A Cross System Machine Translation

Thepchai Supnithi, Virach Sornlertlamvanich and Thatsanee Charoenporn
Information Research and Development Division
National Electronics and Computer Technology Center
112 Thailand Science Park, Paholyothin Rd.,
Klong 1, Klong Luang, Pathumthani 12120, THAILAND
{thepchai, virach, thatsanee}@nectec.or.th

ABSTRACT - The rapid growth of Internet Technology, especially user friendliness approach, helps increase the number of Internet users and the amount of information in the cyberspace. There is a countless amount of information in languages. This has spread developments of MT systems. The focus of our approach is to increase the reusability of those MT systems by using Cross System machine translation. Using natural language as an intermediate language, such as English, will help us use the information in Internet qualitatively. In this paper, we point out some problems that may cause the efficiency to decrease when a sentence is translated from a second language to a third language. A novel method is proposed to solve this problem.

KEY WORDS – machine translation, intermediate language, multilingual, cross-language, XML, reusability

1. Introduction

Machine Translation (MT) is an automatic system that provides an ability to convert a message written in one language (source language: SL) to another (target language: TL)[1]. The interlingua approach [2,3], a methodology of constructing an intermediate language, is a dominant approach in standalone system to support multi-language. Many products such as, SYSTRAN [4], BESTILAND [5], are implemented using this approach. Interlingua approach is helpful for a central server, but it is difficult to complete concepts in Interlingua.

The rapid growth of Internet Technology, especially user friendliness approach, helps increase the population of users who access the Internet and the amount of information in the cyberspace. With the increasing amount of online information and the rapid growth of non-English speaking Internet hosts, it is becoming incteasingly important to offer users universal access to valuable information resources in different languages. The European Multilingual Information Retrieval (EMIR) project [6], the MULINEX project [7], the TwentyOne project [8], and the cross-language retrieval track in TREC [9] conference all reflect people's interest in providing interoperability among different language processing environments and multilingual information retrieval.

Distributed system technology plays an important role to enable us to manage information from various places. This makes it unnecessary to access only the central server. It helps machine translation developers to work individually. Yasuhara [10] wrote that many machine translation systems were developed, especially from local language to English, and the language has an important role as an intermediate language.

Our paper tries to apply a distributed technique by using English language, which is mostly used by non-English speakers as a second language to be an intermediate language. Our approach is not aimed to show that it is better than the interlingua approach, but it is another solution for us to use existing resources in cyberspace. We hope that it is possible to help developers build the machine translation that will support all languages taking into account of cost, quantity, and time consumption.

In section 2, we show cross system MT approach. In section 3, an example of our approach is given. In section 4 we illustrate drawbacks of this technique and give an example about how to examine these problems.

2. Cross System MT

The major significance of Asian languages is the variation of languages in the region; most of which use their own unique set of characters. In terms of

This article is a reprint of a paper presented at the COLING 2002 – Proceedings of the Workshop on Machine Translation in Asia (August 2002, Taiwan)

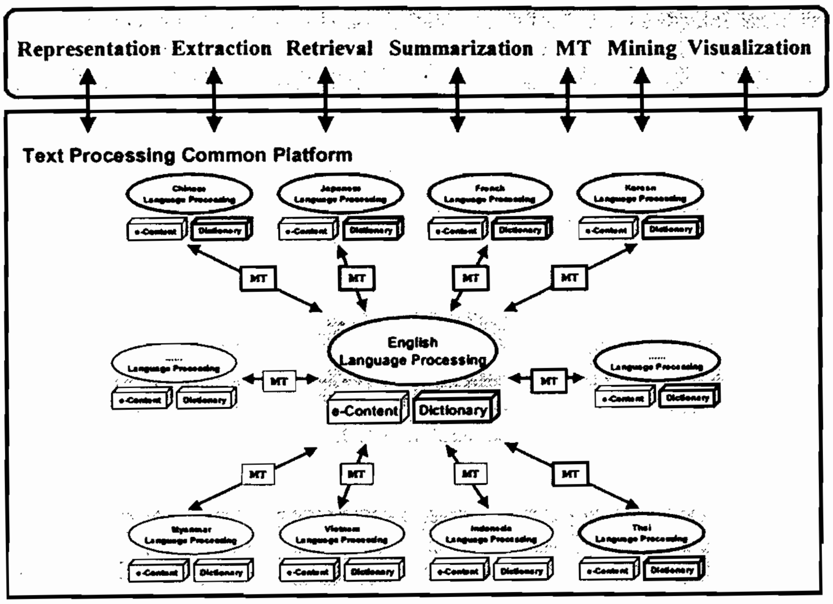
Representation
Extraction
Retrieval
Summarization
MT
Mining
Visualization
Text Processing Common Platform
Chinese Language Processing
Japanese Language Processing
French Language Processing
Korean Language Processing
...... Language Processing
English Language Processing
Myanmar Language Processing
Vietnam Language Processing
Indonesia Language Processing
Thai Language Processing
e-Content
Dictionary
MT

Figure 1. Cross System Architecture

grammar, some (Thai, Laotian, Japanese, Chinese, etc.) do not indicate word boundary, some (Thai, Laotian, etc.) do not inflect while others (Japanese, Korean, etc.) provide particles to indicate the word grammatical function, some are not distinguishable between sentences and phrases, etc. These are the basic difficulties that interest the researchers in the field of machine translation and the application.

