meat, vegetables, and everything, but cannot cut them cleanly. If we slice raw fish with the versatile knife, the cut surface would not be clean. Similarly, currently-available Japanese-English translation systems are versatile but their translation quality is not very high. As a result, people have an impression that automatic translation systems are useless.

For speech translation ("Speech Translation Technology in MASTAR Project", Chapter **7-1**) we focused on conversation for traveling, and for text translation we focused on explanatory sentences about products in e-commerce. We actually succeeded in establishing a new business of automatic translation. In this Chapter we give an overview of the technology and links to related Chapters.

2 Automatic translation of multiple languages

Language is one of the highest barriers for communication among people. Automatic translation is expected as an ultimate tool to overcome this barrier.

For example, owing to the popularization of search engines, we can easily access global information from home. However, information written in foreign language looks like ciphers to many Japanese people, and few people can use the information. According to a survey of the languages used on the Internet, the top ten languages (English, Chinese, Spanish, Japanese, French, German, Portuguese, Arabic, Korean, Italian) make up 84% (See Fig. 1). However Japanese only makes up 7%. Therefore, if we have a high-quality automatic translation system for the top ten languages other than Japanese to Japanese, we could understand 84% of the information on the Internet, which would significantly enhance Japanese people's ability to collect information. If we



Automatic translator for top 10 languages enable for Japanese people to utilize 84% of Internet Information.

Fig.1 *Percentage of multi-lingual information on the Internet*

have a high-quality automatic translation system from Japanese to these nine languages, on the other hand, Japanese people's ability to transmit information would be enhanced remarkably.

Then, what should we do to achieve this? These ten languages are largely different in characters, words, and grammar. It is therefore necessary to develop a high-quality automatic translation technology that does not depend strongly on the characteristics of the languages.

3 Corpus-based translation technology

In this section, we describe our corpus-based translation technology, which is used to achieve a multi-lingual high-quality automatic translation system and solve the problem mentioned above.

Corpus-based translation technology is technology that automatically creates knowledge databases for translation systems (knowledge database for a translation dictionary with occurrence probabilities, necessary for translation systems), on the basis of the bilingual corpus (collection of pairs of original sentences and translated sentences) (Fig. 2). With this automatic creation, there are two advantages. (A) In the development of a translation system in a new field, high-quality can be reached by collecting of that field. If we collect parallel corpora from, for example, newspapers, patents, instruction manuals, information released from local governments, medical and nursery information, Web, blogs, or other fields, we can develop a high-quality translation system dedicated to that corresponding field. In fact, we focused on explanatory documents about products in e-commerce and succeeded in commercializing the translation technology. Conversation for travelling, as mentioned above, is also an example of a specialized field. (B) If we use multi-lingual translation corpus for N languages, we can automatically develop N (N-1) translation systems for all combinations of the languages. We constructed a multi-lingual translation corpus (N = 21) for travel conversation and developed translation systems for all combinations (420 combinations). We confirmed that every translation system achieved a high enough translation quality for practical use. These systems were released as the iPhone application "VoiceTra/TexTra."

