MT R&D IN ASIA

Hozumi Tanaka
Department of Computer Science, Tokyo Institute of Technology
2-12-1, Oookayama, Meguro-ku, Tokyo 152
Tel: +81-3-5734-3046 Fax: +81-3-5734-2915
tanaka@cs.titech.ac.jp

Abstract

There is a big shift in MT R&D in this region after many large-scale projects conducted in the past ten years. Multi-lingual Machine Translation (MMT) project is one of the significant R&D projects that increased a great number of NLP related researchers and research activities which can be seen in the increasing number of the research institutes in the recent years. We learned a lot from the collaboration research across languages and we still hope that it will be a rigorous step for the future MT R&D in this region. Though the MT systems are still far from the extreme goal of the perfect translation, it can be observed that the MT systems are actually used to support information retrieval from the Internet.

1 Introduction

Machine translation R&D activities became in sight after the Japanese national MT project, the Mu-project [7]. Multi-lingual Machine Translation (MMT) project was one of the multi-national R&D activities that made a great contribution to the natural language processing community in Asian countries. Many institutes for MT R&D in this region have been established thereafter. Both government and private sectors realize the necessity and the possibility in developing the MT.

Though there is a bottleneck in raising the translation accuracy of the MT, users can be satisfied by the speed of translation and user-friendly interface which can somehow support in retrieving information through the Internet. In Japanese market, it is observed that users also use MT in the way of getting outline information from the English sources for quick reference, whereas the full function of high accurate translation is still the extreme goal for MT development. With the current MT technology and the advantage of hardware and network environment, MT is another essential tool for surfing through the Internet.

With the experience in the MMT project, many countries in this region start developing a more practical MT system. The bilingual MT between English and their own languages is highly required especially in the present Internet environment. We give a brief introduction of the MMT project and discuss some problems in developing the multi-lingual system in the next section. Section 3 presents the state-of-the-art MT in Asia and the present state of R&D in some remarkable countries. Section 4 gives the information about NLP research activities and some of the available NLP resources.

2 Multi-lingual Machine Translation Project (1987-1994)

Japanese government launched a multi-lingual machine translation (MMT) project in the late of 1980s, in collaboration with China, Indonesia, Malaysia and Thailand. The past success in

commercialized machine translation systems seemed to be a strong driving force of the project. The project started by the gathering of seven leading private companies and Center of the International Cooperation for Computerization (CICC) of Japan, as organized in Figure 1 [5].

