ParSit: Online English-Thai Machine Translation Service

Virach Sornlertlamvanich, Paisarn Chareonpornsawat and Monthika Boriboon

The growth of Internet connections is playing an important role in facilitating human-human, human-computer and computer-computer communication throughout the world. It seems that information as a product of human intelligence can be created and exchanged without limitation on the infrastructure called Internet. This is completely correct if we all speak in the same language. The language barrier problem becomes significant when we do not understand the language that is widely used.

Thailand being one of the non-English-speaking countries, to help the Thai people in assimilating the huge amount of information scattering over the Internet, National Electronics and Computer Technology Center (NECTEC) has spent great efforts to develop an English-Thai machine translation system. The first trial of an automatic English-to-Thai machine translation service, called ParSit, has been put onto the web. ParSit enables the Thai people to surf English webpages as they are browsing the web-pages in the Thai language. Through the ParSit service, the requested English web-pages are automatically translated into Thai by keeping the web-page original layout. The current ParSit trial service can be accessed via one of the following URL's http://go.to/parsit, http://i.am/parsit and http://www.links.nectec.or.th/services/parsit. ParSit has been developed under the collaboration between NECTEC, Thailand and NEC Corporation, Japan. Thai to English and other multilingual translation systems are under development. We are aiming at establishing a cooperative language translation service among the non-English-speaking countries to play an active part in reducing the digital gap that may occur in the future.

How did Machine Translation Come to Thailand?

The first version of Machine Translation (MT) system was based on a direct translation scheme with a limited syntactic grammar. The mechanism was quite simple—simply replacing the equivalent words in the source language to produce the target language. It works for the language pairs that have similar grammar and word formation rules. In 1966 ALPAC report, the Automatic Language Processing Advisory Committee concluded that the US government should not further invest in MT because it was slower, less accurate and twice as expensive as human translation. Although MT in the US was deterred by the report, MT activities still gradually progressed outside the US for decades. Many new MT approaches have been proposed such as syntax/semantic transfer method, interlingual method implemented in rule-based, example-based, statistical-based and corpus-based paradigms. The system architecture was also expanded from bilingual to multilingual architecture with the reason of the cost efficiency in dictionaries and grammar preparation. In bilingual architecture, all pairs of languages have to be prepared while each single pair of target language and the common interlingual is needed in the multilingual architecture. The advances in computer technology and computational linguistics research have made great contributions to overcome the MT obstacles. However, the best task that current MT systems can undertake has been limited to technical text and sub-language domain.

In Thailand, with technical support from Grenoble University of France and auspices of the Ministry of the Universities' Affairs, the English-to-Thai machine translation emerged in 1981. It is reported that the project ended up with a prototypical system. In 1986, the Multilingual Machine Translation for Asian Countries project was initiated by CICC (the Center of International Cooperation for Computerization), Japan. It was the internationally collaborative project between Japan and four other Asian countries, namely, Indonesia, Malaysia, People's Republic of China and Thailand. The interlingual (IL) approach was considered as an appropriate technique for realizing a multilingual machine translation system efficiently and providing a common research topic for researchers. At the onset, the research partners other than Japan were inexperienced and had very few researchers in the field. Thailand was no exception. With the financial and technical support from Japan, MT technology in the region was established. Not only did the project result in the first workable IL-based MT system, but also encouraged the research and application of technology in the field. The number of research projects has become significant and many language resources (such as electronic dictionaries, grammars, corpora, etc.) have been created. Many language processing algorithms and approaches such as word segmentation, full text search, word processing and the like have been proposed. These fundamental resources resulted from the efforts in Natural Language Processing (NLP) and Computational Linguistics (CL) research which are now the crucial tools and knowledge for processing the huge amount of information existing in the Internet.

Thanks to the research experience in the Multilingual Machine Translation project, NECTEC has succeeded in implementing the first English-to-Thai machine translation service on the web, called ParSit. Through the intensive collaboration with NEC Corporation, NECTEC developed the English-to-Thai machine translation on the NEC's machine translation engine. On 24 June 2000, ParSit was launched to provide automatic translation for both web-pages and short messages. The service has been well accepted since then and now we are planning to develop a Thai-to-English translation system which will support the Thai people in disseminating original information to the world.

What is "ParSit"?

ParSit (pha:-sit) is a Thai word that means a traditional saying; expressing common experience or observation from generation to generation. The automatic machine translation is one of the ultimate goals of humankind. We have been told from generation to generation that one day we could overcome the language barrier since the era of Babylon. On the other hand, ParSit is coined to rhyme with "parse it" which means to analyze (a sentence) in terms of grammatical constituents, syntactic relations, etc. Consequently the English-to-Thai machine translation system is named ParSit.

