PARSIT^{TE}: Online Thai-English Machine Translation

Teerapong Modhiran, Krit Kosawat, Supon Klaithin,
Monthika Boriboon and Thepchai Supnithi
Text Processing Section, Division of Research and Development on Information,
National Electronics and Computer Technology Center
112 Paholyothin Road, Klong 1,
Klongluang, Pathumthani, 12120, Thailand
{teerapon.modhiran, krit.kosawat, supon.klaithin,
monthika.boriboon, thepchai.supnithi}@nectec.or.th

Abstract

This paper presents an online Thai-English MT system, called PARSIT^{TE}, which is an extension of PARSIT English-Thai one. We aim to assist foreigners and Thai in exchanging more easily their information. The system is a rule-based and Interlingua approach. To improve the system, we concentrate on pre-processing and rule analysis phases, which are considered necessary because of some specific problems of Thai language.

1 Introduction

Research on machine translation in Thai has been started since 1980. The first English-Thai one was set up under ARIANE project, followed by five Asian languages machine translation joint research under CICC project. After having finished these two projects, we had a prototype of machine translation system and a basic knowledge of Thai language.

Knowledge-based MT [7] is an approach that uses linguistics knowledge in the form of rule. The experience from two mentioned projects helped us in developing PARSIT, an online English-Thai MT system [5, 15], under the collaboration with NEC corporation Japan in 2001. The statistics showed that using PARSIT as a translation tool is gradually increasing. Currently, there are about 1,000 users and up to 10,000 translated pages per day. Thai-English MT system is one of the most requests in our web board.

The important phases for developing Thai-English MT system are morphological analysis and rule analysis. Unlike English, which has concrete separators for sentence and word boundaries (period and space), Thai has nothing. Furthermore, there are lexicon ambiguities in Thai, due to some characteristics such as no inflection, no article and no singular-plural form. These cause a difficulty in analyzing rule.

This paper illustrates the rule-based approach for Thai-English MT system. Section 2 explains the major problems of development. Section 3 shows its architecture. Section 4 talks about the implementation. And finally, discussion, conclusion and future work are presented.

2 Problems in developing Thai-English MT

In this section, we illustrate some problems of Thai language in machine translation viewpoint.

2.1 Word segmentation problem

Word segmentation is a major problem for languages that have no word boundary, such as Thai, Japanese, Chinese, etc. For example, the string "ตากลม" in Thai can be defined in two senses: "ตา-กลม (ta-klom)" means "round eye" and "ตาก-ลม (tak-lom)" means "expose to the wind". The correct sense can not be determined by the string itself. It is necessary to know the context around it. There are some researches on word segmentation which mainly focus on statistical based approach. Currently, the accuracy of word segmentation is around 95-99% [18].

2.2 Sentence segmentation problem

Since Thai has no explicit sentence boundary, Sentence Segmentation is raised as one of problematic issues. Space can be determined as a hint for sentence marker, but it has more than one function [2]. There are very few researches on sentence segmentation [4, 6,9]. Most of them apply machine learning to determine whether a space is functioned as sentence segmentation marker or not. Currently, the accuracy of sentence segmentation is about 89% [19].

2.3 Lexicon ambiguity

In Thai, one word may have more than one meaning and/or one syntactic category. A word like "ที่" has 6 categories: noun, pronoun, conjunction, classifier, preposition and prefix. It is difficult to handle this problem in rule analysis phase. In our approach, the disambiguating rules are abstracted from the restricted word position in sentence. For example, in the sentence "คุณลุง มา จาก บ้าน (uncle come from home)", as shown in figure 1, the words "คุณลุง (uncle)" can be categorized as "common noun" or "title noun", "มา (come)" as "active verb" or "auxiliary verb", and "จาก (from)" as "active verb" or "verbal preposition", respectively. We reduce lexical ambiguity by applying analysis rule that concentrates on the position of words in sentence.

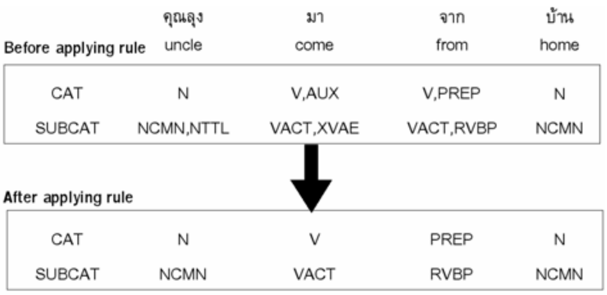


Figure 1: Example of lexicon ambiguity

3 PARSIT$^{\text{TE}}$: online Thai-English MT system

We apply the same approach that we have developed in English-Thai MT system to Thai-English one. This system is the first Thai-English MT that will be provided to public users.

