Promoting Science and Technology Exchange using Machine Translation

Toshiaki Nakazawa nakazawa@pa.jst.jp
Japan Science and Technology Agency (JST), 5-3, Yonbancho, Chiyoda-ku, Tokyo 102-8666
Japan

1 Introduction

There are plenty of useful scientific and technical documents which are written in languages other than English, and are referenced domestically. Accessing these domestic documents in other countries is very important in order to know what has been accomplished and what is needed next in the science and technology fields. However, we need to surmount the language barrier to directly access these valuable documents. One obvious way to achieve this is using machine translation systems to translate foreign documents into the users' language. Even after the long history of developing machine translation systems among East Asian languages, there is still no practical system. We have launched a project to develop practical machine translation technology for promoting science and technology exchange. As the starting point, we aim at developping Chinese ↔ Japanese practical machine translation system. In this talk, I will introduce the background, goals and status of the project. Also, I will give you the summary of the 2nd Workshop on Asian Translation (WAT2015)[1] where Chinese ↔ Japanese scientific paper translation subtasks has been carried out.

2 Background

Figure 1 shows the number of scientific papers in the world which are written in "English". We can presume that the number of papers written in each language has the similar proportion to this graph. You can see that the number of papers from China is rapidly growing in recent years, which means we have a large number of "Chinese" papers.

Some of them may include important and useful information but are never published in English. We, not Chinese native speakers, cannot get any information from such papers as they are, and throw them to the MT engines to translate into our mother tongue. Chinese-to-English and Japanese-to-English MT systems have been developped for years and the quality are sufficient enough for gisting the paper. However there is little MT system between Asian languages such as Chinese-to-Japanese and vice versa which is good enough for gisting. Therefore, we have launched a project to develop practical machine translation technology between East Asian languages.

3 Goals

We have 3 goals in the project and they are summarized in Figure 2.

[1] http://lotus.kuee.kyoto-u.ac.jp/WAT/

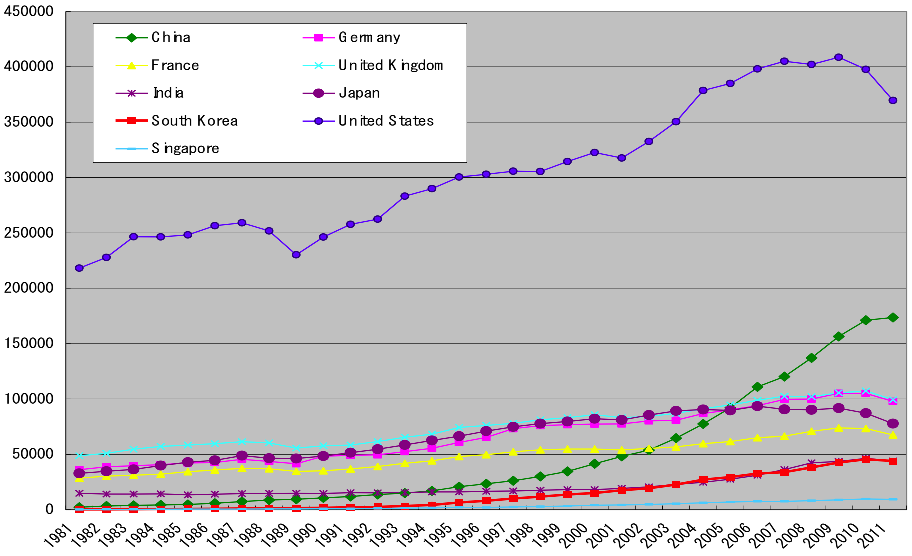

Figure 1: The number of "English" scientific papers in the world (aggregated by JST from Web of Science by Thomson Reuters).

1. Language Resource Construction
   Most of recent corpus-based machine translation systems require parallel corpus where translation rules are acquired. In addition, parallel dictionary is necessary to cover technical terms because the number of technical terms keeps growing.

2. Sentence Analyzers (especially for Chinese)
   Both Chinese and Japanese require the word segmentation technology because they do not have white spaces between words. The word segmentation of Chinese is more difficult than that of Japanese because Chinese sentences are basically composed of only Chinese characters. Also, Chinese and Japanese have different language characteristics, so the high-level sentence analysis such as parsing is important to achieve high translation quality.

3. MT Engine Development
   Our MT engine should be able to use the rich information from the deep sentence analysis. We choose dependency-to-dependency example-based machine translation method in our project.

4 Workshop on Asian Translation

The Workshop on Asian Translation (WAT) is a new open evaluation campaign focusing on Asian languages organized by NICT[2] and JST. The 2nd WAT has 6 subtasks: English ↔ Japanese scientific paper translation, Chinese ↔ Japanese scientific paper translation, Chinese → Japanese patent translation and Korean → Japanese patent translation. The distinctions of WAT are:

[2] National Institute of Information and Communications Technology

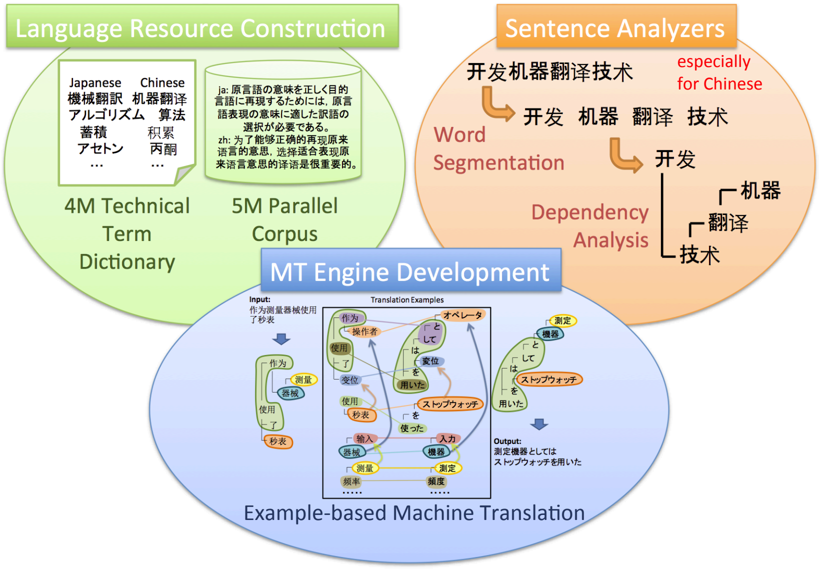

Figure 2: The goals of the JST Chinese-Japanese MT project.

- Open innovation platform
  The test data is fixed and open, so you can repeat evaluations on the same data and confirm changes in translation accuracy over time. WAT has no deadline for the automatic translation quality evaluation (continuous evaluation), so you can submit translation results at any time.

- Domain and language pairs
  WAT is the world's first workshop that uses scientific papers as a domain and Japanese-Chinese as a language pair. In the future, we will add more Asian languages, such as Korean, Vietnamese, Indonesian, Thai, Myanmar and so on.

- Context-aware evaluation
  The test data of WAT is prepared using the paragraph as a unit, while almost all other evaluation campaigns use the sentence as a unit. We would like to consider how to realize context-aware evaluation in WAT.

- Evaluation method
  Evaluation will be done by both automatic and human evaluation. For human evaluation, WAT will use crowdsourcing, which is low cost and allows multiple evaluations.

I will introduce the results and some insights acquired from this year's workshop (WAT2015).