Due to these varieties, it is difficult to build an MT system that supports all languages taking into account of cost, quantity, and time consumption. Cross system machine translation approach is, therefore, an essential concept that helps reduce these problems by reusing the large amount of information existing in Internet.

Figure 1 shows an idea of our cross system machine translation approach. Since the technology of building MT can be transferred from us to other countries in this region and we know that English is broadly used as a bridge to communicate among different languages. It is simpler for a local developer to build an MT system from his/her local language to English (L1←→E). If all countries have their own Ln←→E MT system, sharing English as an intermediate representation language reduces problems shown above. Moreover, there are many different ways to develop a MT system. Our approach is to encapsulate the type differences among MT systems. Thus we can decrease the gap among languages by connecting the MT system of each local language.

Our cross system MT also offers a good infrastructure for many future applications such as e-commerce, digital archive, e-publishing, and so on as shown in figure 1. Next we show an example of the usage of a cross system MT.

3. A Usage Sample of Cross System MT

This chapter shows an application of using our cross system MT. Figure 2 shows an example of our expected application tool for information retrieval. We have two MT systems in our workgroup, a bilingual Thai←→English MT system and Japanese←→English MT system.When a user starts to search by input a keyword in Thai, such as a word “คอมพิวเตอร์ [kom pyu ter] ”(step1). The word “คอมพิวเตอร์ ”will be sent to the Thai→English MT system to translate into “computer”(step2). The word

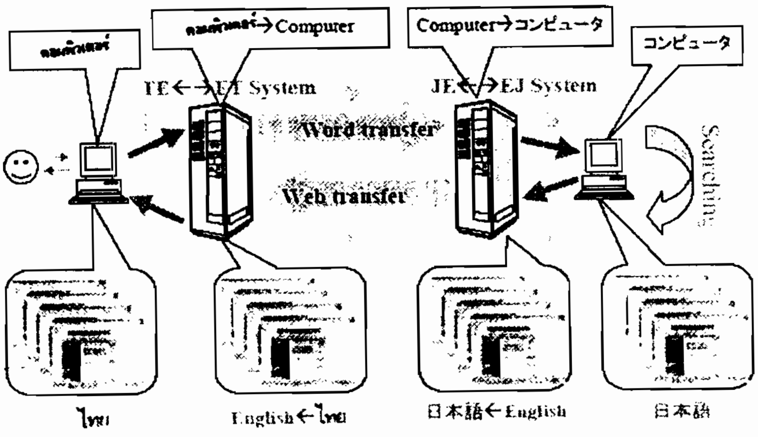

Figure2 Cross System Information Retrieval

"computer" will be sent to the English→Japanese MT system to translate into "コンピュータ"(step3). The word "コンピュータ" will be used as a keyword to search for Japanese web pages by a Japanese search engine (step4). The result of Japanese web pages from the search engine will be sent to Japanese →English MT system to translate into English web pages (step 5). The result of English homepages will again be sent to English→Thai MT system to translate into Thai pages (step 6). Finally, the output of the workgroup is web pages that contain the keyword "คอมพิวเตอร์ " (step 7). These web pages are selected from Japanese web pages.

This approach helps us to develop a MT system that supports all languages taking into account of cost, quantity, and time consumption. If each pair of languages can be translated perfectly, it should produce a satisfactory result for cross system technique. There is, however, a major problem that we have to consider about the efficiency due to the fact that the efficiency of each pair of machine translation is not completed. The more languages we include in our system, the less efficient the system becomes. In order to find out the solution for this problem, we show the linguistic problems and an example for solving those problems in the next section.

4. Problems and Solution to Improve Efficiency from Cross System Approach

The cross system MT approach seems to be another solution to develop an MT system that is possible to connect to other languages. However, it has a major problem of efficiency decreasing. When we consider the efficiency of translation from the SL to TL, we find that the machine translation cannot be transferred completely. Section 4.1 gives the problem that is possible in linguistics and in section 4.2 our approach to transfer information from the first SL to TL in order to examine these problems is described.