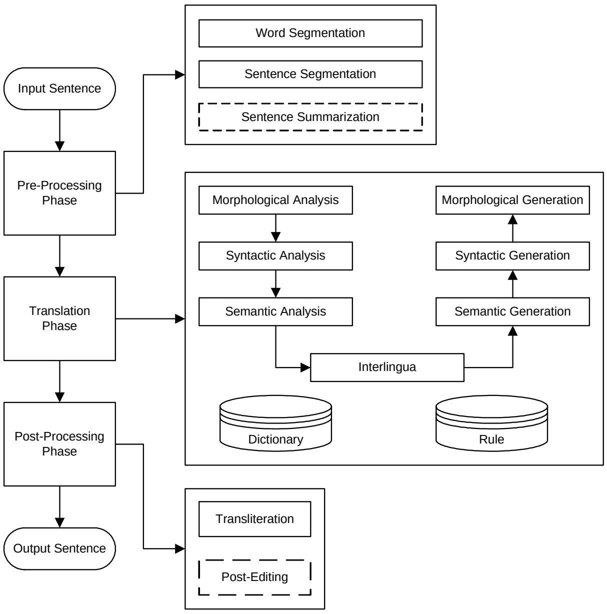


Figure 2: Architecture of PARSIT$^{\text{TE}}$ system

3.1 System overview

Figure 2 shows the overview of architecture for Thai-English MT system. The dash-boxes are the modules that we will develop in the next step. Our system is composed of three main parts.

Firstly, pre-processing phase which is assigned for preparing Thai input, because Thai has no sentence and word boundaries. Currently, we have developed word and sentence segmentation modules to handle these tasks. This problem, however, can not fully be solved by word and sentence segmentation. There are a lot of redundancies in Thai. If we can extract the important information from each sentence, it will help us translate more correctly. In the future, we plan to apply Thai sentence summarization to this system.

Secondly, translation phase which is composed of two main parts: analysis module and generation module. Analysis module is assigned for interpreting Thai sentence to Interlingua representation. Generation module is assigned for interpreting Interlingua representation to English sentence. We apply dictionary, rule database and linguistics information such as verb pattern, etc., in translating process.

Finally, the last phase is preparing formally English sentence. We use transliterate module to convert Thai unknown words (as a sound) to English words (i.e. “สมชาย” to “Somchai”). Since the current version is assumed to have lower accuracy than English-Thai MT system, we plan to add the post-editing module to improve the translation result.

In order to increase the accuracy of translation, which may be decreased from preprocessing phase, we have a choice for users to translate in automatic or semi-automatic mode, where users can verify the output results from word segmentation and sentence segmentation before sending them to translate.

The detail of each module is explained in the following sections.

3.2 Pre-processing module

Pre-processing module is assigned to solve the problem of no-word boundary and no-sentence boundary characteristics. We start from doing word segmentation and then using the output to apply to sentence segmentation module.

3.2.1 Word Segmentation module

Word segmentation is a prerequisite for developing Thai-English MT. In our system, we apply SWATH which is an automatic Thai word segmentation tool [8]. SWATH has been developed in RDI laboratory. It has two automatic functions: automatic word segmentation and automatic Part of Speech (POS) tagging. It provides longest matching technique [17], maximal matching technique [14] and POS bigram technique [3, 10] for developers. Developers can apply POS bigram technique, if they want to get the automatic POS tagging. Our system uses longest matching technique for segmenting words.

3.2.2 Sentence segmentation module

Sentence Segmentation is also a prerequisite for developing Thai-English MT. We have applied machine learning (SVM algorithm [1]) to train our sentence segmentation model. To increase its accuracy, we analyse words that have potential to be clue words for sentence boundary such as “อย่างไรก็ดี (however)”, “ยิ่งไปกว่านั้น (moreover)” and so on. The sentence segmentation algorithm requires segmented words as learning information.

3.3 Translation module

Translation module is composed of two main modules: analysis and generation as shown in figure 2. The functionality of each module is as follows:

1) Morphological Analysis module: this module serves as a pre-process of translation. It includes dictionary loading and some morphological disambiguation. For example POS disambiguation, Thai Unknown Word, Word sense disambiguation.