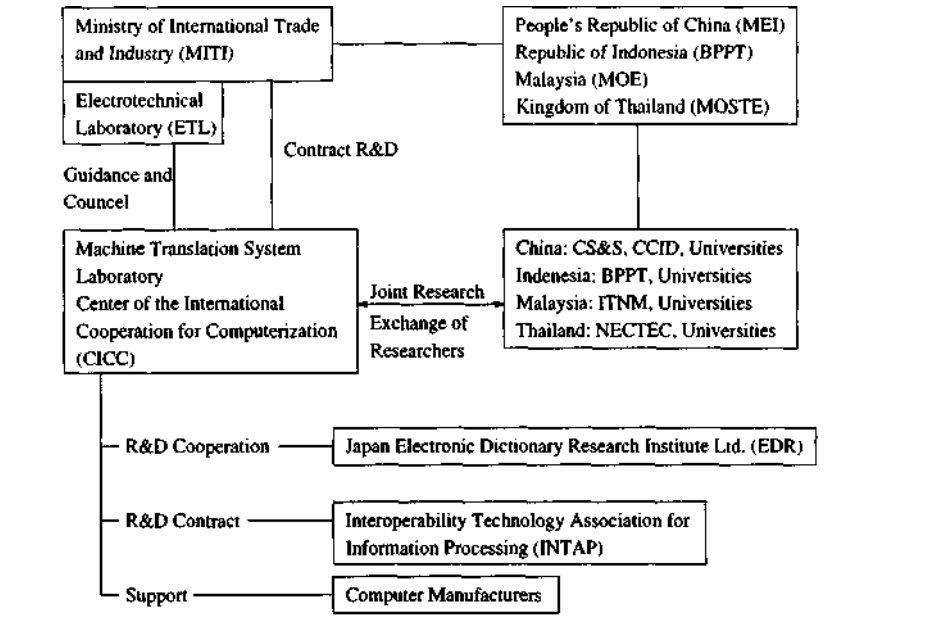

Figure 1: Collaboration in MMT project

The system used an interlingua as the intermediate representation for the languages of Chinese, Indonesian, Malay, Thai and Japanese. For each language, the analysis system for analyzing an input sentence to an interlingua representation and the generation system for generating a target language sentence from an interlingua representation were developed, see Figure 2.

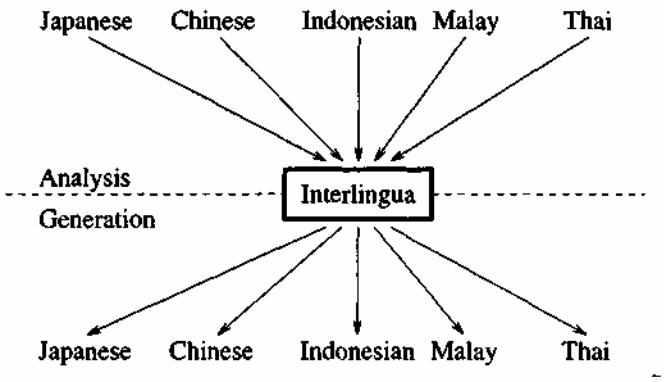

Figure 2: Interlingua-based MMT system

From the very different starting point of the technology of each country, the project shared the same theme of developing an interlingua. The project aimed at developing an universal knowledge representation scheme for all languages, at least for the above mentioned five languages. The

interlingua [4] is composed of two parts, namely, the part of concept classification and the part of concept relation. Within a language, it seems possible to define a set of concepts descriptively or by sets of their synonyms such as the synsets used in Wordnet. The problems occur when it needs to link between the concepts of different languages. The project used the set of concepts in EDR concept dictionary as the initial set for building the linkage between languages. Only about 25% of the total concepts can be linked between two languages and about 15% among the multiple languages. One of the typical problems of concepts linking across languages is the inequality of the concepts. It still needs further adjustment in the concept linking procedure.

Concerning the set of concept relations, the project proposed a set of 24 types of concept relations as the final version [3] of the interlingua. It was used in developing the MMT system. The mismatching between analysis and generation systems occurred when the definitions for the concept relations were unstable. The project had tried to re-adjust the set of concept relations and finally proposed a new set of 17 types of concept relations managed in a hierarchy. Together with the definitions, a set of case-frames to constrain the concept relations was also collected. With this approach, the concept relations can be fixed by the case-frames of each verb but the cost of collecting the prototypical case-frames of the verbs will be very high. In our feasibility study, we had prepared about 100 basic verbs of each language, see [4] for the details.

The project has been ended in the early of 1995. The interlingua approach looks like to be the appropriate solution for developing a multi-lingual machine translation systems. Idealistically, any target languages can share the resources by only having a link to the interlingua. But in practice, in addition to the problems on designing the interlingua as mentioned above, developing the multi-lingual machine translation system has its own particular problems:

- **Resource availability:-** This includes human and language resources. Human resource as the system developer is very limited. Grammar developers for MMT system indeed have to be fluent in the target language, otherwise, they will not be able to adjust the grammar to suit the interlingua. One reason is that we have no a good inference engine to handle the interlingua yet. In case of language resource, it needs a large amount of corpora for language study. Building a collection of language corpora and a large scale word dictionary have just been started. It is a labor intensive task because most of the languages in Asia still need keyboarding for data input.

- **System evaluation:-** It is hard to evaluate the system performance when it is a multi-lingual system. The evaluator has to be fluent in all the accounting languages. Most of the detail problems cannot be solved because they are unrecognizable by the evaluator.

