Development of a Japanese-Chinese Machine Translation System

Hitoshi Isahara, Sadao Kurohashi[1], Jun'ichi Tsujii[2], Kiyotaka Uchimoto[3], Hiroshi Nakagawa[2], Hiroyuki Kaji[4] and Shun'ichi Kikuchi[5]

Computational Linguistics Group
National Institute of Information and Communications Technology (NICT)
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289
Japan
isahara@nict.go.jp

Abstract

We are aiming to overcome language barriers by creating a high-performance natural language processing technology, which will enable the processing of human language using computers. We have embarked on a five-year project, starting in 2006, to develop a Japanese-Chinese translation system for scientific and technological papers, as the cooperation among the National Institute of Information and Communications Technology, the Japan Science and Technology Agency, Kyoto University, the University of Tokyo, and Shizuoka University.

Introduction

Today, thanks to progress in science and technology, the world can instantly share all sorts of information, and a variety of data collected from around the world now forms an increasingly significant part of daily life. Since much of this information is expressed in words, differences in languages can be an obstacle to the distribution and use of knowledge.

We are aiming to overcome language barriers by creating a high-performance natural language processing technology, which will enable the processing of human language using computers. We have embarked on a five-year project, starting in 2006, to develop a Japanese-Chinese translation system for scientific and technological papers, as the cooperation among the National Institute of Information and Communications Technology (NICT), the Japan Science and Technology Agency (JST), Kyoto University, the University of Tokyo, and Shizuoka University. Part of this research and development is carried out as a study under the auspices of the "Special Coordination Funds for Promoting Science and Technology - Research and Development Program for Resolving Critical Issues," as "R&D for Japanese-Chinese and Chinese-Japanese Language Processing Technology."

Overview

While not as problematic among Western nations, the distribution of information in English is met with difficulty in Asia. We believe that, as an Asian nation, we should develop a machine translation system for Asian languages. As the first step in this endeavor, we have begun development of a machine translation system, mainly for scientific and technological materials in Chinese and Japanese, to keep pace with the significant progress we are seeing in various fields. We have carried out cooperative research projects with China, India, and other Southeast Asian nations, and in the future, we plan to expand the target languages of the system to include an even larger number of Asian languages.

Methodology

In the past, the implementation of machine translation has adopted a range of approaches, including the transfer method and the interlingua method. In the transfer method, the input text in the original language is first analyzed, and then the sentence structure is mapped out in accordance with the grammar of the original language. This sentence structure is then converted into that of the target language using transfer rules, to create a corresponding sentence. In the interlingua method (pivot method), the input sentence undergoes a deeper analysis, and is converted into an expression described in an intermediate language that is not dependent on a specific language. The sentence in the target language is then created based on the structure of the intermediate expression. Since the interlingua method carries out translation based on the identification of meaning, it allows for a more liberal translation and results in more natural phrasing. However, this demands that processing provide a deeper understanding of meaning, while at the same time handling massive volumes of information. On the other hand, the transfer method requires the description of a great number of conversion rules, which results in a proportional increase in the number of required rules when multiple languages are involved. Both methods involve compilation from various sources (grammar rules, lexicons, thesaurus, etc.), which must be performed manually, and the establishment of a coherent approach to this task of compilation is extremely difficult. Recently, Statistical Machine Translation (SMT) is widely studied and shows its promising features. However, the capability to handle pairs of languages with very different grammatical and/or lexical structure is still unclear.

---

[1] Kyoto University
[2] University of Tokyo
[3] National Institute of Information and Communications Technology (NICT)
[4] Shizuoka University
[5] Japan Science and Technology Agency (JST)

