A Practical of Memory-based Approach for Improving Accuracy of MT

Sitthaa Phaholphinyo, Teerapong Modhiran, Nattapol Kritsuthikul 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
{Sitthaa.Phaholphinyo, Teerapong.Modhiran,
Nattapol.Kritsuthikul, Thepchai.Supnithi}@nectec.or.th

Abstract

Rule-Based Machine Translation (RBMT) [1] approach is a major approach in MT research. It needs linguistic knowledge to create appropriate rules of translation. However, we cannot completely add all linguistic rules to the system because adding new rules may cause a conflict with the old ones. So, we propose a memory based approach to improve the translation quality without modifying the existing linguistic rules. This paper analyses the translation problems and shows how this approach works.

1 Introduction

"ParSit" is English to Thai ruled-based Machine Translation [2],[3]. It has been launched in Thailand since 2000. This approach requires much linguistic knowledge. However, it is difficult to completely create all rules. Moreover, adding new rules may conflict with the old ones. Because of this limitation, we propose another approach to improve the translation quality.

In this paper, we will discuss the ParSit's errors in details. Then we show the overview of Post-Editing tool and how Post-Edit technique improves on ParSit's translation quality.

2 Characteristics of ParSit Errors

Charoenpornsawat et al. [4] claim that the common errors of machine translation can be classified into two types: incorrect meaning and incorrect ordering. The former, covering 81.74% of all errors, can be classified again into missing some words, generating over words and using incorrect words.

As said by Meechoonuk and Rakchonlatee [5], there are 12 types of linguistic problems found in ParSit translation. Table 1 shows a summarization of the mentioned problems

Types of Linguistic problems	***Source Sentence***	***Human Translation***	***ParSit Translation***	***Analysis***
Mismatched concepts	You will be ten years old next year.	คุณ-you|**จะมีอายุ-will be**| 10ปี-ten years old| ปี -year.|หน้า-next	คุณ-you|**กลายเป็น-will be**| 10ปี-ten years old| ปี year.|หน้า-next	**WILL BE** is translated as "become", but when it collocates with OLD, it should be translated as "age".
Misplaced modifiers	He lives quite near here.	เขา-he|อยู่-live|ใกล้-near| นี่-here|**ทีเดียว-quite**	เขา-he|อยู่-live|**ทีเดียว-quite**| ใกล้-near|นี่-here	**QUITE** occurs in the wrong position.
Inappropriate literal translation	Until recently, the company was in the red.	จนกระทั่ง-until| เร็วๆนี้-recently| บริษัท-the company| อยู่-was|ใน-in| **ภาวะขาดทุน-the red**	จนกระทั่ง-until| เร็วๆนี้-recently| บริษัท-the company| อยู่-was|ใน-in| **แดง-the red**	**RED** in this sentence refers to "lost" but ParSit translates it as "color".
Addition of words or phrases	The native of Java are like the Malays.	คนท้องถิ่น-the native| ของ-of|เกาะชวา-Java| **คล้าย-are like**| ชาวมลายู-Malays.	คนท้องถิ่น-the native| ของ-of|เกาะชวา-Java| **กำลังคล้าย-are like**| ภาษา-Malays.	**ARE LIKE** is used in present simple tense but ParSit translates it in the form of present continuous tense.