3 State-of-the-Art MT R&D in Asia

it is not doubt to say that there is a technology shift occur in Asian countries after the MMT project. Especially in the MMT project participating countries, many MT related research activities are initiated and become significantly supported by the governments. Most of the countries started with a limited MT related researchers and language resources. Constructing a machine readable dictionary was one of the most labor intensive tasks. Though there were a lot of classical printed dictionaries available, number of word entries was very limited and most of the word information were omitted and left to reader's intuition. The project achieved not only in realizing the multi-lingual machine translation system but also building the infrastructure for MT R&D in this region. About 50,000 word entries of basic term dictionary and 25,000 word entries of information science technical term dictionary are available in electronic form for each language. In addition, corpora for grammar acquisition and evaluation are also gradually increased.

Thanks to the widespread availability of Internet in the recent days, the expectation from MT technology has been shifted to be more practical on networking environment and large amount of information. English language gains a high potential through the wide spread of computer network and has a great effect on the MT R&D in this region. Following issues are raised to note the remarkable paradigm shift in MT R&D especially in this region:

- **Knowledge representation to knowledge acquisition:-** Study of interlingua for the universal representation in knowledge-based MT was the main topic in the past ten years. The availability of information in the electronic form poses us another topic of knowledge acquisition. In addition to the research on knowledge representation, many recent research topics are conducted in the framework of statistic or probabilistic based approach and machine learning as well.
- **Multi-lingual MT to Bilingual MT:-** Gaining the potential from developing MMT system, many countries look for a more practical system to match the present needs of MT. With the growth of English language needs especially through the Internet, the priority of MT development is bent to the translation system between English language and their own languages.
- **Total to partial use of MT technology:-** Instead of applying the NLP technology for MT system as a whole, some partial technology are able to be applied directly to the language specific problems, for example, word processing, word segmentation, spell checker, grammar checker, etc.
- **Workstation to personal computer:-** The rapid growth of hardware performance has offered another choice of computer platform with a satisfactory cost performance for personal use. Many commercial systems are developed for windows system on personal computer platform in addition to the relative high-cost of workstations.
- **Translation accuracy to cost, speed and user interface:-** The high translation accuracy is not the only goal for MT development any more. Current technology assures the translation accuracy somehow. On the other hand, it seems to be difficult to make a significant different with the current technology. Responding to the present needs, other than the high translation accuracy, the system must be good in cost performance, able to translation in an acceptable speed and have a user-friendly interface.

3.1 Japan

As to the report [1] on commercial MT product at the end of 1996, there are 21 companies released 64 products with the price less than 1,000,000 yen (about 8,500 US dollars). The lowest price is as low as 6,000 yen (about 51 US dollars). The price about ten years ago was around 7,000,000 yen (about 60,000 US dollars) which was hardly compared with the present lowest price. According to the report in [1], the needs of MT in Japan are as high as the needs of word processor. But, because of the surprisedly high price of it and the reliability of translation result obstructed it from general use during the past years. Not only the low price which a company can offer, followings are also the support for wide spread of MT in Japanese market:

- Translation accuracy is raised to an acceptable level.
- Accompanied with the use of web browser saves the trouble in data inputting.

- Translation speed is acceptable with the present high-performance hardware.
- Translation is adjustable corresponding to the field of user's requirement.
- Fault tolerance according to the input in a mess.

It is noticeable that almost all of the mentioned products are for the translation between Japanese and English languages. Few of then are for European, Korean and Chinese languages.

Other than the above home-user oriented MT systems which have to be low-price and compact, there are also some remarkable middle range systems developed and used within the organizations. MT system developed in NHK is used for broadcasting related translation such as news capture, superimposition and so on. The system has to be able to handle various fields of contents with the frequent used of proper noun. NTT developed a system called ALT-J/E focusing on translating in the most natural way. It introduced a case-frame transferring method restricting the translation selection of the target language. ATR is aiming at developing a speech-to-speech translation system which is expected to be able to support in interpretation in the international conference or telephoning. JICST, having a new organization called JST, succeeds the MT system of μ project providing abstract translation service for Japanese research papers.

