World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:2, No:6, 2008

Estimating Word Translation Probabilities for Thai – English Machine Translation using EM Algorithm

Chutchada Nusai, Yoshimi Suzuki, and Haruaki Yamazaki

***Abstract*—**Selecting the word translation from a set of target language words, one that conveys the correct sense of source word and makes more fluent target language output, is one of core problems in machine translation. In this paper we compare the 3 methods of estimating word translation probabilities for selecting the translation word in Thai – English Machine Translation. The 3 methods are (1) Method based on frequency of word translation, (2) Method based on collocation of word translation, and (3) Method based on Expectation Maximization (EM) algorithm. For evaluation we used Thai – English parallel sentences generated by NECTEC. The method based on EM algorithm is the best method in comparison to the other methods and gives the satisfying results.

***Keywords*—**Machine translation, EM algorithm.

I. Introduction

SELECTING the word translation from a set of target language words, one that conveys the correct sense of source word and makes more fluent target language output, is one of core problems in machine translation.

In Thai-English Dictionary a Thai word can have many senses and each of those senses can be mapped into many English words. For example, Thai word "เงิน" has two different senses, one of which refers to "money" and can be translated into many English words "*money, currency, cash, coin*", and the other one refers to "silver" and can be translated into two English words "*silver, atomic number 47*". Moreover, some of these English words are also the translation of the other Thai words in Thai - English dictionary. The example of multiple translations per word shows in Fig.1.

In machine translation, selecting the word translation is difficult work. Human translators select word translations that accurately describe the source meaning, but they also want to generate fluent target language output. That means a certain word translation may be preferred if it fits in well with other word translations. Also, the target language may have finer sense distinctions than can be foreseen in the source language [1].

In this paper we compare the 3 methods of estimating word translation probabilities for selecting the word translation in Thai – English Machine Translation. The 3 methods are (1) Method based on frequency of word translation, (2) Method based on collocation of word translation, and (3) Method based on Expectation Maximization (EM) algorithm. The method based on EM algorithm is the best method in comparison to the other methods in our experiment. The EM algorithm alternatively scores the possible English word for each Thai word (the expectation step) and estimates translation probabilities based on this (the maximization step) until convergence. The best method uses 3 knowledge sources which are a Thai – English dictionary, an English corpus (target language), and a Thai corpus (source language) while the other methods had not used a Thai corpus.

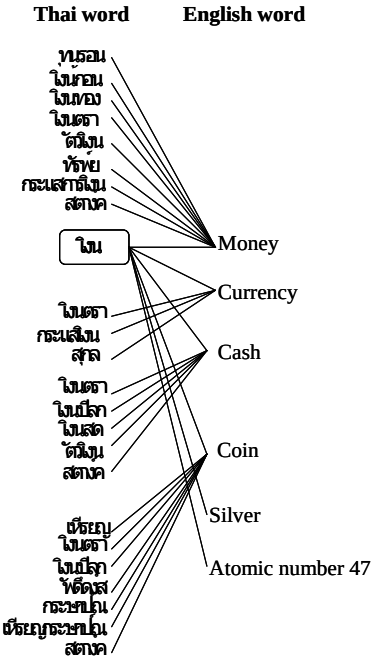

Fig. 1 example of multiple translations per word

In our research we focus on Thai nouns in a given sentence and construct the word translations. For testing purpose we used Thai – English parallel sentences (bitexts) in education domain generated by NECTEC [2]. The method based on EM algorithm gives satisfying results.

II. Related Work

At present, there are very few researches on machine translation of Thai language especially from Thai to other languages due to the several characteristics of Thai language make difficulty in translation.

In this paper, our approach is using "**collocation**" refers to the habitual co-occurrence of individual lexical items [3]. In other words, collocation embodies the relationship of words

that co-occur in a discourse. However, it is difficult to establish such related sets of words, sometimes called collocation sets, in a principled way [4]. The idea of semantic collocation is adapted and derived by [5] as an unsupervised learning algorithm that is based on two powerful constraints that words tend to have one sense per discourse and one sense per collocation.

