English-Thai Translating Algorithm by Subject Category using Neural Network

Tongchana Oupatcha, and Nithinant Thammakoranonta

***Abstract*—**English is one of the most important language. It is hard to translate English into Thai. Because there are many meaning in Thai for one English word or the specific meaning for a technical term which different from general English term, it is interesting to use subject category of the paper to help identify the right meaning of a word. This paper proposed a algorithm to translate English into Thai using the context-based concept. The document will be prepared first by text operation algorithm, then all its keywords are fed into neural network to get the subject category before submitted to the translator process to match with the words in the dictionary database. 50 documents in NLP, environment sciences, and sociology are used to trained the neural network. 15 documents in these three areas are used to test the proposed algorithm. The results show that there is a statistical significant difference between the proposed algorithm and N-Gram algorithm when translating.

***Keywords*—**Translator, concept-based, classification, neural network

I. Introduction

ENGLISH is one of the formal languages that was used all over the world. There are a huge amount of books written in English. These books contain a lot of useful knowledge. Even the government requires all students in Thailand to learn English. Many people still cannot read and understand English well. Due to the complication of language from grammars structures and roots, it is hard to translate English into Thai correctly. Also one English word has many meaning in Thai. Thus is the reason why there one many translation tools such as talking Dictionary and Google translator, etc .

For translation Computer and IT text books, there are three major problems. The first one is there are many meaning in Thai for one English word, for example the word "account". This word in business area mean "finance and accounting" but in computer and IT area, it means "the list of users" The second problem is there are many specific terminologies in computer and IT area. It is not suitable to translate them into Thai, for example "Oracle". This is not sight to translate into Thai as "Foresee Angle" The last reason is there are many technical terms, which don't require translation, for example the word "Pivot". Most translation engine will translate it into "screw", which has no meaning in computer and IT area.

There are many language translators. These translators trains computers to learn words and grammars. However, most languages are too complicated, for example there are many meanings for one word, so most translators may pick the wrong words and meaning when translating. One example of language translator is Google translator, that uses "statistic machine translation" as the main model. [2] This Model use alignment paralleled corpus as the knowledge base and also N-gram to help translating from one language to the other language. The N-gram statistic indicates the frequency of appearance of a set of words. There statistics will be kept in knowledge base and will be used to translate by comparing the set of words and frequency. This translation model has no concern with grammars, so it can be used to translate any language. The only problem with this model is the relative large Corpus, used to stored statistic values and sets of words. [4][5]

Another translating model is example-based machine translation. [1] This model used a lot of samples sentences which stored in database to calculate the statistic, so the database is quite large. Also the model does not concern about grammars. Without a large database and enough sample sentences, it will be hard to learn and find the statistic values, which will affect the efficiency of translation. One problem of this model is a sentence many have different meaning due to different categories of subjects. It affect the efficiency of learning from examples. [3] To solve the problem mentioned above, content analysis concept from Information Retrieved operation is brought to consider. The concept is used to expand the query for something documents in the database. But before that the document has to classified to know which subject it is in. Thus will help the translator to pick the right word efficiently, because there are more than one meaning for translated word, To classified the document into any subject. Neural Network is used However, it is hard for computers to translate as good as human does. It can just help human to translate the repetitive words faster In this study ,Neural Network used back-propagation and multilayer perceptron as learning method.

II. Translating Process

These one three main processes for translating from English to Thai.

Tongchana Oupatcha is a student in School of Applied Statistics, national Institute of Development Administration, Bangkok, Thailand.

Nithinant Thammakoranonta is an assistant professor in School of Applied Statistics, national Institute of Development Administration, Bangkok, Thailand.

http://dx.doi.org/10.15242/IIE.E0714035

A. Text operation process

This process eliminates all stopwords such as “a”, ”an”, “the”, or general words which are found in every document. These it will eliminate the prefix and suffix of the rest of the words until getting to the root of each word. After finding the root of all words in the document, the frequency of each root word is counted and the calculated to find the standardized frequency. Standardized frequencies along with corresponding words will be used to classify the document into suitable subject category.

B. Classification process.

The keywords and their corresponding standardized frequencies are led into the trained neural network to get the result, which is the subject category.