4 Two major issues from research

In order to achieve a high-quality automat-

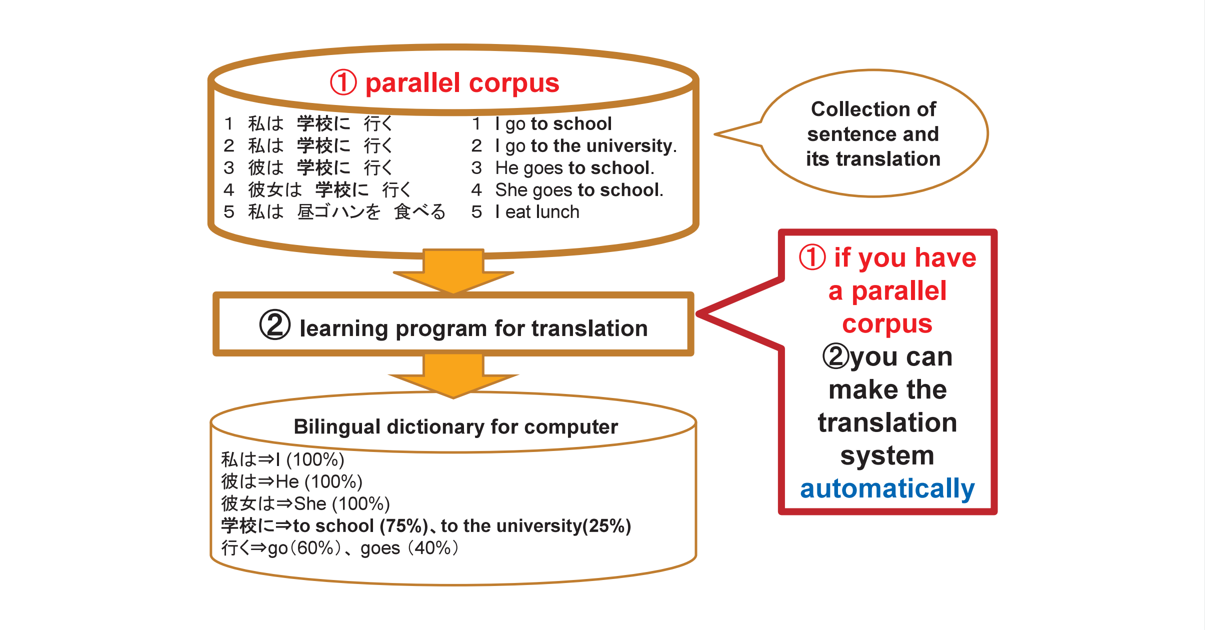

① parallel corpus
1 私は 学校に 行く
2 私は 学校に 行く
3 彼は 学校に 行く
4 彼女は 学校に 行く
5 私は 昼ゴハンを 食べる
1 I go to school
2 I go to the university.
3 He goes to school.
4 She goes to school.
5 I eat lunch
Collection of sentence and its translation
② learning program for translation
① if you have a parallel corpus ②you can make the translation system automatically
Bilingual dictionary for computer
私は⇒I (100%)
彼は⇒He (100%)
彼女は⇒She (100%)
学校に⇒to school (75%)、to the university(25%)
行く⇒go（60%）、goes（40%）

Fig.2 *Basis of corpus-based translation technology*

ic translation system using this corpus-based translation technology, there are two major issues to be solved.

(1) Collection of parallel corpora: It is known that the translation quality would reach a practical level if more than a certain volume of parallel corpora are available. It is hence important to establish a method of collecting parallel corpora efficiently in a short period of time.

(2) Sophistication of translation algorithm: Translation performance changes depending on the translation algorithm, even using the same amount of parallel corpora. It is therefore important to develop a superior algorithm.

We introduce an example of each issue below.

4.1 Collection of parallel corpora

Since parallel corpora are a major knowledge source for the corpus-based translation technology, the efficient collection of data is essential. We took two complementary approaches to this. (1) Computer-assisted approach, through Web crawling, creating bilingual translation data from a monolingual corpus, and a comparable corpus, and (2) human-based or society-based approach, collecting bilingual translations from the Web, hosting services of voluntary translation, and collaboration with external organizations. For details, see Chapter **4-2** “Efficient Technologies for Creating Parallel Corpora”.

4.2 Sophistication of translation algorithm

There are also many sub-themes in the sophistication of the translation algorithm, such as the precision improvement of word segmentation, the precision improvement of word alignment programs, the processing of proper nouns, the transliteration processing (see Chapter **4-3** “A Transliteration System Based on Bayesian Alignment and its Human Evaluation within a Machine Translation System”), the automatic acquisition of technical terms, adjustment to specific fields or topics, the use of syntax, the modeling of scenes, situations, and context, the method of appropriately mixing multiple translations, and the parallelization of model training.

Here we will explain the use of syntax. The order of words is a problem when we consider two completely-different languages such as Japanese and English, although it does not create a serious problem between similar languages such as Japanese and Korean, or Spanish and Italian. The basic order of Japanese words is SOV, while that of English words is SVO. In this case, translation in correct word orders is difficult. We therefore do not simply allow all possible word orders. Instead, we propose a method of using input syntax to restrict word orders and pick up only those orders that satisfy the requirement. We confirmed that the number of possible translations could be reduced significantly with this method and the error rate of Japanese-English translation could be decreased. For the method of using syntax for the unification of multiple hypotheses of translation, see Chapter **4-4** "Direct Use of Syntactic Information for Machine Translation System Combination".

5 Examples of high-quality specialized translation

5.1 e-commerce

One of the fields that require high-quality translation systems is the electronic commerce (e-commerce). This is a growing industry in the process of going international. Automatic high-quality translation systems are essential for the industry since it handles a tremendous number of products with a high merchandise turnover ratio. By combining (1) efficient translation backed up NICT's translation support technology, (2) efficient creation of a technical term dictionary using its automatic dictionary creation technology, and (3) syntax-based statistical machine translation, a high-quality translation system for e-commerce was developed. This technology was transferred to a business party and is being utilized in a global site for one of Japan's largest apparel e-commerce companies.

5.2 Patent translation

As pointed out in the Patent Agency's Global IP Initiative (concrete policies for global development of infrastructure of intellectual properties) in July 2011*[1], the number of patents from China and Korea has been increasing and more and more patents are in dispute. In this situation, it would be beneficial for Japan to develop a search system of foreign patent literature with a translation function from Chinese or Korean to Japanese.