Our approach uses both monolingual corpuses are a source language corpus and a target language corpus. The idea of using a target language monolingual corpus is exploited by [6]. They use a target language model to find the correct word-level translation. Reference [7] also proposed an approach using statistical information gains from a target language monolingual corpus in English – Persian machine translation. Parallel corpus could also be especially created for the training of machine translation system, but this is a costly option. The monolingual corpuses are readily available for most languages, while parallel corpuses rarely exist even for common language pairs. As in Thai language, we do not have Thai – English large parallel corpus, since the resources of Thai language for natural language processing (NLP) research few exist.

Currently, one of the most successful lines of machine translation research is unsupervised method by using statistical machine translation. The EM algorithm [8],[9] was applied in several tasks in many machine translation researches. Reference [10] employs the EM algorithm to estimate translation probabilities from an Indonesian - English bitext. Reference [11] employs the algorithm to learn the binarization bias for each tree node from the parallel alternatives. The EM-binarization yields best translation performance. Reference [12] proposed a new and effective method for Base noun phrase translation by using web data and the EM Algorithm. For our research uses the EM algorithm to estimate translation probabilities for Thai – English machine translation.

III. Method of Estimating Word Translation Probabilities

This research proposes 3 methods of estimating word translation probabilities for selecting the word translation in a Thai sentence. The all methods and knowledge sources that they used are shown in Table I.

TABLE I
THE METHOD OF ESTIMATING WORD TRANSLATION PROBABILITIES

No	Method	Knowledge source
1	Method based on frequency of word translation	Thai–English Dictionary (Lexitron dictionary) English corpus (Reuter news corpus)
2	Method based on collocation of word translation	Thai–English Dictionary (Lexitron dictionary) English corpus (Reuter news corpus)
3	Method based on EM algorithm	Thai–English Dictionary (Lexitron dictionary) English corpus (Reuter news corpus) Thai corpus (Thai concordance corpus)

A. Method based on Frequency of Word Translation

This method uses the frequencies of the word translation in a target language corpus to estimate word translation probabilities, regardless of context. In this form

$$freq(e_i : t) \quad (1)$$

The frequencies of English word translation e_i of Thai word t that occur in the English corpus.

This research, we focus on Thai noun in sentence to simplify our experimental setup.

For example, we look at the Thai noun "เงิน", our dictionary list 6 possible English word translations: *money*, *currency*, *cash*, *coin*, *silver*, and *atomic number 47*. We obtain the frequencies of each English word in our English corpus as show in Table II.

TABLE II
THE FREQUENCIES OF THE WORD TRANSLATIONS OF THE THAI NOUN " เงิน"

Frequencies	English word	Sense (meaning)
2095	*money*	money
1685	*currency*	money
1256	*cash*	money
145	*coin*	money
255	*silver*	silver
4	*atomic number 47*	silver
Total 5440		

So, we can calculate the probabilities that the Thai word "เงิน" will be translated as each English word P_w (e_i | เงิน). Such as the probability that the Thai word "เงิน" will be translated as English word "*money*" = P_w (e_i | t) = P_w (*money* | เงิน) = 2095/5440 = 0.385. Then, we select an English word that has highest probability score.

B. Method based on Collocation of Word Translation

This method uses collocation of word translations in a same given sentence which occur in target language corpus to estimate word translation probabilities. The idea of this method is that the translation of one Thai word will affect to the translation of the other Thai words in a same given sentence. As following example, the translation of Thai word "**ครู**" will affect to the translation of Thai word "**นักเรียน**".