2) Syntactic Analysis module: this module serves as a syntactic level analysis. It analyses a linear sentence and construct a syntactic tree.

3) Semantic Analysis module: this module serves as a semantic level analysis. It analyses a sentence structure and construct a semantic tree.

4) Semantic Generation module: this module serves as a semantic level generation. It interprets the meaning from Interlingua.

5) Syntactic Generation module: this module serves as a syntactic level generation. It interprets the grammatical structure from Interlingua.

6) Morphological Generation module: this module serves as a post-process of translation. It includes reordering module.

Since the generation module is about generating English sentence from Interlingua, we can reuse the generation module from Japanese-English MT system (as shown in arrow 2 in Figure 3). So, the main point of this paper focuses only on the analysis module. We will explain the Thai analysis rule and dictionary, which are the crucial part for developing Thai-English MT system.

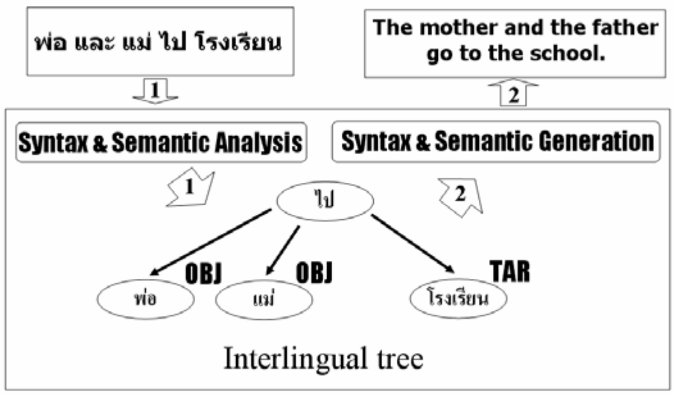

พ่อ และ แม่ ไป โรงเรียน
The mother and the father go to the school.
1
2
Syntax & Semantic Analysis
Syntax & Semantic Generation
1
2
ไป
OBJ
OBJ
TAR
พ่อ
แม่
โรงเรียน
Interlingual tree

Figure 3: Example of Analysis and Generation.

3.3.1 Thai analysis rule

Thai analysis rule has two main functions: 1) syntactic analysis is assigned for the obligatory relations (the relations between verb and its arguments) in terms of grammatical rule 2) semantic analysis is assigned for mapping the output from 1) to semantic case relation.

We apply both top-down and bottom-up parsing for Thai analysis rule. Top-down parsing is applied to generate all possible alternatives. It generates the alternatives under the restriction on related lexicons in a sentence. The bottom-up parsing will be processed when all possible alternatives have been generated. It will conduct the elimination of flawed alternatives or selection of the effective ones, and finally select the best solution. The grammar rules in parser are shown as follows [12]:

```
- top-down method
    S <-- NP VP
    VP <-- V NP PP ADV
    VP <-- V NP S'
    VP <-- V NP
    VP <-- V
- bottom-up method
    V <-- LAUX V RAUX
    V <-- LAUX V
    V <-- V RAUX
    NP <-- N NUM CLAS DET
    NP <-- N VATT CLAS DET
    NP <-- N CLAS DET
    NP <-- N DET
    NP <-- N CLAS VATT
    NP <-- NP PP
    NP <-- NP CONJ NP
    PP <-- PREP NP
    S' <-- S
```

Since Thai has lexicon ambiguity problem. In our approach, the disambiguating rules will be abstracted from the restricted word position in sentence. From the example shown in Figure 1, the sentence "คุณลุง มา จาก บ้าน (uncle come from home)" can be disambiguated by applying three rules: 1) rule for determining between "common noun" and "title noun" for the word “คุณลุง”; 2) Rule for determining "active verb" and "auxiliary verb for the word “มา”; 3) Rule for determining "active verb" and "verbal preposition" for the word “จาก”, respectively.

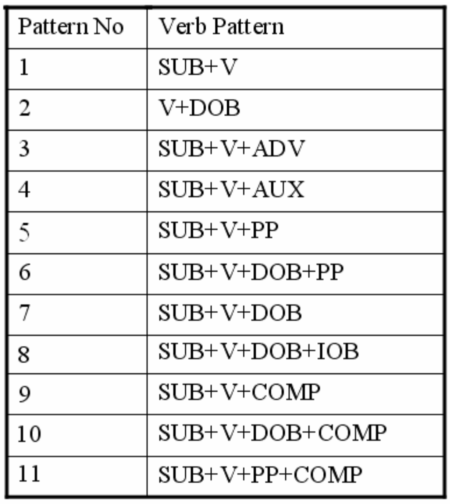
3.3.2 Dictionary

Dictionary is an important component in a translation process. Each word in dictionary may have more than one entry, corresponding to the meaning of word. The information in each entry is divided into three types:

1) Analysis Phase information: the information that relates to Thai words, such as POS [12], Verb pattern (as shown in Table 1).