Types of Linguistic problems	*Source Sentence*	*Human Translation*	*ParSit Translation*	*Analysis*
Omission of words.	He is standing still.	เขา-he**|<u>ยืน-is standing</u>|** นิ่ง-still	เขา-he|นิ่ง-still	**<u>STANDING</u>** is omitted in ParSit translation.
Insufficient definitions of idioms, two-word verbs, and phrasal verbs.	He paid on the nail.	เขา-he|ชำระ-paid| **<u>ทันที-on the nail</u>**	เขา-he|ชำระ-paid| **<u>ตะปู-on the nail</u>**	**<u>ON THE NAIL</u>** is an idiom which means "immediately", but ParSit can not recognize it as an idiom.
Translation which does not conform to Thai grammar	He is wrong to leave.	เขา-he|**ผิด-is wrong|ที่-to|จากไป-leave**	เขา-he**|เดินทาง-leave|ไป-to| ผิด-wrong**	Thai language needs relative marker to express meaning for this sentence but English are not.
Implicit in both source language and target language	It is freezing cold today.	วันนี้-today| **เย็นยะเยือก-freezing cold**	มัน-it| **จับอากาศแข็ง-freezing cold|** วันนี้-today	**<u>FREEZING</u>** refers to temperature but ParSit can not translate this implicit meaning.
Active in source language but passive in target language	He criticized a man who they thought was his enemy.	เขา-he|วิจารณ์-criticized| ชาย-a man|ที่-who| พวกเขา-they| **<u>คิดว่า-thought</u>**|เป็น-was| ศัตรู-enemy|ของเขา-his	เขา-he|วิจารณ์-criticized| บุคคล-a man|ใคร-who| พวกเขา-they|ศัตรู-enemy| ของเขา-his|**<u>ถูกคิด-thought</u>**	**<u>THOUGHT</u>** in SL is a past tense verb but it is incorrectly translated in a passive form.
Insufficient dictionary definitions	Spring has come.	**<u>ฤดูใบไม้ผลิ-spring</u>**| มาแล้ว-has come	**<u>สปริง-spring</u>**|เกี่ยวกับ-has| มา-come	**<u>SPRING</u>** in the dictionary has only one definition.
Different semantic segmentation between source language and target language	She looked very sad, which she really was not.	เธอ-she|ดูเหมือน-looked| เศร้าเสียใจ-sad|มาก-very| ซึ่ง-which|เธอ-she| ไม่ได้-not|เป็น-was|จริงๆ-really	เธอ-she|ดูเหมือน-looked| เศร้าเสียใจ-sad|มาก-very| เธอ-she|ไม่-not|คือ-was| จริงๆ-really	**<u>The comma (,)</u>** is a punctuation in SL. But in ParSit, it is meaningless.
Specific in source language but generic in target language	These properties are used when simplifying algebraic expressions.	**<u>คุณสมบัติ-properties</u>**| เหล่านี้-these|ใช้-are used| เมื่อ-when|ทำให้-simplifying| นิพจน์-algebraic expressions| ง่ายขึ้น-simplifying	**<u>ลักษณะพิเศษ-properties</u>**| เหล่านี้-these|ถูกใช้-are used| เวลา-used|ลดรูป-simplifying| นิพจน์-algebraic expressions	**<u>PROPERTIES</u>** refers to natural quality in general. It is inappropriately translated to specific meaning in TL as particular quality.

Table 1: Types of linguistic problems

MT error types Chaorenpornsawat et al. [4]	**Linguistic Problems in Meechoonuk and Rakchonlatee [5]**
Missing some words	- Omission of words.
Generating over words	- Addition of words or phrases.
Using incorrect words	- Mismatched concepts - Inappropriate literal translation - Insufficient definitions of idioms, two-word verbs, and phrasal verbs. - Translation which does not conform to Thai grammar. - Implicit in both source language and target language - Active in source language but passive in target language - Insufficient dictionary definitions - Different semantic segmentation between source language and target language. - Specific in source language but generic in target language
Incorrect ordering	- Misplaced modifiers

Table 2: Comparison between MT error types and types of linguistic problems

We can map the 12 types of linguistic problems found on ParSit translation with the concept idea of machine translation errors. The re-arrangement in table 2 shows that using incorrect words is comprised of 9 types of linguistic problems. We suppose that Post-Edit technique can improve translation quality by reducing the errors of incorrect words as well as incorrect ordering. We will describe all the processes of our system, experiment and result in the next sections.

3 ParSit with Post-Edit module

3.1 System overview

Figure 1 shows the original ParSit. The English sentence is analyzed and generated to target sentence in Thai. We improve translation quality by adding Post-Edit module to the system as shown in Figure 2. The Post-Edit module detects some translation errors and corrects them to the appropriate form before sending the final output to users.