Our research focuses on noun in the Thai sentence to simplify our experimental setup. For example, consider the following sentence:

(translation: The teacher tells the students to take their books to school), annotated with the English noun translations. The correct translations are in bold type.

	t_1	t_2	t_3	t_4	(t_s)
Thai sentence	ครูบอกให้นักเรียนนำหนังสือมาโรงเรียน				
Candidate English word translation	*teacher* *instructor*	*pupil* ***student*** *learner* *scholar* *disciple* *undergraduate*	*letter* ***book***	*school*	
	e_1	e_2	e_3	e_4	

The number of possible candidate English words sequences e_s are equal to 24 (2*6*2*1) sequences.

For example,	e_1	e_2	e_3	e_4	(e_s)
No. 1.	*teacher*	*pupil*	*letter*	*school*	
No. 2.	*teacher*	*student*	*letter*	*school*	
No. 3.	*instructor*	*learner*	*book*	*school*	
…					
No. 24	*instructor*	*undergraduate*	*book*	*school*	

We estimate probabilities $P(e_s|t_s)$ for each candidate English word sequence e_s by this formula:

$$P_{col}(e_s) = P_{col}(e_1 ... e_n)$$

$$= P_{col}(e_1)P_{col}(e_2 \mid e_1)...P_{col}(e_n \mid e_{n-1}) \quad (2)$$

We count the frequencies of English word and adjacent word in our English corpus. These frequencies allow us to estimate collocation probabilities $P_{col}(e_n|e_{n-1})$. With the resulting of collocation of English word translations we can compute the probabilities for each candidate English words sequence $P_{col}(e_s)$.

Then, we select the English words that occur in highest scoring candidate sequence.

C. Method based on EM Algorithm

This method combines the notion of translation probabilities with the use of context. We use the method based on EM algorithm to estimate word translation probability. This method uses the source language corpus and the target language corpus.

The two ideas of this method are; (1) since the translation of one Thai word will affect to the translation of the other Thai words in a same given sentence (same idea of method No. 2) ; and (2) since a Thai word can be translate into many English words and these English words are also translation of the other Thai words in our Thai –English dictionary. For example (see the example Thai sentence in method No. 2), the Thai word "นักเรียน" can be translate into 6 English words "*pupil*, *student*, *learner*, *scholar*, *disciple*, *undergraduate*" while these English words are also the translation of the other Thai words. Such as the English word "*pupil*" is also the translation of the other 8 Thai words "ตาดำ, รูม่านตา, เด็กนักเรียน, นักศึกษา, ผู้เรียน, ศิษย์, ลูกศิษย์, ธรรมมันเตวาสิก", the English word "*student*" is also the translation of the other 6 Thai words "นักศึกษา, ผู้เรียน, เด็กนักเรียน, ผู้เรียนรู้, ศิษย์, ลูกศิษย์" , etc. Actually, some of the occurrences of "*pupil*" ,and some of the occurrences of "*student*" in the English corpus do not relate to Thai word "นักเรียน".

Thus, we use the method based on EM algorithm to resolve this problem which to get better translation probabilities estimates. We have to take into account which occurrences of the word translation actually relate to the source word in consideration.

For explanation of this method , we use the example Thai sentence same as the one described in method No. 2, the number of possible candidate English words sequences are equal to 24 (2*6*2*1) sequences.

We compute probabilities for each candidate English words sequence of e_s. First, we use Bayes rule [13] ;

$$P(e_s \mid t_s) = \frac{P(e_s)P(t_s \mid e_s)}{P(t_s)} \quad (3)$$

But, the factor $P(t_s)$ can be discarded for the purpose of comparing different English noun sequences, since it is equal for all possibilities.