2) Concept information: the information that relates to Interlingua representation as shown in Figure 4.

3) Generation Phase information: the information that relates to English word, such as POS, Verb pattern, number, English surface.



Pattern No	Verb Pattern
1	SUB+V
2	V+DOB
3	SUB+V+ADV
4	SUB+V+AUX
5	SUB+V+PP
6	SUB+V+DOB+PP
7	SUB+V+DOB
8	SUB+V+DOB+IOB
9	SUB+V+COMP
10	SUB+V+DOB+COMP
11	SUB+V+PP+COMP

Table 1: Thai Verb pattern

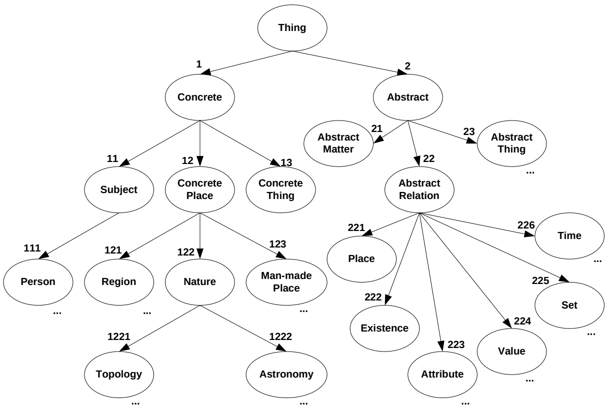


Figure 4 Interlingua Representation

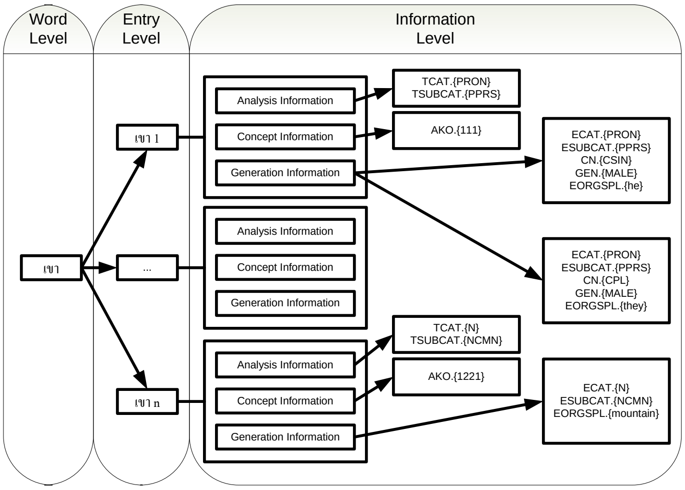


Figure 5 Example of Dictionary Data

Figure 5 shows an example of word "เขา". In entry “เขา 1”, the word “เขา” is a pronoun and can be interpreted as “he” or “they”. In entry “เขา n”, the word “เขา” is a noun and can be interpreted as “mountain”.

Currently, there are about 25,000 words in our dictionary. All words can be categorized into noun, pronoun, verb, verb attribute (adjective), auxiliary verb, determiner, adverb, classifier, conjunction and preposition.

3.4 Post-processing module

When translating an unknown word, the output will be Thai characters, which are mostly not be recognized by foreigners. Therefore, we provide the transliterate output in English to at least help foreigners pronounce the words.

Transliterate module uses Context Free Grammar(CFG) algorithm to construct a word in syllable tree form, and then apply Probabilistic Generalized Left-to-Right Parser (PGLR) to select the best tree. After that, syllable-to-phoneme is applied to change it into phonetic form [16]. Finally, the phonetic form will be mapped to English syllable by using the regulation from the Royal Institute of Thailand [11].

4 Implementation

We implement the PARSIT^{TE} system in order to provide to public users. To prevent the overloading problem due to many requests, we design architecture to support a large amount of its.

4.1 Translation architecture

Figure 6 shows the overall architecture of Client-Server based PARSIT^{TE} system, called T-E Module. This module receives any Thai sentences as an input for translation.