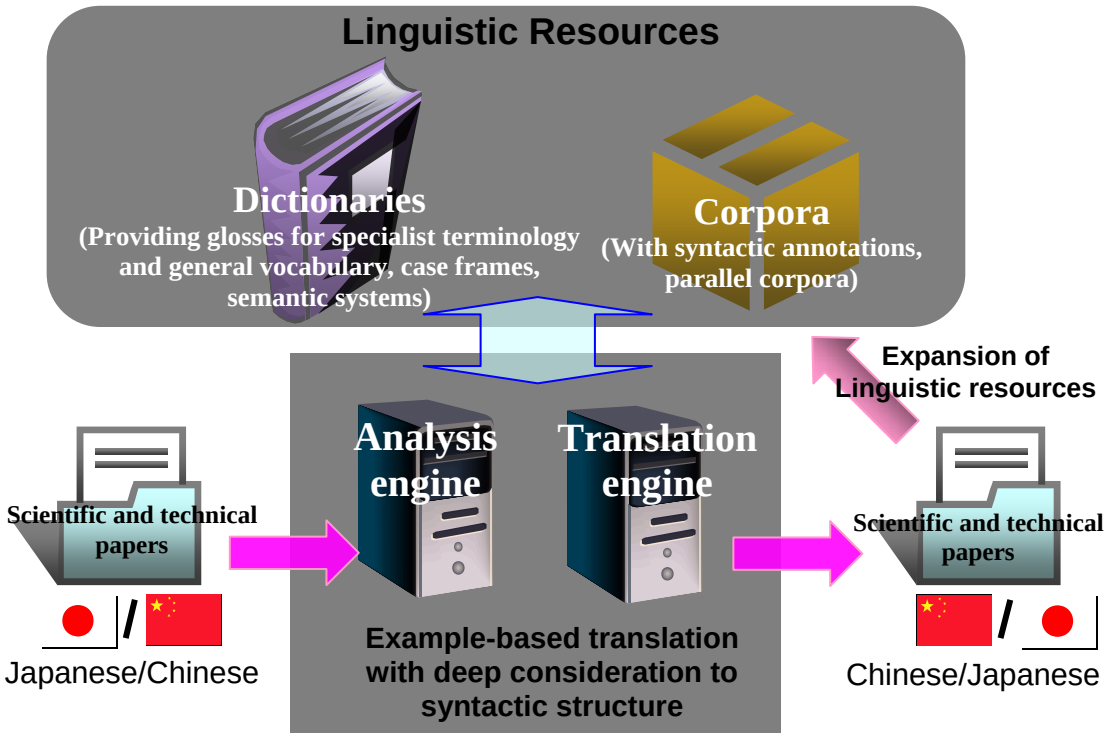

Figure 1. Delivering a practical machine translation system within a new framework

In contrast, it is believed that when a human performs translation, he or she is not strictly applying such knowledge, but is instead translating sentences through combinations of recollected phrases in the target language. Based on this hypothesis, Dr. Makoto Nagao proposed an Example-Based Machine Translation (EBMT) in 1981. In an example-based machine translation system, translation is executed based on the similarity between the input sentence and an example contained in a voluminous parallel corpus. This process demands a method of judging the similarity between the two—yet without the need to create a large collection of transfer rules. Furthermore, the quality of translation may be improved simply by adding examples to the database, and since the example translations will naturally reflect contextual differences in translations, these differences will also be reflected in the machine translation results.

At the time of proposal of EBMT, computing capacity was insufficient to produce a practical system under this approach. However, rapid improvements in computer performance in recent years, in addition to the development of a method for judging similarity between examples (through reference to a database of syntactically analyzed sentences accumulated in the system) has now formed the foundation for the establishment of a practical example-based machine translation system.

Toward the Development of a Practical Machine Translation System

As we mentioned above, our final goal is to develop system which covers wide variety of Asian languages. However, because the world pays attention to China because China is making a remarkable progress in science and technology, we started our project from Japanese-Chinese translation.

Our objectives are;

- Making scientific and technological information in China and other Asian countries easily usable in Japan.
- Promoting the distribution of documents to China and other countries about science and technology in which Japan is at the forefront.
- Contributing to the development of science and technology in Asian countries by the information exchange through machine translation.

The goal of this project is to develop, within a period of five years, a practical machine translation system between the Japanese and Chinese languages focused on scientific and technological materials. In this endeavor we have adopted the example-based translation method, which provides a better reflection of linguistic structures, and syntactic information will be used in many parts of the translation engines.

Figure 1 presents an outline of the system under development. Currently our target domains are information science, biological science and environmental science.