The needs of English to Japanese language translation can be obviously observed in the JEIDA (Japan Electronics Industry Development Association) investigation on the requirements for English use on the Internet in Japan. The committee on text processing technology at JEIDA is a subcommittee of JEIDA's committee on natural language processing, and has been developing its bilingual corpus for research on machine translation systems since the 1996 Japanese fiscal year [6]. The committee has surveyed the present situation and requirements for information use on the Internet using a questionnaire.

To accurately analyze the needs of new users of the computer network according to the contents and the uses of the Internet and assistance functions, the committee designed a questionnaire on (1) the use of document data, (2) electronic newspapers, (3) digital libraries, (4) English use on the network, and (5) information retrieval. The survey was done from April 1996 to May 1996. Of the 214 answers, 50% were gathered from engineers or scientists and 35% were gathered from teaching staff and students. All questions in JEIDA's questionnaire, results and the analysis can be found in the JEIDA homepage, http://www.jeida.or.jp/committee/textsyori/sec-0.html.

Concerning the results of asking, “How do you use English on the Internet?”, we can summarize the analyses as follows:

- Ninety percent of the answers stated they used the WWW for information gathering and 80% of the WWW users accessed it several times a week. This indicates that due to the spread of the WWW, the frequency of the access of information written in English is increasing. On the whole, the access for information gathering is much more frequent than the access for information dissemination in Japan.
- As for the purpose of using English, “to understand the outline of the next” got the highest score. People tend not to and do not want to simply surf on the network. They gather necessary information from the huge amount of information written in English.
- As for the things which are inconvenient for the users concerning using English, only less than 15% say that there is nothing which makes them feel inconvenient.
- As for the way of solving the inconveniences, printed or electronic dictionaries are frequently used; however, the percentage of the persons who use machine translation systems is less than 10%.

Concerning machine translation systems, almost half of the answers state the experience in using machine translation systems. Though this survey was done one and a half years ago when there were not so many low-cost machine translation software as in present, the percentage of having ' experience in using machine translation systems was higher than we have expected.

Of the people who had used machine translation, no one replied that “nothing is dissatisfied (or completely satisfied)”. This embarrassed us because it implied that machine translation systems have not reached a level at which users are fully satisfied with them.

Because many users come across English text on some occasions on the Internet and almost all of the users experience difficulty in using English, the demand for machine translation systems is rather high. Actually, the percentage of users satisfied with the assistance functions of English sentence reading for understanding outlines is high. At the same time, because printed or electronic dictionaries are used to resolve the problems with English, current machine translation systems are insufficient even for "assistance". We have to identify the dissatisfied items for each user on each occasion which English is necessary, and identify the user needs for machine translation systems.

3.2 China

Other than the R&D for the project of MMT, many bilingual MT projects, such as English-Chinese, German-Chinese, Russian-Chinese and Japanese-Chinese were parallel conducted. The development for a particular pair of languages seems to be more practical and straightforward to the interesting language pair.

Chinese government plays an important role in supporting the national MT R&D project. The support focuses on the follow-up research of the MMT project in both improving the translation quality and the evaluation method. They aim to make the system possible for practical use. The government also pays attention on constructing bilingual MT systems for such the pairs of languages between Chinese and any of English, Japanese or German. Above that, dynamic translation through the network is also within the scope of supporting.

MT R&D activities increase in recent years. It can be observed from the increasing of the number of research institutes. Most of them are aiming at developing a practical bilingual MT. Some are extending the achievement of MT R&D in part to other applications, such as word segmentation, part-of-speech tagging, identifying phrase boundary and bilingual parallel text alignment.

Regarding the commercial activities of MT, there are at least three systems on the market at present. All of them are English to Chinese MT systems which can work on personal computer platform. The unsatisfactory translation accuracy (70-80%) is the main reason of suppressing the MT market of the country.