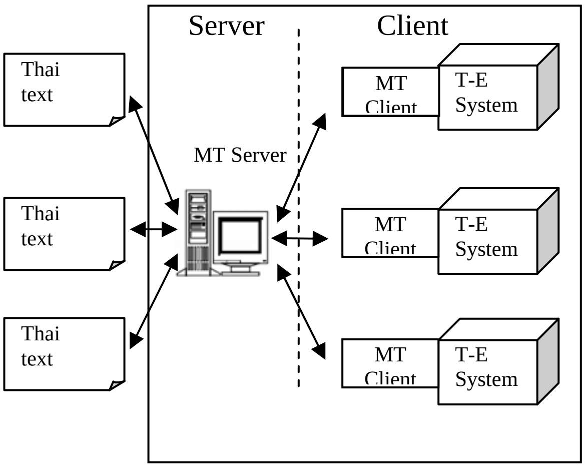
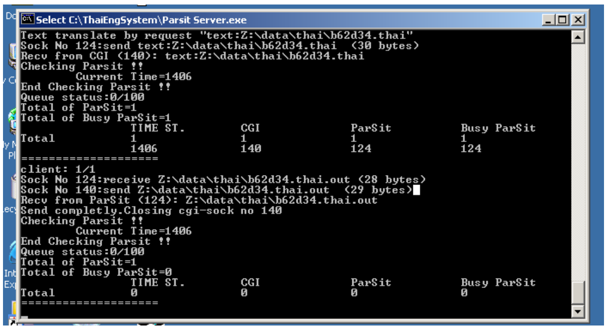


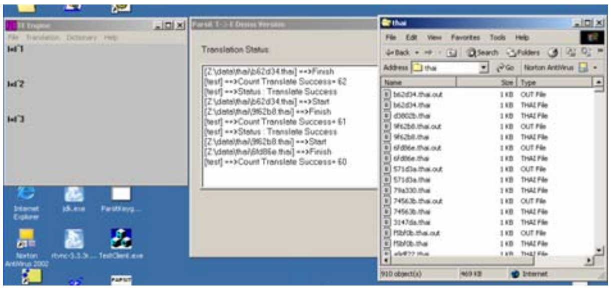
Figure 6: Client-Server based PARSIT^{TE} system (T-E Module)

There are two main parts in the T-E module. First, the MT-Sever is used for receiving our output in Thai sentence and distributing tasks to the MT-Client. Second, the MT-Client is a module that receives tasks from the MT-Server. It manages the incoming file (translation input) and the outgoing file (translation output) between the MT-Server and the T-E system (PARSIT^{TE} system).



```
Text translate by request "text:Z:\data\thai\b62d34.thai"
Sock No 124:send text:Z:\data\thai\b62d34.thai  (30 bytes)
Recv from CGI (140): text:Z:\data\thai\b62d34.thai
Checking Parsit !!
        Current Time=1406
End Checking Parsit !!
Queue status:0/100
Total of ParSit=1
Total of Busy ParSit=1
                TIME ST.        CGI             ParSit          Busy ParSit
Total           1               1               1               1
                1406            140             124             124
====================
client: 1/1
Sock No 124:receive Z:\data\thai\b62d34.thai.out (28 bytes)
Sock No 140:send Z:\data\thai\b62d34.thai.out  (29 bytes)
Recv from ParSit (124): Z:\data\thai\b62d34.thai.out
Send completly.Closing cgi-sock no 140
Checking Parsit !!
        Current Time=1406
End Checking Parsit !!
Queue status:0/100
Total of ParSit=1
Total of Busy ParSit=0
                TIME ST.        CGI             ParSit          Busy ParSit
Total           0               0               0               0
====================
```

Figure 7: Interface of the MT-Server



```
[Z:\data\thai\b62d34.thai] -->Finish
[test] -->Count Translate Success= 62
[test] -->Status : Translate Success
[Z:\data\thai\b62d34.thai] -->Start
[Z:\data\thai\9f62b8.thai] -->Finish
[test] -->Count Translate Success= 61
[test] -->Status : Translate Success
[Z:\data\thai\9f62b8.thai] -->Start
[Z:\data\thai\6fd86e.thai] -->Finish
[test] -->Count Translate Success= 60
```