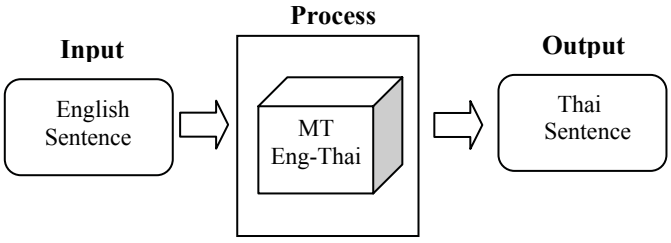

Figure 1: ParSit translation system

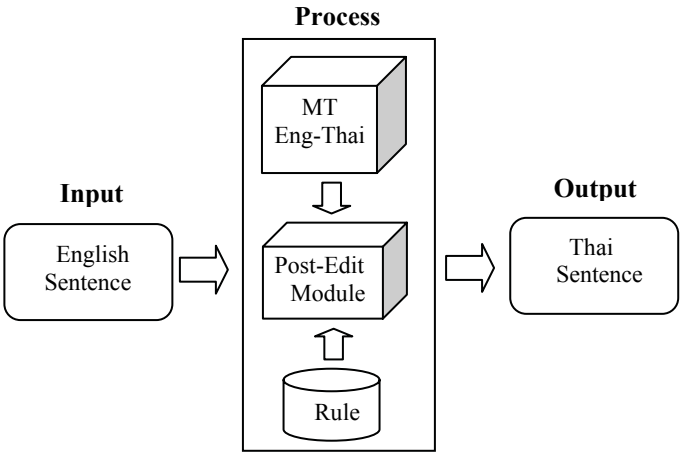

Figure 2: ParSit with Post-Edit module

3.2 Alignment tool

To construct Post-Edit rules, Alignment tool is developed to create input and output alignment. This tool is originated from Cairo program [6], used in Statistical Machine Translation (SMT) [7]. Cairo program concentrates only on the translation pair, while our tool concentrates on translation error correction.

ParSit generates both acceptable and unacceptable output. The input of the tool, called Alignment file is a pair of an English sentence and its unacceptable translation in Thai. We use Alignment tool to modify the translation result which is used to improve the translation quality.

We can open these files with Alignment tool as shown in Figure 3.

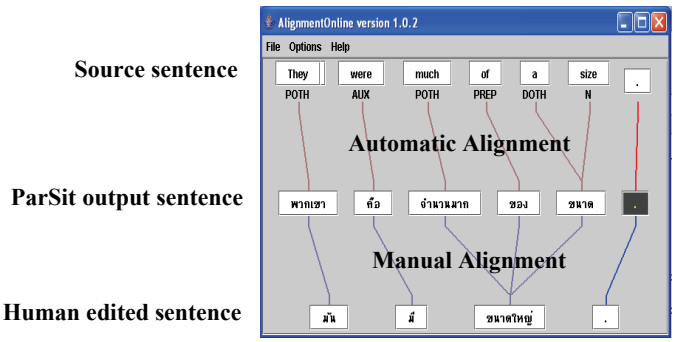

Figure 3: Alignment tool

Considering Figure 3, the topmost level is English source sentence. The middle part refers to ParSit generated output. The bottommost level is provided for human editing. There are two types of connection lines between levels. The upper ones link between the source sentence and ParSit output. These lines cannot be changed. The lower ones express a relation between ParSit output and human edited translation. Alignment file format are shown in Figure 4. Each file contains WORD, POS, ordering information in each pair sentence.

<source_sentence>
Word1|POS1 Word2|POS2 Word3|POS3 Word4|POS4….
<|source_sentence>
<target_sentence>
Word1 Word2 Word3 Word4 Word5….
<|target_sentence>
<Human_sentence>
Word1 Word2 Word3 Word4 Word5….
<|Human_sentence>
<alignment>
Number Number
… …
<|alignment>
<Human_alignment>
Number Number
… …
<|Human_alignment>

Figure 4: Alignment file format.

Alignment tool is used to modify words in the bottommost layer and the connection line between ParSit sentence and Human edited sentence. The commands in alignment tool that provides for linguists are shown as follows:

Add word is used to insert a new word.
Delete word is used to remove an unnecessary word.
Edit word is used to change a word.
Add link is used to insert line between the middle to the bottommost layer.

Delete link is used for removing unnecessary line between the middle and the bottommost layer.

The data collection which is rectified by human is called “Post-Edit Corpus”. It is used for creating rules in Post-Edit module.

The main criteria of editing alignment file are

- Edit only comprehensible sentence.
- Edit only necessary part in each sentence, not exceed three positions.
- Omit long, complex or incomplete sentence.

3.3 Post-Edit Rule

As mentioned above, Post-Edit rules are generated from collection of edited alignment files. In the current version, we use three types of relations to construct Post-Edit rules.

One-to-one relation is used to create a rule to change any wrong word to the proper one.

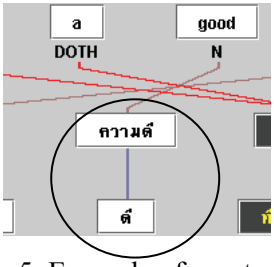

Figure 5: Example of one-to-one relation

Many-to-one relation is used to create a rule to combine multiple words into a new single word.