4.1 Linguistic Problems

Manisara Meechoonuk and Somporn Rakchonlatee [11] evaluated the result of machine translation developed in Thailand, they define the linguistic problems as shown in table 1. In the investigation, they show that the result from MT that is perfect translation is about 29%, comprehensible translation is about 55%, and incomprehensible translation for the remaining. They also state that "Mismatch Concept" is about 34% found and is the most common linguistic

Table 1. List of Linguistic Problems and Meaning

Linguistics Problems	Meaning
Mismatch Concept	Inappropriate concept is selected
Misplaced Modifiers	Wrong position of words, phrases or modifiers in TL resulting in distortion of meaning
Inappropriate Literal Translation	An inappropriate translation that follows closely the form of SL. It can be categorized into 1) part of speech, 2) order, 3) idiom.
Addition of words or phrases	Some words in TL that are not stated in SL are added.
Omission of words	The meaning of a word or words when translating from SL to TL is/are leaved out.
Insufficient definitions of idioms, word verbs, and phrasal verbs	The scope or number of words in electronic storage is either limited or inaccurate according to the meanings of words in SL
Translation which does not conform to Target language grammar	A difference sentence structure in TL that may cause an incomprehensible translation.
Implicit in both SL and TL	The implied meaning of a word in the SL is not expressed clearly or fully in TL
Active in SL but passive in TL	The participles appear in SL as active forms but are translated into passive forms in TL
Insufficient Dictionary Definitions	The scope or number of words in the electronic data dictionary is limited
Different Semantic Segmentation between SL and TL	Using difference marker, such as punctuation or space in SL and TL may cause the incomprehensible translation
Specific in SL but generic in TL	A specific word in SL is referred as a general meaning in TL

problems.

These linguistic problems cause the incomplete translation. An MT system cannot correctly translate from second language to third language if the result of translation from the MT system from first language to second language is not perfect. We, however, find that "Insufficient definitions of idioms, two-word verbs, and phrasal verbs" and "Insufficient Dictionary Definitions" problems cannot be fixed by the cross language system because of the lack of information before the translation in SL.

We examine this problem by adding the information from the first language together with the result of second language. When the MT system translates from the second language to the third one , it can request additional information that is attached from the first language as a reference. Next we show some examples of using our method.

4.2 Examples of Information Transfer

In section 4.1, we give linguistic problems that cause a decrease of efficiency of translation. In order to increase the efficiency, we illustrate how to give the additional information in order to help the translation when the second language functions as a source language. Our approach is that the information we receive from the first language is the most appropriate information. If we can add additional information from the first language as much as we can, it will help us increase the efficiency of translation. We use an XML as a language to transfer from first language to other languages.

For example, we have two MT systems, a Thai←→English MT system and English ←→ Japanese MT system.

(1) Looking at the first sentence,
"เด็กดื่มยา (dek duum ya)"
means "A child drinks a medicine".
But it is translated into "A child drinks a drug" by the Thai→English MT system.

The problem of this sentence is classified as a "Mismatch Concept" problem. A word "ยา" has several meanings, such as medicine, drug, cure, tonic and so on. For this problem we can add all concepts as a reference as follows.

"A child <AGT> drinks<... > a drug <OBJ:c#drug, c#medicine, c#pill, c#tonic> "

This will help the second MT system not to translate "ยา" as "drug", but refer all concepts of "ยา" before the translation. The result of translation should be, "子供が薬を飲む" by English →Japanese MT system.

(2) Looking at another sentence in Japanese.
"Internet Explorer 6 では、幅広いインターネットセキュリティとプライバシー基準を導入しています。"
means "A private standard and a wide range internet security has been introduced in Internet Explorer6".
But it is translated into "A private standard has been introduced as the Internet security whose it is wide range in Internet Explorer6".

The problem of this sentence is classified as a "Misplaced Modifiers" problem. "Internet security" which is defined as a private standard causes the misunderstanding. In this problem we can add a tag that refers to the functions of Internet security (TAR) as a reference as follows.

A private standard<TAR> has been introduced<...> as the Internet security<TAR> whose it is wide range<...> in Internet Explorer6<...>.

This will help the second MT system use the meaning of the CASE relation as TAR as a reference. The result of translation should be," มาตรฐานที่เป็นส่วนตัวและระบบความปลอดภัยอินเทอร์เน็ตในวง กว้างถูกบรรจุใน Internet Explorer 6 " by English →Thai MT system.

(3) Looking at another sentence in Japanese.
"野山には食べられる草があります。"
means "There is grass that can eat in hills and fields".
But it is translated into "There is grass eaten in hills and fields" by the Japanese→ English MT system.

The problem of this sentence is classified as an "Active in SL but passive in TL" problem. A word " 食べられる " which shows a possible shape is translated as a passive form. In this problem we can add a tag that refers to a possible form as a reference as follows.