Name	Size	Type
b62d34.thai.out	1 KB	OUT File
b62d34.thai	1 KB	THAI File
d3802b.thai	1 KB	THAI File
9f62b8.thai.out	1 KB	OUT File
9f62b8.thai	1 KB	THAI File
6fd86e.thai.out	1 KB	OUT File
6fd86e.thai	1 KB	THAI File
571d3a.thai.out	1 KB	OUT File
571d3a.thai	1 KB	THAI File
79a330.thai	1 KB	THAI File
74563b.thai.out	1 KB	OUT File
74563b.thai	1 KB	THAI File
3147da.thai	1 KB	THAI File
f5bf0b.thai.out	1 KB	OUT File
f5bf0b.thai	1 KB	THAI File
a9a877.thai	1 KB	THAI File

Figure 8: Interface of the MT Client

Figure 7 and 8 illustrate the MT-Server Program and MT-Client Program respectively. The translation process can be described as follows:

1) The MT-Server receives Thai input sentences.
2) The MT-Server sends command in order to specify input filename (Thai Sentence) to MT-Client.
3) The MT-Client receives an input file from MT-Server, sends the content of input to T-E system to translate, waits for the output translation, and then sends the message to MT-Server when translation is completed
4) The MT-Server sends the output file (English Sentence).

In case there are multiple requests simultaneously, the MT-Server will manage the translation queue by checking which MT-Client is available and sending task to that client. If all MT-Clients are busy, the MT-Server will keep the input file in a queue and send to the available MT-Client later.

4.2 Web-Based service

We provide this system for web-based service. Public users can use our system via the internet. As described in the previous section, there are pre-processing and post-processing phases to accomplish the translation. Users can select between automatic translation and semi-automatic one. If they select the automatic translation, the translation result will be sent automatically. If they select the semi-automatic one, the output from pre-processing phase will firstly be shown. In this step, they can edit the output from the segmenting phase to avoid the fault translation due to the segmentation errors.

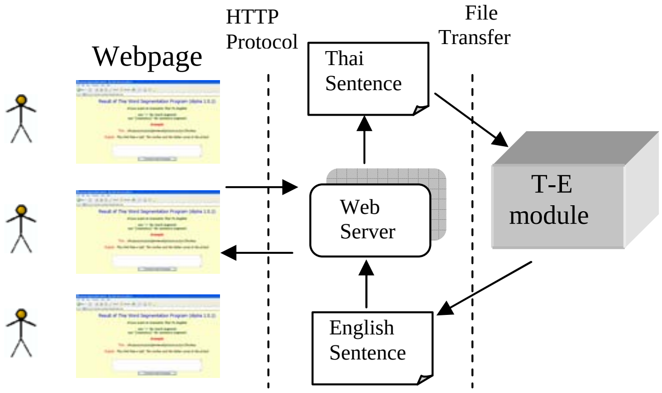
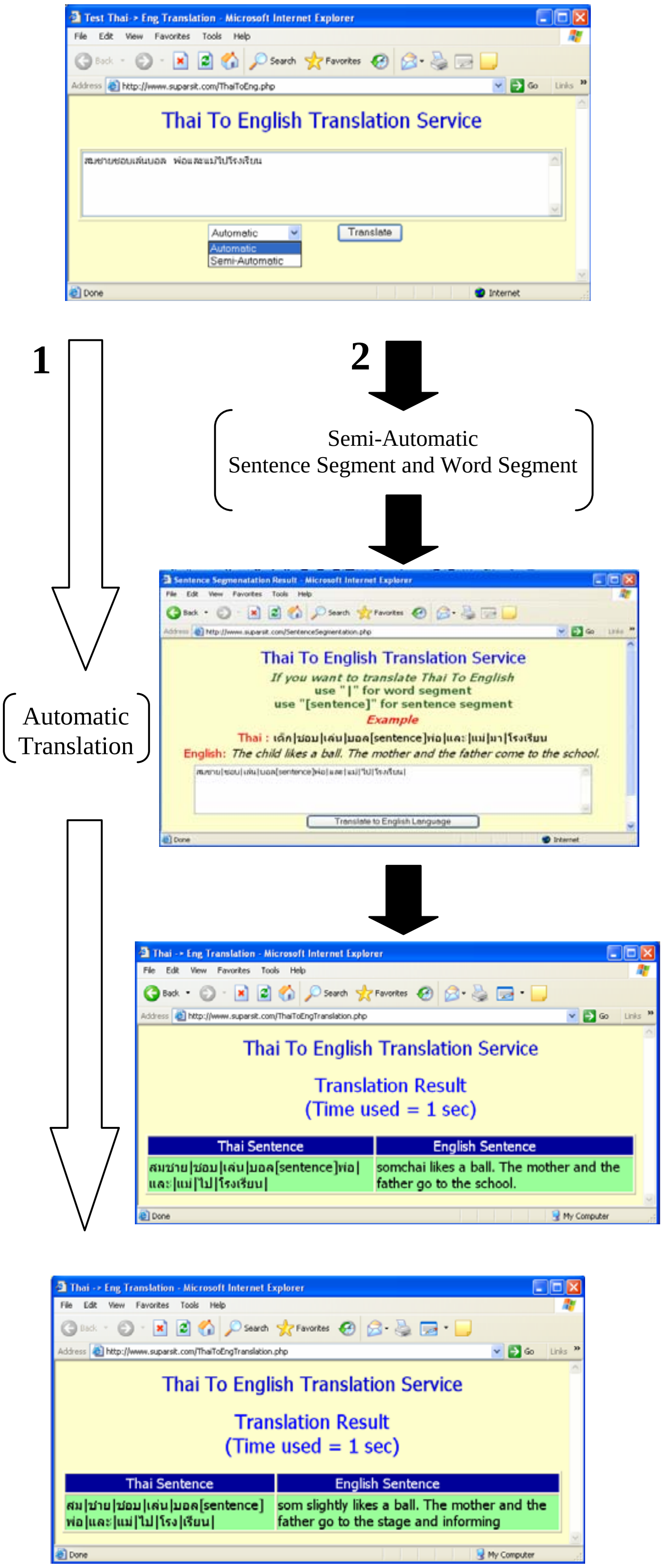