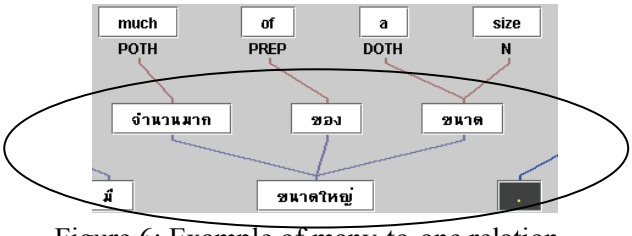

Figure 6: Example of many-to-one relation

One-to-many relation is used to create a rule to split a single word to multiple words.

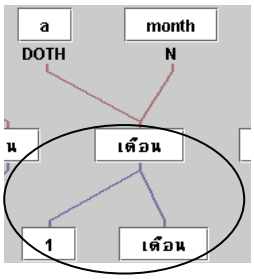

Figure 7: Example of one-to-many relation

3.3.1 Structure of Memory-based Rule Information

We apply Memory-based to assist rule generation. We consider word and POS, in source sentence as necessary features for creating rules.

The structure of information is:

Left#Center#Right#Source_word#Target_word#Human_word

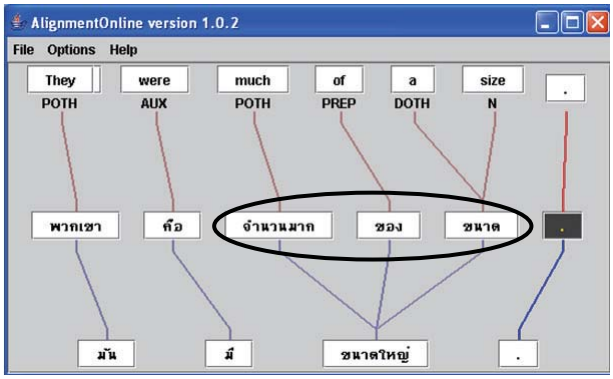

Figure 8: Example of alignment file.

Figure 8 shows an example of alignment file. we can create a rule to correct the word "จำนวนมาก-much/ ของ-of/ ขนาด-size/" to be "ขนาดใหญ่ –large size/" by 2 types of features.

First, we focus on two left and two right neighboring words. We will get "they"," were" and "." as word features. From this example, the information for constructing rule is.

Left : *They were*
Center: *of a*
Right: .
Source_word: *(much|POTH)(of|PREP)(a|DOTH size|N)*
Target_word: *จำนวนมาก ของ ขนาด*
Human_word: *ขนาดใหญ่*

Second, we use POS of two left and two right neighboring POS as feature. The information for constructing rule will be;

Left : *POTH AUX*
Center: *PREP DOTH*
Right:
Source_word: *(much|POTH)(of|PREP)(a|DOTH size|N)*
Target_word: *จำนวนมาก ของ ขนาด*
Human_word: *ขนาดใหญ่*

In some cases, two types of features can be combined to get better rules.

We store the information in trie structure[8] so that we can access rapidly. Trie structure is also appropriated to large amount of information.

In our data structure, Trie composed of two types of data; keyword and a collection of

Post-Edit rule. keyword is a data that point to its own representative Post-Edit rule collection.

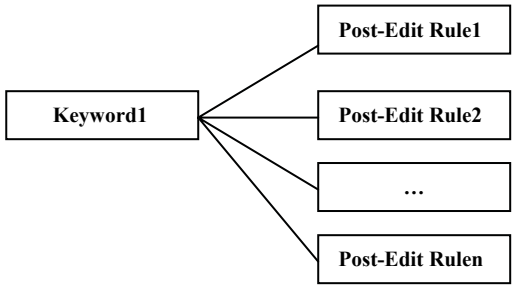

Figure 9: Data structure of rule information

We use production rule to represent Post-Edit Rule. It contains <*condition*> terms and <*action*> terms as shown in the following pattern.

