THE RECOGNITION SYSTEM OF SENTENTIAL NOUNS PHRASE STRUCTURE USING CONDITIONAL RANDOM FIELD MODEL

Phuridech Phopiphat[1,*] and Rachada Kongakchandra[2]

[1,2]Department of Computer Science, Faculty of Science and Technology, Thammasat University, Klongluang, Patumthani, 12120, THAILAND

ABSTRACT

The important problem of word segmentation in Thai language is sentential noun phrase. The existing studies try to minimize the problem. But there is no research that solves this problem directly. This study investigates the approach to resolve this problem using conditional random fields which is a probabilistic model to segment and label sequence data. The results present that the corrected data of noun phrase was detected more than 78.61 % based on our technique.

KEYWORDS

Sentential noun phrase, conditional random field, recognition system

1. INTRODUCTION

Recently, the computer science has developed rapidly by integrate several knowledge in order to create new knowledge. Language is very important to create public benefits, such as creating electronic dictionary program, searching information from the internet. However, the differences in the grammar of language between Thai and foreign languages can cause limitation of search engine or display an error from different languages. Although today there are many researches that attempts to resolve these limitations of language, the problem are still exist.

In Thailand, research project about translation English into Thai using computers has begun since the beginning of the year 2524. From then onwards, the researchers still developed and analyze language seriously that aim to create natural language processing especially in English language more than Thai language. Because, Thai language has several specially characteristics that are very different from the English language. According to the study Pornprasertsakul et al. (1989) [1] have discussed about the nature of grammar in Thai. Thai sentence structure has specific nature of the problem especially when compared to the English vocabulary such as, no tense, verbs are not transformed follow by subject of sentence , no plural noun and neglect terms of the sentence etc. Moreover, noun phrase is important in natural language processing because the structures of noun phrases look similar to sentence structure.

The study of Kawtrakul and Boonkwan (2004)[2] have discussed the main problems in the functioning of machine translation of analyzing sentence structure in Thai language including, the issue of extension extravagant (modifier attachment) 14.70%, gerund look similar to structure (sentence nominalization) 9.67% ,noun phrases look similar to sentence (Sentential-Noun Phrase-

SNP) 15.10% and omission 41.00%. This study found that the phrase is significantly importance to analyze the distribution clause using machine translation (MT).

In addition, study of Kawtrakul et al. (2004) [3] also studied the problems in distribution and syntactic analysis. This study concluded that nouns phrase has a problem that needs to resolve seriously. After survey in 1,000 sentences, researcher found problems in distribution and noun phrase analysis as follow, noun phrase boundary ambiguity 14.9%, verb phrase structural ambiguity 18.5%, zero anaphora 27.0%, case transition 8.6% and excessive nomialization 4.2%. The researchers described problem with noun phrase boundary ambiguity (14.9%), referring to the confusion of natural language processing to find out the true boundary of the noun phrase when the nouns phrase has expanded with phrases such as verbs and prepositional phrases.
Moreover, ambiguous in describing the structure of the phrase occurs. The ambiguity of the phrase boundaries can be broken down into three problems: 1. sentential noun phrase 9.3% 2. sequential noun phrase 4.4% 3. noun phrase topicalization 1.2%.

For the reasons above, we can realize that the noun phrases are important issues that affect translation. The researchers including, computer engineers, computer scientists and computer linguists try to identify the most effective and how to analyze noun phrase. In this study aim to studied and developed a system that can segment the word of noun phrases look similar sentence structure. The noun phrase is a critical issue. This study analyzes data using statistical model Conditional Random Fields (CRFs) in process of name recognition. The study of Lafferty et al. (2001) [4] found that CRFs are more advantages when compared to (Hidden Markov models: HMMs) and (Maximum entropy Markov models: MEMMs). Additionally, study of Sha and Pereira (2003) [5] also showed that the results from CRFs got the best percentage of detected when analyzes the noun phrase compared with CoNLL method.

In the past, several studies have related with processing the natural Thai language using CRFs model. Study of Kruengkrai et al. (2006)[6] compared with other methods, including Longest Matching and Maximum Matching. The results showed that CRFs got high accuracy and study of Haruechaiyasak, Kongyoung and Dailey (2008) [7] conducted a study comparing the efficacy of segmentation from models that separate in four types including, Naive Bayes, decision tree, Support Vector Machine, and CRFs. The results revealed that CRFs use the word Thai is better than other models. Moreover, study of Supnithi et al. (2010) [8] found that CRFs can resolve sentential noun phrase better than CoNLL task.

In this study aim to compare the performance of the recognition system in implicit sentential nouns phase structured like sentence using statistical Conditional Random Fields model (CRF).

2. RELATED WORKS

The literature review was performed to know the concepts, theories and existing study in three topics as follows; problem of noun phrases, concepts and theories of technique and Conditional Random Fields model.