Figure 9: T-E web-based Architecture

Considering the web-based architecture designed as in figure 9, user sends a request through a web server, which is connected to the T-E module. After the T-E module finishes its process, the translation result will be sent back to the web server and then display on the web page.

An example of web-based service is illustrated in figure 10. Flow 1 corresponds to the automatic translation process, whereas, flow 2 corresponds to semi-automatic one. The later approach is useful when users require accurate translation, while the automatic approach is suitable for massive translation.

In this example, when a user uses the semi-automatic approach, the translation is correct. However, when a user uses the automatic translation. The result is deviated due to some mistakes in word segmentation.



Test Thai -> Eng Translation - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.suparsit.com/ThaiToEng.php
Go Links
Automatic
Automatic
Semi-Automatic
Translate
Done
Internet
1
2
Semi-Automatic
Sentence Segment and Word Segment
Automatic
Translation
Sentence Segmenatation Result - Microsoft Internet Explorer
Address http://www.suparsit.com/SentenceSegmentation.php
Translate to English Language
Thai -> Eng Translation - Microsoft Internet Explorer
Address http://www.suparsit.com/ThaiToEngTranslation.php
My Computer

Thai To English Translation Service

สมชายชอบเล่นบอล พ่อและแม่ไปโรงเรียน

Thai To English Translation Service

If you want to translate Thai To English
use "|" for word segment
use "[sentence]" for sentence segment
Example
Thai : เด็ก|ชอบ|เล่น|บอล[sentence]พ่อ|และ|แม่|มา|โรงเรียน
English: *The child likes a ball. The mother and the father come to the school.*

สมชาย|ชอบ|เล่น|บอล[sentence]พ่อ|และ|แม่|ไป|โรงเรียน|

Thai To English Translation Service

Translation Result
(Time used = 1 sec)

Thai Sentence	English Sentence
สมชาย|ชอบ|เล่น|บอล[sentence]พ่อ|และ|แม่|ไป|โรงเรียน|	somchai likes a ball. The mother and the father go to the school.

Thai To English Translation Service

Translation Result
(Time used = 1 sec)

Thai Sentence	English Sentence
สม|ชาย|ชอบ|เล่น|บอล[sentence]พ่อ|และ|แม่|ไป|โรง|เรียน|	som slightly likes a ball. The mother and the father go to the stage and informing

Figure 10: Example of MT Web Service.

5 Discussion

Since there has no Thai-English MT before, a test set for evaluating Thai-English MT system have never existed. To evaluate our system, we started by picking up examples from verb pattern which included complex sentence, compound sentence and sentence with conjunction. We found that simple sentence worked quite well in most of patterns. However, sentences that had the same form as noun phrase were still unsolvable. For example, the sentence "คอมพิวเตอร์ พิมพ์ รายงาน", which means "A Computer prints a report", is translated as "The Computer of the mold reports". This phenomenon occurred from the ambiguity between noun and verb because the word "พิมพ์" could be translated as "print (V)" or "mold (N)". It is necessary to solve this kind of pattern.