IF <*condition 1*> ... <*condition n*> THEN
< *action 1* >... < *action n* >

The Post-Edit rule pattern is defined as follows:

IF *Match(Target_word)* THEN
IF*(Match(Left_word)*
&&Match(Right_word)
&&Match(Center_word))
THEN
Assign(Target_word Human_word);
}

From the example shown in Figure 8. Post-Edit rule pattern (two left and two right neighboring words) will be

IF (Match (Target_word "จำนวนมาก ของ ขนาด")) THEN
IF (Match (Left_word "They were"))
&& (Match(Right_word "."))
&& (Match (Center word "of a"))
THEN
(Assign Target_word ขนาดใหญ่);

Post-Edit rule pattern (two left and two right neighboring POS) will be

IF (Match (Target_word "จำนวนมาก ของ ขนาด")) THEN
IF (Match (Left_word "POTH AUX"))
&& (Match(Right_word "PUNC"))
&& (Match (Center word "PREP, DOTH"))
THEN
(Assign Target_word ขนาดใหญ่);

4 Result and Discussion

We have tested this technique with our 6,000 bilingual sentences and found that Post-Edit technique can reduce some important errors which frequently occur in ParSit's translation result. These improvements are shown below.

4.1 Improvement in incorrect meaning

The improvement in this section is categorized in two types: incorrect words reduction and over words reduction.

4.1.1 Incorrect words reduction

Source Sentence	ParSit Translation	ParSit Translation with Post Edit technique	Improvement in incorrect meaning
A serious accident occurred last year.	อุบัติเหตุ-accident| **ที่เคร่งเครียด-serious**| เกิดขึ้น-occured| ปี-year|ที่แล้ว-last	อุบัติเหตุ-accident| **ร้ายแรง-serious**|เกิดขึ้น-occured| ปี-year|ที่แล้ว-last	-Inappropriate literal translation
An overpowering burst of ammonia assaulted my lungs.	ระเบิด-a burst| พิชิต-overpowering| ของ-of| แอมโมเนีย-ammonia| **โจมตีอย่างหนัก-assaulted**| ปอด-lung|ของฉัน-my	ระเบิด-a burst| กำลังมหาศาล-overpowering| ของ-of| แอมโมเนีย-ammonia| **ทำลาย-assaulted**| ปอด-lung|ของฉัน-my	-Inappropriate literal translation
He is a nailer at lying.	เขา-he|คือ-is|**ผลิตตะปู-a nailer**|ที่-at| เวลามีสภาพ-lying	เขา-he|**เก่ง-is a nailer**| ใน-at|การโกหก-lying	-Insufficient definition of idioms -Addition of words or phrases -Insufficient dictionary definition
This subject lies near my heart.	สิ่งของ-subject|อันนี้-this|อยู่ที่-lies| **เป็นที่รัก-my heart**	เรื่อง-subject|อันนี้-this| อยู่-lies|ใกล้-near|**ใจ-heart|ฉัน-my**	-Mismatched concepts -Inappropriate literal translation
Death is a necessary end.	ความตาย-death|คือ-is| การจบ-a end| **ที่จำเป็น- necessary**	ความตาย-death|คือ-is| การจบ-a end| **ที่แน่แท้-necessary**	-Inappropriate literal translation

Table 3: Incorrect words reduction

4.1.2 Over words reduction

Source Sentence	ParSit Translation	ParSit Translation with Post Edit technique	Improvement in incorrect meaning
There isn't any logic to his move.	ไม่มี-there isn't|**ก็**| เหตุผล-logic|ใดๆ-any| ที่|ขยับ-move|ของเขา-his	ไม่มี-there isn't| เหตุผล-logic| ใดๆ-any| ต่อ-to| การเคลื่อนไหว-move| ของเขา-his	- Addition of words or phrases
He nakedly declared his intentions.	เขา-he| **ประกาศว่า-declared**| มุ่งมั่น-intentions| ของเขา-his| ไม่มีใบ-nakedly	เขา-he| **ประกาศ-declared**| ความมุ่งมั่น-intentions| ของเขา-his| อย่างเปิดเผย-nakedly	- Addition of words or phrases - Insufficient dictionary definition
To envy our betters is human nature.	การอิจฉา-to envy| เรา-our| **ถูกปรับปรุง-betters**| และ|คือ-is| ธรรมชาติ-nature| มนุษย์-human	การอิจฉา-to envy| **ผู้ที่เหนือกว่า-betters**| เรา-our|คือ-is| ธรรมชาติ-nature| มนุษย์-human	- Addition of words or phrases - Mismatched concept - Translation does not conform to Thai grammar
Success is near at hand.	ประสบความสำเร็จ-success|คือ-is| **อย่างนั้นใกล้จะถึง-near at hand**	ความสำเร็จ-success| **ใกล้เข้ามาแล้ว-is near at hand**	- Addition of words or phrases