We now compute the remaining probabilities $P(e_s)P(t_s|e_s)$ using the collocation of English word translation probabilities $P_{col}(e_s)$ (Same Method No.2) combine with word translation probabilities that these English words will be translate as Thai words $P_w(t_s|e_s)$:

$$P(e_s)P(t_s \mid e_s) = P_{col}(e_s) \bullet P_w(t_s \mid e_s)$$

$$= P_{col}(e_1,..., e_n) \bullet P_w(t_1 \mid e_1)...P_w(t_n \mid e_n)$$

$$= P_{col}(e_1)P_{col}(e_2 \mid e_1)...P_{col}(e_n \mid e_{n-1}) \bullet$$

$$P_w(t_1 \mid e_1)...P_w(t_n \mid e_n) \quad (4)$$

Note : we only explain $P_w(t_1|e_1)…P_w(t_n|e_n)$

$P_w(t_1|e_1)$ mean the probabilities that English word e_1 will be translated as Thai word t_1

$P_w(t_2|e_2)$ mean the probabilities that English word e_2 will be translated as Thai word t_2

. . .

$P_w(t_n|e_n)$ mean the probabilities that English word e_n will be translated as Thai word t_n

For explanation, we use the example Thai sentence same method No.2.

To compute $P_w(t_2|e_2)$ of candidate English translation words sequence No.1.

e_1	e_2	e_3	e_4
teacher	*pupil*	*letter*	*school*

(in this example Thai sentence, t_2 is Thai word "นักเรียน")

First, we find the Thai words which can be translate into the English word e_2 "*pupil*". Our dictionary list 9 Thai words. We obtain the frequencies of each Thai word in our Thai corpus as show in Table III .

TABLE III
THE FREQUENCIES OF THE THAI WORDS

Frequency	Thai word
234	**นักเรียน**
10	ตาดำ
0	รูม่านตา
31	เด็กนักเรียน
1538	นักศึกษา
507	ผู้เรียน
62	ศิษย์
60	ลูกศิษย์
0	ธรรมมันเตวาสิก
Total 2442	

So, we can estimate that $P_w(t_2|e_2) = P_w($ นักเรียน $| pupil) = 234 / 2442 = 0.095$.

While we compute $P_w(t_2|e_2)$ of candidate English translation words sequence No.2.

e_1	e_2	e_3	e_4
teacher	*stuent*	*letter*	*school*

that $P_w(t_2|e_2) = P_w($นักเรียน $|$ *student*$) = 234 / 2432 = 0.096$

The $P_w($นักเรียน $|$ *student*$)$ is higher than $P_w($นักเรียน $|$ *pupil*$)$, because of main reason that English word "*pupil*" has two senses(meaning). One of these senses means "*a student supervised by a teacher*" and the other one means "*The dark circular aperture in the center of the iris of the eye*" which not relate to Thai word "นักเรียน".

With the collocation of word translation probabilities $P_{col}(e_s)$ and word translation probabilities $P_w(t_s|e_s)$ in place, we can find the best word translations for a given Thai sentence t_s by using the Bayes rule :

$$\arg\max_{e_S} P_S(e_S \mid t_S) = \arg\max_{e_S} P_{col}(e_S) P_W(t_S \mid e_S) \quad (5)$$

Then, we select the English words that occur in highest scoring candidate sequence.

IV. EXPERIMENTS

A. Knowledge Sources

The knowledge sources in our experiment consist of (1) English corpus used Reuter news corpus [14] compose of 4 domains are : No.1 Education/Government/Social, No.2 Corporate/Industrial, No.3 Economics, and No.4 Markets (4,030,596 words), (2) Thai corpus used Thai concordance Corpus [15] (3,476,000 words; education domain), and (3) Thai - English dictionary [16] (35,000 Thai words).

B. Test Data

For testing purposes we used Thai – English parallel sentences in education domain contain 1,000 sentences generated by NECTEC [2].

C. Experiment Methodology

After the Thai sentence aligning them we use our Thai - English dictionary to identify how the nouns in the Thai sentences were translated.

D. Evaluation

We measure how accurate the 3 methods match these word translation pairs in our Thai – English parallel sentences.