3.3 Indonesia

BPPT (Agency for the Assessment and Application of Technology) of Indonesia individually expands the achievement of MMT project by developing a tool for linguistic research called LOP (Linguistic Operation Panel). The LOP is developed for working on PC unix based machine. The system provides a user friendly interface for supporting the development of grammar and dictionary. The prototype of MMT system is now available through network service.

Though there is no commercial MT service at present, the MT related technologies are applied in some applications such as hyphenation function in word processor, grammar checker and so on.

It is also reported that English language is the highest requirement for MT development. Many other MT R&D activities are conducted in the universities and most of them are concerning Indonesian-English MT.

3.4 Malaysia

ITNM (Malaysian National Institute of Translation), the previous partner of the MMT project, has re-organized itself to focus on translation services. MT R&D activities are then transferred to other research institutes of the universities. There are some NLP related research focusing on constructing bilingual dictionaries (Malay and English) and thesauri.

Many commercial products are coming up with the applied NLP technology for example spell checker, grammar checker, word processor, bilingual dictionaries among Malay, English, Chinese and French, and so on. There is also a release of talking dictionary for Malay, Chinese and English.

Though there is no any evidence of MT R&D, the requirement of NLP technology is obvious through the wide spread of dictionary development and grammar research.

3.5 Thailand

LINKS (Linguistic and Knowledge Science Laboratory) of NECTEC (National Electronics and Computer Technology Center) has prepared Thai language resources for studying language processing to develop either a whole system of MT or NLP applications. So far it has been reported, there are two projects to develop English to Thai MT systems. One is conducted in LINKS and another one is the collaboration project between a university and an organization from Singapore. However, there is no commercial plan for any MT development yet.

LINKS in collaboration with CRL (Communications Research Laboratory) of Ministry of Posts and Telecommunications of Japan, have constructed a tagged corpus for Thai language, called ORCHID corpus. The corpus is tagged with its original part-of-speech tagset which is the result from improvement after using the tagset in the MMT system. The corpus is consisted of about 2MB (or about 400K words) of the proceedings of NECTEC annual conference. It is scheduled to release in the end of 1997.

The widely use of computers in the present day stimulates the needs of standard processing for Thai language. As a national project, NECTEC has set up a plan for developing Thai standard software library. It covers all the fundamental technologies for processing Thai language, including character coding, font design, word segmentation till voice synthesis and recognition.

3.6 Singapore

ISS (Institute of Systems Science) of National University of Singapore makes a significant progress in development MT for English, Chinese and Malay languages. It is a transfer-based MT which is now available for service. They also plan to increase the language pair for translation in the future. Besides the successful in bringing MT into service, ISS has also developed a multi-lingual supporting system called MASS (Multi-lingual Application Support Service) which is now available under both Unix and Windows environment.

Many other kinds of multi-lingual processing projects are ongoing in the institute. The service through Internet is also within the scope of development.

3.7 Korea

Many efforts can be seen in the development of Japanese-Korean MT systems. There are 5 systems released in the market, only one system is for the mainframe computer and other four systems are for personal computer environment. All systems are developed using the advantages in the similarity of languages. Actually it is so, but to gain the more accuracy in the translation, perhaps it needs to study more about the difference between the languages [2].

Other than the above countries, there are also MT R&D activities conducted in the areas such as in Hong Kong and Taiwan. India also has MT systems for translating between Hindi and English.

4 NLP Related R&D and the Available NLP Resources

Number of research papers and conferences show that research activities in this region increase year by year. The conferences or workshops are held both locally and regionally. Followings are not all but some of the remarkable conferences held in this region:

- NLP, The Association for Natural Language Processing, Japan, held annually since 1995.
  http://www.kyutech.ac.jp/nlp/index.html
- PACLING, Pacific Association for Computational LINGuistics, Japan-Australia, held in every 2 years since 1989.
- AAMT, The Asia-Pacific Association for Machine Translation, Japan, MT activities supporting journal.
  http://www.jeida.or.jp/aamt
- ROCLING, Research on Computational Linguistics, Taiwan, held annually since 1987.
- SNLP, Symposium on Natural Language Processing, Thailand, held in every 2 years since 1993.
- NLPRS, Natural Language Processing Pacific Rim Symposium, Region, held in every 2 years since 1991.