Firstly topic is problems of noun phrases. The researchers limited the scope of review, only two main problems of the noun phrase. There are the noun phrase that is structurally similar sentence (Sentential Noun Phrase) and Sequential Noun Phrase.

2.1. Sentential Noun Phrase - Sentential NP

The noun phrase that extended with verb or other phrases usually have structurally similar sentence because Thai language is no indication or upon terms that indicate function of phrase. Study of Kawtrakul et al. (2004) [3] divided this problem into two groups as follow;

2.1.1 Explicit Sentential Noun Phrase

This group compounds words using the same structure as the sentence (noun + verb + noun). Therefore, if we do not consider in the detail, we might misunderstand that is sentence. A sentence that could be analyzed in this problem, for example;

แปลง	เพาะ	กล้า
[+ N]	[+ V]	[+ N]

First Meaning: plant nursery culture young plant---act (the sentence).
Second Meaning: plant nursery. (Noun phrases, verb phrases with extension).

This example is very clearly to explain problem. The structure of the noun phrase structured like sentence that consist of noun followed by verb and objective. But when human analyze it, human can see that the word is not a sentence. Because, human can recognize and discern immediately that word would be a sentence or a verb. The landmark is verb itself from this example can be seen that the verb "เพาะ" is verb that requires subject is alive. If noun that place before verb "เพาะ" is not alive. This information can be concluded that this is noun phrase. However, computer cannot analyze this information immediately same as human. If the computer knows this landmark as well as human know, it is not a difficult to process. Computer will analyze results same as the human mind. Computer must need lexical feature, which is indeed an important tool to significant increase in the processing and capacity of the computer.

2.1.2 Implicit Sentential Noun Phrase

Study of Kawtrakul et al. (2004) [3] explained that this type of noun phrase is specific words that not come from compound words. There are only the names of scientific careers, time or event, tools, etc., for example.

คน	ขับ	รถ
[+ N]	[+ V]	[+ N]

2.2 Sequential Noun Phrase

Sequential Noun Phrase consists of compound nouns. This is one of the issues about the noun phrase boundary ambiguity that we have inability to find the end of structure. There is example of this problem as below;

เสา	ไฟ	ฟ้า
N1	N2	N3

This noun phrase consists of three nouns. Natural language processing issues can interpret in difference meaning. For example show in Table 1 below.

Table 1. Analysis of phrase "เสาไฟฟ้า".

Structure	example	meaning
1. N1 + N2 + N3	เสา + ไฟ+ ฟ้า	All the equal importance of meaningful share.
2. N1 + [N2N3]	เสา+ไฟฟ้า	[ไฟฟ้า] extend [เสา] which is the main noun.
3. [N1N2] + N3	เสาไฟ + ฟ้า	[ฟ้า] extend [เสาไฟ] which is the main noun.
4. N1N2N3	เสาไฟฟ้า	This is one word that cannot be separated from each other.

The phrase "เสาไฟฟ้า" was analyzed various meaning that shown above. But the accuracy meaning is second meaning, with the "เสา" which is a main noun as the first word. "ไฟฟ้า" expand the meaning of the "เสา" .The meaning of this noun phrase is electronic pole.

The researchers limited the scope of study; we focus on sentential noun phrase because the important problem of word segmentation in Thai language is sentential noun phrase. The existing studies try to minimize the problem. But there is no research that solves this problem directly. Second topic is concepts and theories of the study that usually use three approach as follow, Rule-based

The rules-based of the language used most often by an expert linguist or an analyst that present Pros and Cons in Table 2.

Table 2. The advantages and disadvantages of the use of the rule.

Advantages	Disadvantages
The performance recognition accuracy is quite high.	It takes longer time to develop.
	Requires specialized knowledge to create the rules.
	The rules are usually written with the language experiment. Therefore, it is very difficult to adapt with other languages.

Machine learning

Human will training computer to know the rules that may be learn characterized by the presence of group and order. Computer can recognize the unique name that is similar to the traditional name. The feature will different in each model to learn. Statistical models usually be used, such as Support Vector Machines (SVMs), Decision Tree, Hidden Markov Models (HMMs), Maximum Entropy (MaxEnt, ME), Conditional Random Fields (CRFs), etc. Table 3 present pros and cons of machine learning.

Table 3. The advantages and disadvantages of approaches to statistical modeling.

Advantages	disadvantages
Take time to develop more quickly than the rule-based.	Using big database for training.
Do not need experts to analyze the rules.	

The mixed (hybrid)

This method combines between rules and statistical methods. This is to reduce the limitations of both methods. If the rule does not recognize the new name, it could use a statistical method based on the frequency displayed. However, for this method need adequate time and resources.

This study focused on machine learning method due to limitation of time and human resources. For machine learning, we selected Conditional Random Fields (CRF) model because in study of Lafferty (2001)[4] presented that CRF showed several advantages over hidden Markov models and stochastic grammars for such tasks, including the ability to relax strong independence assumptions made in those models. Conditional random fields also avoid a fundamental limitation of maximum entropy Markov models (MEMMs) and other discriminative Markov models based on directed graphical models, which can be biased towards states with few successor states.