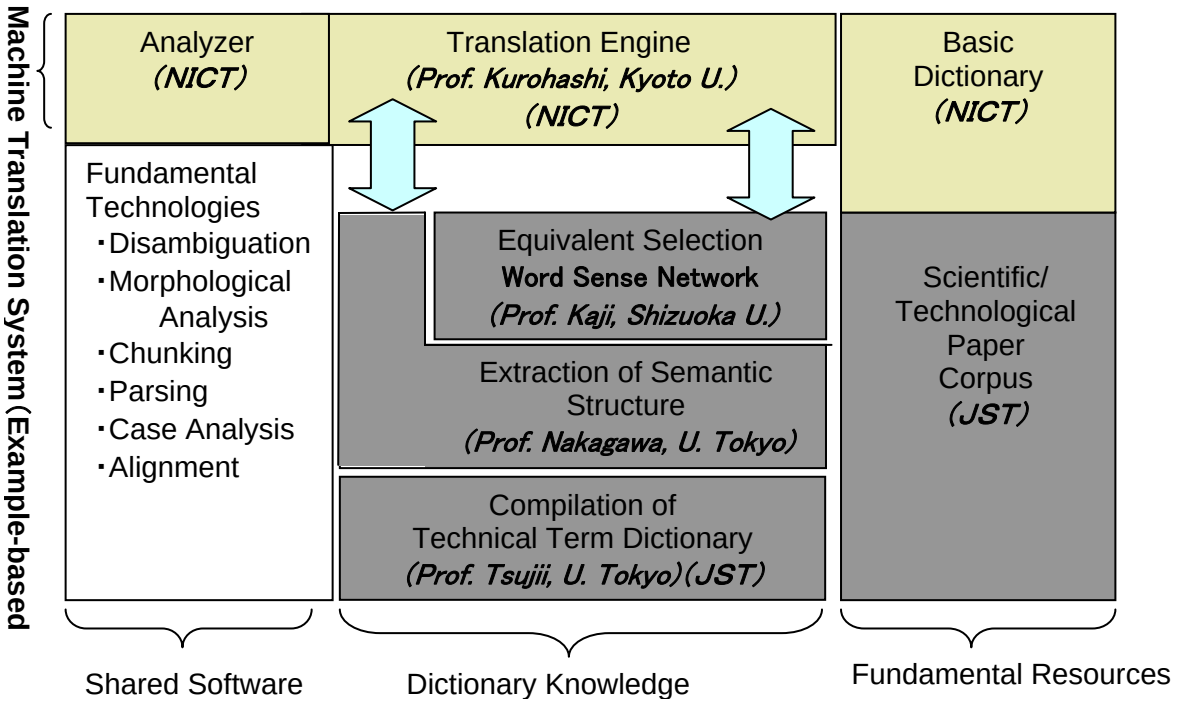

Figure 2. Collaboration between System development and Dictionary development

Research Teams

This project consists of five members, i.e. NICT, JST, Kyoto University, University of Tokyo and Shizuoka University. The relations between these five teams are shown in Figure 2. Kyoto University and NICT are mainly responsible for R&D of software such as translation engine and analyzers. University of Tokyo, Shizuoka University and JST are mainly responsible for R&D of resources such as dictionaries and corpora.

The R&D of the basic technology will involve improving analyzers of the Japanese and Chinese languages. Further, we plan to modify example-based translation to accommodate the lengthy and complex sentences often seen in scientific and technological materials.

EBMT requires the accumulation of voluminous examples; accordingly, we are planning to develop a parallel corpus on the scale of 10 million sentences. We plan to extract parallel sentences from existing comparable texts semi-automatically and align words and phrases semi-automatically. We also plan to make the best use of existing linguistic resources and language processing technology owned by our institutes.

Goals

In this five year project, there are goals for the 3rd year and the 5th year.

- Goal for the 3rd year
  Evaluate the Japanese-Chinese machine translation prototype system for specific target domains.
- Goal for the 5th year
  Improve the Chinese analysis performance, and complete demonstration experiments on the Japanese-Chinese and Chinese-Japanese machine translation prototype system

Even during the course of the project, we plan to publicize the language resources (such as the corpus) to the full extent possible, for research purposes. We will also work to publicize the contents and results of our research widely, as part of our outreach activities.

Resource Compilation

In this project, we will compile Japanese-Chinese word dictionary and parallel corpus.

As for word dictionary, NICT has huge Japanese-English electronic dictionary (EDR dictionary) with 400,000 words and is now expanding it into Japanese-Chinese-English dictionary. So far, NICT proposed a method to semi-automatically develop a Japanese-Chinese dictionary with English as the intermediary language.

As for preparation of a parallel corpus, we combine manual translation and automatic gathering. For manual translation, we are selecting Japanese texts to be translated and started to translate them manually into the target Chinese language. We will translate Japanese texts (mainly scientific literature) into Chinese to create a large parallel corpus (over 1 million sentences) in 5 years. For automatic gathering, we gather parallel corpora from Web and also extract parallel text from non-parallel corpora by using NLP techniques.