Some recent activities can be seen in standardizing for data exchangeability, such as, UPF (Universal Platform), an AAMT activity aiming at setting up a standard format for dictionary encoding, and MLIT (Standardization of Multi-lingual Information) aiming at setting up a standard for character codes, keyboards, input/output methods, fonts and so on.

NLP Resources Some electronic dictionaries and corpora are available for research use. Followings are some collections of the resources. It can be directly approached to each site for the details.

- CICC, the achievements from MMT project. Besides the technical reports of interlingua and the MMT system, there are also electronic dictionaries and corpora available for the languages of Chinese, Indonesian, Malay and Thai. For details see
  http://tyo-cc-server.cicc.or.jp/homepage/english/about/act/mt/mt.htm.
- EDR, electronic dictionaries and corpora are available for both English and Japanese languages. In addition to word dictionaries, there are also co-occurrence dictionaries and concept dictionaries available in a very large scale.
  For details see http://www.iijnet.or.jp/edr.
- IPA, both Japanese electronic dictionary and corpus are available.
  For details see http://www.ipa.go.jp/STC/NIHQNGO/IPAL/ipal.html.

Many other resources can also be found developing for NLP research such as ATR, RWC and so on. It is evident that NLP research in this region will be increasing and continues making a great contribution for the NLP community.

5 Conclusions

It is evident that English language gets involved in MT R&D in every Asian country. The needs of using English language especially increase when Internet can be freely connected from almost every where. Most of the information flowing on the Internet is written in English. To absorb the flood of information from the Internet, we need a sort of efficient tools to extract what we want and provide us in the message that we can understand in quick. Translation from English language to our mother tongue is highly required in this region. Though the accuracy of translation is hardly acceptable without human interaction, other factors of such low-price, high speed translation or user-friendly interface have made MT widely used as being observed in Japanese market. Countries other than Japan are now starting to develop bilingual MT for English language after gaining potential from the MMT project in the past ten years.

NLP resources such as online dictionaries and language corpora are also increased year by year. It is expected to be another valuable resources for language study and MT research in the future. Serving online dictionary through the Internet can be another element that fills the needs of MT for both gathering and dissemination of information.

6 Acknowledgments

We would like to thank Dr.Hitoshi Isahara of CRL for the helpful JEIDA's analysed results and Virach Sornlertlamvanich, a Ph.D. student of our laboratory, for valuable supports in the study of R&D status in this region.

References

[1] AAMT. 1996 *AAMT Journal,* The Asia-Pacific Association for Machine Translation, No.16.

[2] Choi, K. S. and Kim, T. W. 1996 Present Status of Japanese-Korean MT Systems and the Analysis, *Proceedings of The Second Annual Meeting of the Association for Natural Language Processing,* pages 433-443, (in Japanese).

[3] CICC. 1995 *Technical Report on Interlingua Final Edition,* 6-CICC-MT36, Center of the International Cooperation for Computerization.

[4] CICC. 1997 *Technical Report on Interlingua and Concept Classification,* 8-CICC-MT31, Center of the International Cooperation for Computerization.

[5] Komurasaki, M. 1994 Profile of International R&D Cooperation Project on Multi-lingual Machine Translation (MMT) System, *MMT'94,* Center of the International Cooperation for Computerization, pages 11-23.

[6] Isahara, H. 1997 The Present Situation and Requirements for Information Usage on the Internet - Based on a Questionnaire by JEIDA -, *MT Summit VI.*

[7] Nagao, M., Tsujii, J. and Nakamura, J. 1985 The Japanese Government Project for Machine Translation, *Computational Linguistics 11,* 2-3, pages 91-110.