MT Research in Thailand and Linguistics and Knowledge Science Laboratory (LINKS)

It was more than 40 years ago that researchers in Thailand had revealed their interest in putting an ability of processing Thai language onto a computer system. The attempts have been first started on the development of a word processor for Thai together with a draft character code assignment. The development have been done on the mainframe computer and with the great efforts of the researchers from various fields, we now feel no inconvenience in creating or reading Thai document through computer terminals.

The first project on development of Machine Translation system (MT) for English-Thai translation was conducted by a group of Thai researcher with the technology supports from Grenoble University of France. The project, known as ARIANE project, started in the year of 1981. The peculiarity of Thai language bottlenecked the progress of project and left many language specific problems to be overwhelmed. At the same time, the research on modern language among the linguists have partially proposed some systematic models of Thai. Though we have no language model to represent Thai exhaustively at the present time, a lot of researchers strongly express their enthusiasm to realize a computer system that has the ability in parsing and generating any of Thai sentences.

To be a part of the Multi-lingual Machine Translation system (MMT) for Asian languages, National Electronics and Computer Technology Center (NECTEC) and Center of The International Cooperation for Computerization (CICC) agreed on the cooperation in research and development of machine translation system for Thai. NECTEC and CICC launched the project in 1987 and expected to a system which can translate among five languages of Thai, Chinese, Malay, Indonesian and Japanese through an designated intermediate repre sentation called Interlingua. We are assured that the Interlingua translation method is plausible for designing a system for multi-lingual machine translation. The project is scheduled to be ended in the beginning of 1995.

The project started with anxiety about the research background of Thai language and the infrastructure of Thai researchers in the field. Computers and linguists from universities conjoined to discuss on how to prepare linguistic data and conform it to the parse algorithm. The research and development plan has been detailed into the components of Thai sentence analysis, Thai sentence generation, Thai electronic dictionary for basic term and technical term, Translation support and Thai input/output system and Integration system. With the research cooperation of the King Mongkut's Institute of Technology Thonburi (KMITT) and the King Mongkut's Institute of Technology Ladgrabang (KMITL) the translation system is gradually improved quantitatively and qualitatively.

Thanks to the NECTEC-CICC's MMT research and development project, a lot of research institutes and universities actively conduct the research on natural language processing. Many research focus on how to parse Thai sentence such that Chulalongkorn University develops CUPARSE for Thai syntactic parser with the dependency grammar approach, AIT develops a syntactic parser based on the GPSG, etc. Others are the research on some specific characteristics of Thai language. A lot of works on Thai language are paving the way to assure the possibility in building a large scaled Machine Translation system. In the very near future, we hope that various kinds of MT which can handle Thai sentence as a source or a target language will be unveiled.

Linguistics and Knowledge Science Laboratory, known as Links, is founded as a national research laboratory in 1992. Links, a laboratory in NECTEC, is fully subsidized by Thai government to conduct the research in the field of natural language processing and applied AI. Moreover, Links also makes up the research strategy which itself will be the information center providing resources to stir up the research activities in Thailand.

The goals of Links are as followings

1. To study and investigate the research in information processing, NLP, Large-scaled database for instance.
2. To research in NLP and AI to do feasibility study of system development.
3. To join the research with other research institutes or universities.
4. To join the system development with private sectors hence, the private sectors will gain the potential in frontier technology development.
5. To be a center for providing facilities, resources and funds for advanced research.

In MT development, Links joins the NECTEC-CICC's MMT project, guides the development for Thai language system to be able to connect with other 4 languages through Interlingua. With the great support from CICC, the system can handle some specific Thai sentences which are carefully selected to be a set of corpus sentence for system development. We are now making the improvement and expecting an usable MT system. Therefore, the MT system will be a tool to prepare the information for data exchanging among the participat ing countries.

In the developing process, we prepare a set of corpus sentence to evaluate the accuracy of translation result. This corpus base as well as text base, word dictionary base and grammatical rule base are considered to be a great resources for further research of the related field. In the final state of development, we plan to do system evaluation regularly by considering the actual demand of document translation.

The MT technology is planted in these recent years, computers and linguists work closely to resolve the problems occur in both of computer and linguistics. On top of computational linguistic problems in general, we cannot easily recognize Thai text into a sentence each because we are not accustomed to writing a text sentence by sentence. This causes problems in word and sentence recognizing process. We do not have an effective algorithm to perform fully automatic word segmentation. We are still facing the difficulty in expressing the syntactic structure. These are the fundamental issues that we have to go over to reach the next step of development of MT.

Mr. Virach Sornlertlamvanich
Chief of Linguistics and Knowledge Science Laboratory (LINKS)
National Electronics and Computer Technology Center
Ministry of Science, Technology and Environment
Email: virach@nwg.nectec.or.th