"There is <...> grass<...> eaten <style:possible shape> in hills and fields<...>. "

This will help the second MT system to translate "eat" with a possible shape. The result of translation should be, "มีหญ้าที่สามารถรับประทานในเนินเขาและทุ่งหญ้า" by English→Thai MT system.

5. Conclusion

Due to the growth of information in digital world, the digital divide problem becomes much more severe. We need a system that is possible to be developed in a short period of time. Our approach, cross system MT is a key for developing such a system by considering only a bilingual MT system.

The cross system MT is an alternative solution to help the countries in the region to create their own MT systems by using existing resources when considering the cost, quantity, and time consumption. This will help them have a chance to strengthen their economic competitiveness and keep their culture preservations.

Since our method tries to give necessary information from the first language, it is appropriate to solve existent linguistics problems, such as, "mismatch conceptions", "inappropriate literal translation" and so on. It cannot, however, solve some linguistic problems because of the lack of information in the first language, such as "insufficient dictionary definitions".

In this paper, we give an outline of our approach, cross system MT, some usages of this system and define linguistics problems that may cause the efficiency decrease. We show an example of solving those problems.

The future work is to analyze essential information for each of the linguistic problems and subsequently implement our system.

References

[1] Briem, S. (2001): Machine Translation. http://www.simnet.is/stbr/vela/html

[2] Tucker, Allen B. (1987): Current Strategies in Machine Translation Research Development in Sergeri Nirenburg, Machine Translation Cambridge University Press (pp24).

[3] Sornlertlamvanich, V. (1995): Paper on Natural Language Process: Multi-lingual Machine Translation and Related Topics (1987-1994). Bangkok: National Electronics and Computer Technology Center.

[4] SYSTRAN: http://www.systransoft.com/

[5] BESTILAND: http://www.bestiland.com/

[6] Fluhr, C. and Radwan, K. (1993): Full Text Databases as Lexical Semantic Knowledge for Multilingual Interrogation and Machine Translation. In EWAIC' 93.

[7] Erbach, G. Neumann, G. and Uszkoreit H. (1997): MULINEX: Multilingual Indexing, Navigation, and Editing Extensions for the World Wide Web, in Working Notes of AAAI Spring Symposium on Cross Language Text and Speech. Standford California.

[8] Hiemstra, D.A (1998): Linguistically Motivated Probabilistic Model of Information Retrieval, In: C. Nicolaou and C. Stephanids(eds.) Proc. Of the second European Conf. On Research and Advanced Technology for Digital Libraries, pp 569-584.

[9] Hiemstra, D.A(1995): The 4th Text Retrieval Conference(TREC-4), Gaitherburg, Md., Nov 1-3.

[10] Eiji Komatsu, Cui Jin, and Hiroshi Yasuhara. (1993): A Mono-lingual Corpus-based Machine Translation of the Interlingua Method. In *Proceedings of TMI 93.*

[11] Meechoonuk M, and Rakchonlatee S. (2001): An Analysis of Text Translation by Machine, A master Thesis in Faculty of Arts, School of Language and Communication, NIDA

Thepchai Supnithi received his B.Sc. (Mathematics) from Chulalongkorn University in 1992 and received M.Eng. and Ph.D. (Electronics Engineering specific in Knowledge Engineering) from Osaka University in 1997 and 2001 respectively. His fields of expertise include Education System, Computer Supported Collaborative Learning, Knowledge Engineering, Natural Language Processing and Machine Translation. He joined Text processing Technology Section, NECTEC from 2001. He joined Linguistics group, CRL Japan from October 2002 to March 2003 as an invited researcher. Now he is a researcher and chief of Text Processing Technology Section, Information and Knowledge Division, NECTEC

Virach Sornlertlamvanich received his B.Eng. (Precision Mechanics) and M.Eng. (Precision Mechanics) from Kyoto University in 1984 and 1986 respectively. He received his D.Eng. (Computer Science) from Tokyo Institute of Technology in 1996. He was a researcher involving MT R&D in NEC Corporation (Japan) during 1988-1992. He founded LINKS Lab, NECTEC in 1992, was a Project leader of the cooperating project of MMT. He has been with Information Research and Development Division as the director since the start of the division. Now he is a research supervisor of Thai Computational Linguistics Laboratory, Communications Research Center, Japan.

Thatsanee Charoenporn received her BA.(Thai) from Chulalongkorn University in 1988, and MA. (Linguistics) from Thammasat University in 1992. Before moving to be a research associate of Thai Computational Linguistics Laboratory, Communications Research Center, Japan, she has been with Information Research and Development Division, NECTEC as a researcher and chief of Text Processing Technology Section.