Table 4: Over words reduction

4.2 Improvement in incorrect ordering

Source Sentence	ParSit Translation	ParSit Translation with Post Edit technique	Improvement in incorrect ordering
The tree are naked of leaves.	ต้นไม้-the tree|ใบไม้-leaves|**ไม่ทำ-are naked of**	ต้นไม้-the tress|**ปราศจาก-are naked of**|ใบไม้-leaves	- Misplaced modifiers - Insufficient definitions of idioms, two-word verbs, and phrasal verbs.
We are a democracy only in name.	เรา-we|คือ-are|ประเทศระบอบประชาธิบไตย-democracy|**เท่านั้น-only**|ใน-in|ชื่อ-name	เรา-we|คือ-are|ประเทศระบอบประชาธิปไตย-democracy|ใน-in|นาม-name|**เท่านั้น-only**	- Misplaced modifiers
An old man was fascinated by her lily hands.	ชายชรา-an old man|ถูก-was|มือ-hands|ลิลลี่-lily|ของเธอ-her|**หลงใหล-fascinated**	ชายชรา-an old man|**หลงใหล-was fascinated**|มือ-hand|ซึ่งขาวเหมือนลิลลี่-lily|ของเธอ-hand	- Misplaced modifiers
There is a lack of naval officers.	มี-there is|**ไม่เพียงพอ-a lack|ของ-of|**นาวิกโยธิน-naval officers	มี-there is|นาวิกโยธิน-naval officers|**ไม่เพียงพอ-a lack of**	- Misplaced modifiers
Some of the provincial newspaper are as influential as the national newspaper.	**หลายๆ-some|ของ-of|**หนังสือพิมพ์ส่วนภูมิภาค-the provincial newspaper|มี-are|อำนาจ-influential|เท่ากับ-as|หนังสือพิมพ์นานาชาติ-the national newspaper	หนังสือพิมพ์ส่วนภูมิภาค-the provincial newspaper|**หลายๆ ฉบับ-some of**|มี-are|อำนาจ-influential|เท่ากับ-as|หนังสือพิมพ์ระดับชาติ-the national newspaper	- Misplaced modifiers

Table 5: Incorrect ordering reduction

Charoenpornsawat et al. [4] had improved the qualities of translation by applying machine learning to the machine translation results. Focused on verb-to-be, they claimed that the accuracy had been increased about 25.1-29.7 per cent compared to the original results. We assimilate an experiment by applying these rules to Ripper [9] to generalize rule. However, the numbers of rules are too few, because the sparseness of data caused only partial features are analyzed. Finally we got too general rule.

However, Post-Edit Technique cannot improve some error such as missing word(s). Because this technique are based on context information, if the target sentence lack of the meaning or translation, the word's link on alignment tool also cannot link to the source word. Considering the Post-edit rule, we currently use the output from the corpus to generate rules based of word and POS features.

5 Conclusion and Future work

The advantage of this technique is to improve some weak point of machine translation such as incorrect meaning reduction. This technique requires a large size of input data to succeed. Our next step is to collect a larger size of corpus , handle with a long, complicate data set and apply an machine learning technique for automatic rule generation without sparseness problem.

References

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

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

[3]ParSit : http://www.suparsit.com

[4]Paisarn Charoenpornsawat ,Virach Sornlertlamvanich and Thatsanee Charoenporn. 2002. *Improving Translation Quality of Rule-based Machine Translation*. In the 19th International Conference on Computational Linguistics (Coling 2002): Workshop on Machine Translation in Asia. Taipei, Taiwan.

[5]Manisara Meechoonuk and Somporn Rakchonlatee. 2001. *An analysis of text translated by machine*. National Institute of Development Administration.

[6]Smith, Noah A. and Michael E. Jahr (2000). *"Cairo: An Alignment Visualization Tool,"* in the proceedings of The Second International Conference on Language Resources and Evaluation (LREC 2000), Athens, Greece

[7]Al-Onaizan, Yaser, Jan Curin, Michael Jahr, Kevin Knight, John Lafferty, I. Dan Melamed, Franz-Josef Och, David Purdy, Noah A. Smith, and David Yarowsky (1999). *"Statistical Machine Translation: Final Report,"* Johns Hopkins University 1999 Summer Workshop (WS 99) on Language Engineering, Center for Language and Speech Processing, Baltimore, MD, USA

[8]Trie algorithm: http://linux.thai.net/~thep/

[9]Cohen, William W. 1995.. *Fast effective rule induction*, In Proceedings of the Twelfth International Conference on Machine Learning, Lake Taho, California, Morgan Kauffman.