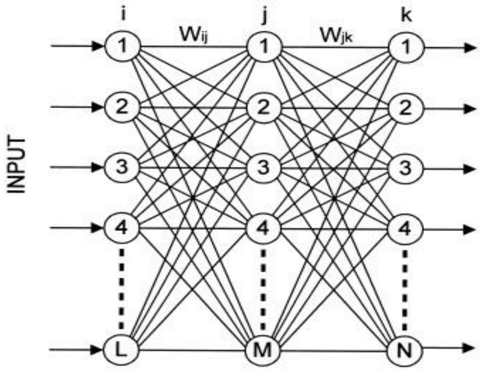

Fig1. Back-Propagation Architecture

TABLE I
PARAMETER OF TRAINING NEURAL NETWORK

Parameter in NN	Parameter value
Learning rate (η)	0.1
Training cycle (t)	500
Input data (Pi)	2391 words
bias (b)	0.1
Hidden Layer	12

C. Translation process.

The list of words from a document and the subject category are led into this process. The program will call up the list of words which have the same subject category as the input from the dictionary database. Then the word in Thai are assigned to every word from the input list.

To improve the efficiency of process. The document, its words and the corresponding standardized frequencies and its assigned subject category are shown to the user. If the assigned subject category is wrong, the program will ask the user to correct the subject category. All data will be stored in a storage. Every month the data from this storage will be used to train the neural network for getting the values for node parameter.

III. EVALUATING THE SYSTEMS

50 documents from sociology, IT and environment subject categories one used to train the neural network. Also 30 documents are used to test the efficiency of neural network for classifying the documents into subject category. These three subject areas were used because they provide significant different keywords, which will make the result for this study clear. The test results have 40%, 70% and 50% correctness for sociology, IT and environment area respectively. However, the result from this classification process can be overrided by expertise and kept in the storage for training the neural network every month. The performance of the neural network should be better over a period of time.

TABLE II
RESULT OF NEURAL NETWORK TESTING

Area	No. of training document	No. of correctness	% of correctness
Environment	10	5	50%
IT	10	7	70%
sociology	10	4	40%

With the subject category, the words from 15 documents translate into Thai with 96.47%, 96.41% and 96.97% correctness, for sociology, IT and environment area respectively. While using Google Translator, which is an example of N-Gram model to translate the same documents, the average correctness for these three subject categories are 90.39%, 72.27% and 90.86% respectively. A paired t-test statistic provided the value equaled to 5.455 with 0.000 significant level. Thus statistic result confirmed that content-based algorithm can give a better translation performance the N-gram model at 95% significant level.

TABLE III
PAIR SAMPLES STATISTICS

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	V1	96.6340	15	.88977	.22974
	V2	83.7553	15	9.45245	2.44061

TABLE IV
PAIRED SAMPLES CORRELATIONS

		N	Correlation	Sig.
Pair 1	V1&V2	15	.389	.152

TABLE V
PAIRED SAMPLES TEST

		Paired Differences					t	Df	Sig. (2-tailed)
		Mean	Std. Deviatio n	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	V1–V2	12.879	9.143	2.361	7.815	17.942	5.455	14	.000

V1: Percent accuracy in the interpretation of the proposed algorithms.
V2: Percent accuracy in the interpretation of N-gram algorithms.

IV. CONCLUSION

This study proposed a new concept for translating English into Thai. The program consists of 3 main processes, which one text operation process, classification process and translation process. All results are kept in storage and are used to update the knowledge for classifying documents. With the subject category, the correctness breed to translate from English into Thai is higher than N-Gram concept

REFERENCES

[1] Arnold, D. J., Balkan, L., Meijer, S., Humphreys, R. L., and Sadler, L., "Machine Translation: An Introductory Guide," London: NCC Blackwell, 2001.

[2] GoogleTranslate, "About Google Translate,"http://translate.google.co.th/about/intl/th_ALL/, 2011.

[3] McTait, K., "Translation Pattern Extraction and Recombination for Example-Based Machine Translation," Doctoral dissertation. Centre for Computational Linguistics, Department of Language Engineering, University of Manchester Institute of Science and Technology., 2001.

[4] Nattapol K., Arit T., and Thepchai S., "English-Thai Example-Based Machine Translation using n-gram model," *IEEE International Conference on Systems, Man, and Cybernetics*, Oct. 2006.

[5] Peter et al Brown., "A statistical approach to language translation," *in Proceedings ofthe 12th COLING*, 1988, pp. 71–76.