We computed the accuracy of each method as follow:

$$Accuracy \ (\%) = \frac{C}{A} * 100 \quad (6)$$

Where **C** refers to the number of the correct sentences, and **A** refers to the number of all sentences in our Thai-English parallel sentences (1,000 sentences)

Note that since some sentences more than one translation of a word may be fully acceptable, we can not expect 100% accuracy on this task, but it is still a very good metric of relative performance.

V. RESULTS

The results of all experiments are shown in Fig. 2. The experiment was divided into 4 groups by the domains of our English corpus while all groups use same domain Thai corpus (education domain). We estimated the accuracy of 3 methods in each group.

The results of the experiments show that the method based on EM algorithm (method No.3) provides the best accuracy result in all groups, and the group which used our English corpus in Education/Government/Social domain provides the best accuracy result in comparison to other groups.

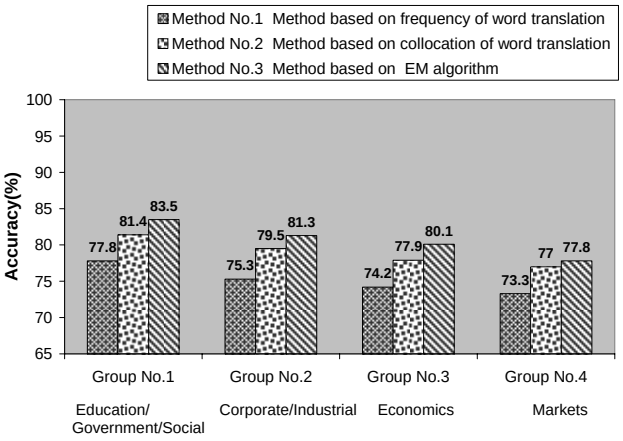

Fig. 2 Results of the experiments

VI. Discussion

A. The Reason why the English Corpus in Education/Government/Social Domain Provides the Best Accuracy Result

Since our experiment use Thai-English parallel sentence in Education domain for testing purposes. Thus, these Thai words are related to English words which occur in the English corpus in Education/Government/Social domain more than the other domains.

B. The Reasons why the Method based on EM Algorithm (method No.3) provides the Accuracy Higher than the other Methods

The reasons can be explained by the theory of language model in term of collocation. To estimate the word translation probability by using the collocation of English words translation probabilities $P_{col}(e_s)$ combine with word translation probabilities that these English words will be translate as Thai words $P_w(t_s|e_s)$ which useful for get better word translation probabilities estimate.

For the example of this explanation, the Thai word "หนังสือ" can be translate into two English words "*book*" and "*letter*", while the English word "*book*" is also the translation of the other 3 Thai words "ตำรับ, ปกรณัม, สมุด" ,and the English word "*letter*" is also the translation of the other 7 Thai words "จดหมาย, อักขระ, อักษร, ตัวเขียน, ตัวหนังสือ, ตัวอักษร, สาสน์".

In our experiment, we obtain the accuracy of the Thai word "หนังสือ" in each method are shown in Table IV.

TABLE IV
THE ACCURACY OF THE THAI WORD "หนังสือ" IN EACH METHOD

No.	Method	Accuracy (%)
1	Method based on frequency of word translation	33.33
2	Method based on collocation of word translation	75.00
3	Method based on EM algorithm	83.33

From the accuracy viewpoint, the method No.1 performs badly. In the method No.1, we use the frequencies of English words "*book*", "*letter*" in our English corpus to estimate word translation probabilities are shown in Table V. Then, we select an English word that has highest probability score.

TABLE V
THE PROBABILITY OF THE ENGLISH WORD TRANSLATION OF THAI WORD "หนังสือ"

English word	Frequency	Probability
letter	521	0.5489
book	428	0.4510

Actually, the most common word translation "*book*" ranks behind the English word "*letter*". This happens because the English word "*letter*" is also the translation of the frequent Thai words.