The sentences used in patent literature are usually long. We hence developed new technology for long-sentence translation. To be more specific, we constructed the following two methods. (1) Sentence division method: Dividing a long sentence by its superficial characteristics and combining the translation results of the divided parts, and (2) Noun phrase capsulation method: Encapsulating noun phrases, shortening the sentence, translating the shortened sentence, and putting the translated noun phrases back into the sentence. By combining these methods we realized significant enhancements in the translation performance (See Fig. 3).

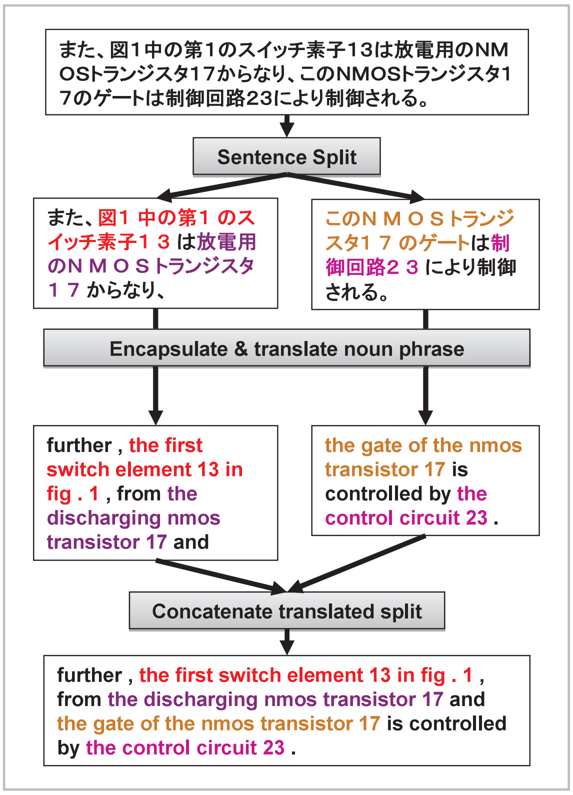

また、図1中の第1のスイッチ素子13は放電用のNMOSトランジスタ17からなり、このNMOSトランジスタ17のゲートは制御回路23により制御される。
Sentence Split
また、図１中の第１のスイッチ素子１３は放電用のＮＭＯＳトランジスタ１７からなり、
このＮＭＯＳトランジスタ１７のゲートは制御回路２３により制御される。
Encapsulate & translate noun phrase
further , the first switch element 13 in fig . 1 , from the discharging nmos transistor 17 and
the gate of the nmos transistor 17 is controlled by the control circuit 23 .
Concatenate translated split
further , the first switch element 13 in fig . 1 , from the discharging nmos transistor 17 and the gate of the nmos transistor 17 is controlled by the control circuit 23 .

Fig.3 *Examples of long-sentence translation*

In the NTCIR project (2010-2011), a com-

*[1] http://www.jpo.go.jp/shiryou/toushin/shingikai/pdf/tizai_bukai_16_paper/siryou_01.pdf

petition-type international symposium Patent-MT was held in collaboration with NII to provide a patent translation corpus and compare translation performance. Twenty one research institutes from the US, Europe and Asia, including IBM and BBN, participated in the symposium. Comparison of the algorithms clarified that statistical machine translation is more promising than rule-based machine translation for English-Japanese and Chinese-English translation.

6 Future study plans

The following three steps are planned for future translation systems to translate any language in any field.

In Step 1, the method and base of the corpus will be established, the translation algorithm will be enhanced, and translation system models for several fields will be developed.

In Step 2, a social and economical mechanism for the realization of a multi-field multi-language corpus will be proposed.

In Step 3, the language translation technology will be made clear to promote collaborative activities with external organizations for the translation of any languages in any fields.

References

1 SUMITA Eiichiro, “Speech Translation Technology in MASTAR Project,” Special issue of this NICT Journal, 7-1, 2012.

2 UTIYAMA Masao, “Efficient Technologies for Creating Parallel Corpora,” Special issue of this NICT Journal, 4-2, 2012.

3 Andrew Finch, YASUDA Keiji, and SUMITA Eiichiro, “A Transliteration System Based on Bayesian Alignment and its Human Evaluation within a Machine Translation System,” Special issue of this NICT Journal, 4-3, 2012.

4 WATANABE Taro, “Direct Use of Syntactic Information for Machine Translation System Combination,” Special issue of this NICT Journal, 4-4, 2012.

(Accepted June 14, 2012)

***SUMITA Eiichiro**, Dr. Eng.*

Director, Multilingual Translation Laboratory, Universal Communication Research Institute

Natural Language Processing, Machine Translation