	Analysis/Generation Technology	Translation Technology	Linguistic Resource Development Technology
Application	Japanese analysis/generation Chinese analysis/*generation* Korean analysis/generation *Thai analysis/generation* *Other Asian language analysis/generation* English analysis/generation European language analysis/generation *Other language analysis/generation*	Translation between languages with similar structures Translation between languages with different structures Translation of short sentences such as travel conversations Translation of long sentences such as scientific and technological document Field-dependent translation *Limited-domain translation (Domain-adaptive, author-adaptive)* *Semiautomatic translation*	Development of a large-scale dictionary for translation Knowledge retrieval Collection and expansion of a corpus Annotation of language information to a corpus *Standardization of a corpus*
Basic	Morphological analysis Syntactic analysis (Parallel structure analysis) Semantic analysis (Chunking, case analysis, *named entity extraction)* *Context analysis (Discourse structure analysis)* *Sentence generation*	Aligning words and phrases in a bilingual corpus Flexible matching Selection of translation equivalents *Automated evaluation of machine translation* *Extraction and translation of named entities, unknown words and idioms*	Recognition of different notations Automated acquisition of translation equivalents Extraction of technical terms Estimation of semantic relationship between words Generation of semantic networks Dictionary development support system Linguistic information annotation support system Aligning sentences
Theory	Grammar theory Leaning theory	Example-based translation Statistics-based translation Rule-based translation Hybrid translation	Clustering Elastic Matching Multivariate statistical analysis

Table 1. Technology Map of Multilingual Translation Technology

Technologies

We utilize example-based translation with further consideration of linguistic structures. EBMT system basically generates sentences in target language by extracting and combining examples in parallel text database whose source language sentence is similar to the input sentence. The specific feature of our system is to utilize deep syntactic analysis. In advance, parallel texts in the example database are analyzed syntactically and aligned with words using syntactic information. In translation phase, we analyze input sentence syntactically, extract parts of sentences from example database and combine them to generate sentences in target language by considering syntactic structures. Finally, the ordering of words in a sentence is decided using information extracted from monolingual corpora.

Theoretical issues, basic NLP technologies and their applications which we will develop are shown in Table 1. Red (underlined) parts mean issues which we will concentrate. Blue (italic) parts mean issues which we will develop but out of this Chinese-Japanese translation project. For example, in the part of “application” and “Analysis/Generation Technology”, we will not focus on Japanese analysis/generation, because they have already reach to practical level. However, we will focus on R&D of Chinese Analysis, because the levels of the accuracy of the output of Chinese analyzer are still low for practical system and the level of analyzer is crucial for the realization of high quality machine translation. We will develop analyzer of Thai and other Asian language at NICT, but it is out of our project of Chinese-Japanese machine translation. As for Korean and English analysis, they have been already reached to the proper accuracy.

The relations between our method and other methods on machine translation are also shown in Table 2. Comparing to other method, i.e., rule-based and statistical based machine translation, our methodology suit to translate longer sentences between different types of languages. Our methodology will be applied easily to extend to other languages, though the language pairs handled are rather different in styles and grammatical phenomena.

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="2">Similarity of linguistic structure</td></tr>
<tr><td>Low</td><td>High</td></tr>
<tr><td rowspan="2">Length of sentence</td><td>Long</td><td>Rule-based translation
<u>Example-based translation with further consideration of linguistic structures</u></td><td rowspan="2"></td></tr>
<tr><td>Short</td><td colspan="2">Statistics-based translation
Example-based translation
Rule-based translation</td></tr>
<tr><td rowspan="2">Extensibility to other languages</td><td>High</td><td><u>Example-based translation with further consideration of linguistic structures</u></td><td>Statistics-based translation
Example-based translation</td></tr>
<tr><td>Low</td><td colspan="2">Rule-based translation</td></tr>
</table>

Table 2. Technology Map of Translation Methods in Multilingual Translation

Future

We have developed prototype system of Japanese-Chinese-English translation system and demonstrated the system in several occasions. Figure 3 shows the screenshot of the prototype system.

In parallel with this project we will apply same technologies to other Asian languages such as Thai. Because NICT has its oversea research laboratory in Thailand (TCL: Thai Computational Linguistics Laboratory), NICT will conduct MT activities at TCL. Because we adopt example-based mechanism, application to other languages will be possible without substantial changes of the system, once their corpora are developed. Findings and technologies obtained in this project can also be utilized in development of corpora in Asian languages making possible production of linguistic resources at low cost.

The goal of science and technology is to provide a comfortable standard of living for everyone, regardless of individual ability or social status. Our goal is to help create an environment in which people everywhere can share information regardless of language barriers; we can accomplish this in part by giving computers the ability to process language. We hope that the present R&D into machine translation system will help us achieve this aim.

Figure 3. Screenshot of the prototype sysmtem