ParSit Inside-Out

The technology for developing ParSit is based on an English-to-Japanese machine translation system developed by NEC Corporation, Japan. ParSit does not translate word by word, but sentence by sentence. ParSit employs the word syntac-

Web-page translation is a service that converts an English web-page into Thai web-page by keeping its original layout. To use this service, a user simply goes to the ParSit top page (http://www.links.nectec.or.th/services/parsit), clicks the 'Enter' button and then clicks the 'WEB translation' button. The user inputs the URL of the desired English web-page in the provided text box then clicks the 'send URL' button and waits. In a few seconds the user will get the web-page with all messages written in Thai except for the images.

Text translation is a service that converts a short English message into Thai. Instead of clicking the 'WEB translation' button, click the 'TEXT translation' button to get to service. There is a text box in the middle of the page. This text box accepts up to 2,000 characters for each time of the translation request. The text length is limited to avoid a long wait for the response. Click 'Translate' button to execute the translation and then wait for the response. The user will shortly get the translated Thai text.

tic and semantic information expressed in grammatical rules and dictionaries to analyze a source language (English) sentence and consecutively generate the target language (Thai) sentence.

Basically, there are four main steps in translating a sentence. First, ParSit looks up words for the syntactic and semantic information from the dictionaries. One word can have a lot of meanings and usages. They all are kept in separated records to maintain the possibilities in interpretation. Secondly, ParSit analyzes the sentence syntactically and semantically to determine the most appropriate interpretation of the sentence. As a result the intermediate representation for that sentence is created. Thirdly, the generation module converts the intermediate representation into the corresponding syntactic structure and selects the appropriate Thai word to express each concept. Finally, ParSit forms a Thai sentence according to word usage under Thai grammar. One of the most difficult tasks is to select the appropriate word interpretation and generate the natural expression in the target language. For example, 'a plane' can be interpreted as a vehicle in which we fly; or a carpenter's tool for smoothing wood; or a flat surface, depending on its context used in the sentence. 'I' is not always the first singular pronoun to refer to oneself as speaker, but it can be interpreted as the symbol for iodine; the first item, etc.

ParSit provides two types of translation services—web-page translation service and text translation service as above.

Some Statistical Views of the ParSit Use

There have been more than 380,000 visitors since the first date of the service in July 2000 until January 2001. The average number of about 2,000 visitors or about 7,000 translation pages per day is good evidence that ParSit has made a significant contribution to Thai society. According to the feedback from the user comments, though the translation quality produced by ParSit is not as good as that of human translation, it is widely accepted as an essential tool for browsing the web-pages written in English. ParSit facilitates a quick information look-up for those who are not familiar with English.

Bridging the Information Gap

Presently, it is reported that more than 65% of web-pages are written in English. Some of the web-pages prepared in Thailand are also written in English without Thai translation. Though the Internet population in Thailand is increasing, the barrier in information access is still high. It is not so fluent at reading messages written in foreign languages. The time consumed in information access could possibly reduce productivity in this information age. It becomes more significant when the speed of Internet connection is getting faster and faster. To bridge the information gap, the machine translation system, like ParSit, is an inevitable tool to help users in making a sketch of the contents. Though the translation result is hardly perfect, owing to the unsolvable problems with the current language processing technology, ParSit can somehow facilitate users in surfing the web with little skill in reading the English text. According to some users' opinion, rewriting the output from ParSit translation into a fine Thai text consumes less time than translating from the original text. It has shown an extensional use of ParSit in preparing Thai documents from the original English text under the Internet environment. ParSit has proven the possibility to be one of the efforts in bridging the information gap between Thailand and the world. The service quality keeps improving in the speed of response time and the legibility of the translation results as well. In parallel, we are preparing for another challenge in developing a Thai-to-English translation system which will help us to publish and send Thai information out to the world. It is our supreme goal to take a part in bridging the information gap for the countries in this region and assisting the creation of the knowledge-based society in the coming era.

Virach Sornlertlamvanich (D. English) is the Director of, **Paisarn Chareonpornsawat** (M. English) and **Monthika Boriboon** (M. Lingistics) are researchers in Information R&D Division, NECTEC. Their research interests are natural language processing, machine translation, information retrieval, etc. ParSit is one of the research projects that has recently been successfully launched for public use. They started the project with their experiences in joining the Multilingual Machine Translation Project (1986-1996). Many basic research issues have been conducted in the division to prepare the basic components for developing the machine translation system, e.g. a large electronic dictionary with full coverage of word syntax and semantic information, morphological and grammatical rules including the user-friendly interface working under the Internet environment.

Information R&D Division, National Electronics and Computer Technology Center (NECTEC), National Science and Technology Development Agency (NSTDA), Ministry of Science, Technology and Environment (MOSTE), 539/2 Sri-Ayudhya Road, Ratchathewi, Bangkok, 10400, Thailand, phone: (66) 2 642 5001,10 ext. 301, fax: (66) 2 642 5015, http://www.links.nectec.or.th/, http://www.links.nectec.or.th/virach