This system supports compound sentences but cannot support complex sentences yet. We can say that from table 1, the current version of Thai-English system can support verb pattern 1-8.

6 Conclusion and future work

We have developed Thai-English machine translation and have provided it online at http://www.suparsit.com/ThaiToEng.php.

Our system works well for simple sentences and some types of compound ones. However, the ambiguity between sentence and noun phrase is still unsolvable.

It is, however, necessary to add more dictionaries and optimize rule for Thai-English MT. For the future plan, there are three important works. First, we aim to increase the vocabulary in the system, add linguistics rule and apply machine learning to do both pre-processing and post-processing phases. Second, we need to improve the fundamental NLP algorithm for Thai such as word and sentence segmentation. Finally, we will develop a test set for evaluating the Thai-English MT system.

7 Acknowledgement

Special thanks to NEC Corporation, Japan for the collaboration in developing PARSIT^{TE} Thai-English MT system.

References

[1]Chin-Chung Chang and Chih-Jen Lin: LIBSVM: a Library for Support Vector Machines, http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf

[2]Nantana Danvivathan. The Thai Writing System, Forum Phoneticum 39, Helmut Buske Verlag Hamburg (1987).

[3]Paisarn Charoenpornsawat. Feature-based Thai Word Segmentation. Master's Thesis. Computer Engineering. Chulalongkorn University, Bangkok, Thailand (1999). (in Thai)

[4]Paisarn Charoenpornsawat and Virach Sornlertlamvanich. Sentence Break Disambiguation for Thai, In Proceeding of the 19th International Conference on Computer Processing of Oriental Languages, (2001).

[5]PARSIT: http://www.suparsit.com

[6]Pradit Mitrapiyanurak and Virach Sornlertlamvanich. The Automatic Thai Sentence Extraction. Proceedings of the Symposium on Natural Language Processing in Thailand (2000).

[7]Sergei Nirenburg, Jaime Carbonell, Masaru Tomita and Kenneth Goodman. Machine Translation: A Knowledge-Based Approach

[8]SWATH . Smart Word Analysis for Thai., http://www.links.nectec.or.th/download.php

[9]Sungkornsarun Longchupole. Thai Syntactical Analysis system by Method of Splitting Sentences from Paragraph for Machine Translation. Master Thesis. King Mongkut 's institute of technology Ladkrabang (in Thai) (1995).

[10]Suraphan Meknavin, Paisarn Charoenpornsawat, and Boonserm Kijsirikul. *Feature-based Thai Word Segmentation*. In Proceedings of the Natural Language Processing Pacific Rim Symposium (NLPRS'97), Phuket, Thailand (1997).

[11]The Royal Institute: http://www.royin.go.th/download/principles-of-romanization.pdf

[12] Thai Analysis Rules: Technical Report: Machine Translation System Laboratory, CICC (1995)

[13] Thai Part-of-Speech Tagged Corpus: Technical Report: Machine Translation System Laboratory, CICC (1995)

[14]Virach Sornlertlamvanich. Word Segmentation for Thai in a Machine Translation System (in Thai) (1993).

[15]Virach Sornlertlamvanich, Paisarn Charoenpornsawat, Mothika Boriboon and Lalida Boonmana. ParSit: English-Thai Machine Translation Services on Internet. 12th Annual Conference, ECTI and New Economy, National Electronics and Computer Technology Center, Bangkok, June (2000). (in Thai)

[16]Virach Sornlertlamvanich, Pongthai Tarsaku, Rachod Thongprasirt, .Thai grapheme-to-phoneme using probabilistic GLR parser., Proc. Eurospeech-01, pp.1057-1060, 2001.

[17]Yuen Poowarawan. Dictionary-based Thai Syllable Separation. In Proceedings of the Ninth Electronics Engineering Conference (1986).

[18]Surapant Meknavin, Paisarn Charoenpornsawat, Boonserm Kijsirikul. Feature-based Thai Word Segmentation. In Proceedings of Natural Language Processing Pacific Rim Symposium, pp. 41-46, 1997.

[19] P. Mitrapiyanurak and V. Sornlertlamvanich. The Automatic Thai Sentence Extraction. In Proceedings of the Fourth Symposium on Natural Language Processing, pp. 23-28, May 2000.