So, when we use method No.2 and method No.3 which use the collocation of English word translations of Thai nouns in the same sentence to consideration. The both methods can get better translation probabilities estimate.

In addition, the method No.3 can estimate word translation probabilities better than method No.2 because the method No.3 also use word translation probability that these English words will be translate as Thai words $P_w(t_s|e_s)$. In this example, the English word "*letter*" is also the translation of the other 7 Thai words "จดหมาย, อักขระ, อักษร, ตัวเขียน, ตัวหนังสือ, ตัวอักษร, สาสน์". Actually, some of the occurrences of "*letter*" in the English corpus do not in fact relate to the Thai word "หนังสือ". Thus, we have to take into account which occurrences of the word translation actually relate to source word in consideration by using $P_w(t_s|e_s)$. In this example, P_w(หนังสือ | *letter*) = 0.577 while P_w(หนังสือ | *book*) = 0.895. With the accuracy results in the experiment, we discuss that $P_w(t_s|e_s)$ is useful to get better translation probabilities estimate.

VII. Conclusion

In this paper we compare the 3 methods of estimating word translation probabilities for selecting the translation word in Thai – English Machine Translation. The 3 methods are (1) Method based on frequency of word translation, (2) Method based on collocation of word translation, and (3) Method based on EM algorithm. The focus of this research is to apply the semantic collocation of words from the target language corpus combine with the use of context in the source language corpus. Our current experimental setup is restricted to nouns, but it will extend to verbs, adjectives, etc. For evaluation, the method based on EM algorithm is the best method in comparison to the other methods with promising results.

References

[1] N. Ide and and J. Veronis, "Introduction to special issue on word sense disambiguation," The stat of the art. *Computational Linguistics*, 1998, 24(1):1-40.

[2] NECTEC: National Electronics and Computer Technology Center, Thailand, http:// www.nectec.or.th.

[3] D. Crystala, *Dictionary of Linguistics and Phonetics*, Blackwell, Oxford, UK, 1996.

[4] R. Wardhaugh, *Introduction to Linguistics*, McGraw-Hill Book Company. a. The study of language, Language in communication, 1972.

[5] D.Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in *Proc. of the 33rd Annual Meeting of the Association for Computational Linguistics*, 1995.

[6] I. Dagan and Itai, A., "Word sense disambiguation using a second language monolingual corpus," *Computational Linguistics*, 20(4):563-596, 1994.

[7] T. M. Miangah and A. D. Khalafi, "Statistical analysis of target language corpus for word sense disambiguation in a machine translation system," presented at the 9th EAMT European association for Machine translation, 2004.

[8] G. McLachlan and T. Krishnan, *The EM algorithm and extensions*. Wiley series in probability and statistics, John Wiley & Sons. , 1997.

[9] A. Mario, *Lecture Notes on the EM algorithm*, 2004.

[10] J. Cathcart and R. Dale, "Producing a Cross-Language Dictionary using Statistical Machine," in *Australasian Natural Language Processing Workshop*, Macquarie University, Sydney, Australia, 2001.
[11] W. Wang and K. Knight, "Binarizing Syntax Trees to Improve Syntax-Based Machine Translation Accuracy," in *Proc. EMNLP-CoNLL*, pp. 746–754, Prague, 2007.
[12] C. Yunbo and L. Hang , "Base Noun Phrase Translation Using Web Data and the EM Algorithm," in *Proc. of COLING-2002*, pp.127-133, 2002.
[13] A. Kaban, *Introduction to Bayesian Learning*, School of Computer Science University of Birmingham, 2004.
[14] Reuter News Corpus, available : http://trec.nist.gov/data/reuters/reuters.html
[15] Thai Concordance corpus, Department of Linguistics, Chulalongkorn University, available: http://www.arts.chula.ac.th/~ling/ThaiConc.
[16] Thai-English Lexitron dictionary, available: http://lexitron.nectec.or.th