Thirdly, this study review concept and theory conditional random fields for more detail at below;
Nowadays, conditional Random Fields (CRF) model is widely accepted that a more effective model, Hidden Markov Models (HMMs), which is a generative models that rely on the joint probability between the input data and results/label released. It is a problem that cannot handle the relationship between the properties, because the information related to various properties independently. The model reduces such problems is Maximum Entropy Markov Models (MEMMs) with a discriminative models that rely on the conditional probability of the results or label sequence in different forms. When MEMMs observe sequence, it can capture the relationship between different properties. However, it also found problems called label biased because the judge at any conditions rely on only current conditions and observation sequence. Other conditions and sequences do not affect in calculation process of this probability models.
Conditional Random Fields (CRF) was proposed by Lafferty et al. (2001)[4] as a model to minimize problems that can explain in three models in Figure 1.

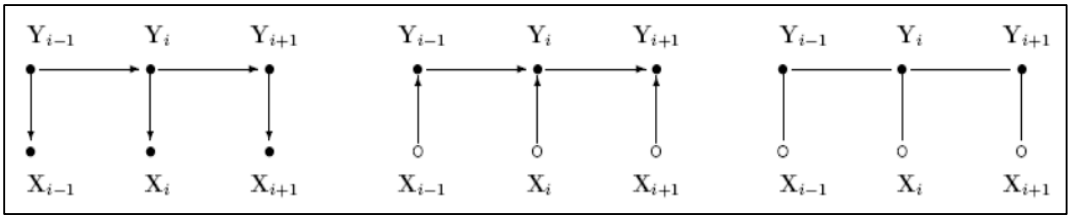

Figure 1. A comparison of models HMMs, MEMMs and CRFs of Lafferty. et al. (2001) X is found from the data line (observation sequence), and Y is the line of work or labels.

Label sequence of HMMs and MEMMs are a directed graphical model. HMMs is a generative model that shown joint probability distribution P (X, Y). Line in HMMs is not involved in this labels or results. So, the result Y is the first order of events or causes X (Sutton and McCallum, 2007)[9] .For MEMMs and CRFs are discriminative models that label sequence in both models are not calculated from processing model, but calculate from specification in condition of label sequence. The probability distribution conditional is P (Y | X). It means that line X occurs before line Y. CRFs model is undirected graphical model which is different from MEMMs. All previous

events with condition will be gotten to reconsider again when CRFs will find probability of the next label. The weights of the various properties in different condition also are considered. So, results are balances with the conditions are.

3. CLASSIFICATION ALGORITHMS AND METHODOLOGY

3.1 Preparing data from BEST2012

Thai text corpus (BEST2012), which is the scope of the 5 million words consisting of four categories Article, Encyclopedia, News and a novel were prepared by the National Electronics and Computer technology Center (NECTEC).

For this study, researchers have taken sentential noun phrase in training technique using Thai text corpus (BEST2012) in only the categories of News for research purposes. Because the news articles are common form of writing that can be found in every day. Study of Warawudhi R. (2006) [10] presented keyword of beginning sentential nouns phrase that divided in five main noun including careers, medical profession, tool, place and time or events. Table 4 present examples of five categories.

Table 4. Examples of main nouns in sentential noun phrase.

Semantic meaning	Examples of nouns	Example of noun phrase
1. Noun phrase about career	คน พนักงาน เจ้าหน้าที่ อาจารย์ ครู ชาว เด็ก	คนขับรถ อาจารย์สอนภาษาไทย พนักงานเก็บตั๋วโดยสาร
2. Noun phrases about medical profession	โรค ไวรัส เชื้อ เซรุ่ม วัคซีน	เข็มฉีดยา วัคซีนป้องกันโรค
3. Noun phrase about tool	เครื่อง ที่ ไม้ ชุด อุปกรณ์	เครื่องซักผ้า ไม้แขวนเสื้อ ที่ตัดกระดาษ กรรไกรตัดเล็บ
4. Noun phrases about building or place	โรง ร้าน ตึก อาคาร ที่ แปลง สวน ศูนย์ ย่าน ห้อง แผง	ร้านขายของชำ โรงเพาะ ศูนย์อนุรักษ์พันธ์สัตว์ทะเล แปลงปลูกผัก แผงขายผ้า
5. Noun phrases about time and event	วัน งาน เทศกาล พิธี	วันสถาปนาโรงเรียน งานเลี้ยงต้อนรับพนักงานใหม่

Thai text corpus (BEST2012) have already used statistical model to tagging type of word. But still are unable to determine the boundaries of sentential noun phrases. Therefore, researchers investigate recognition system method to determine the beginning and end of sentential noun phrase using five main nouns.

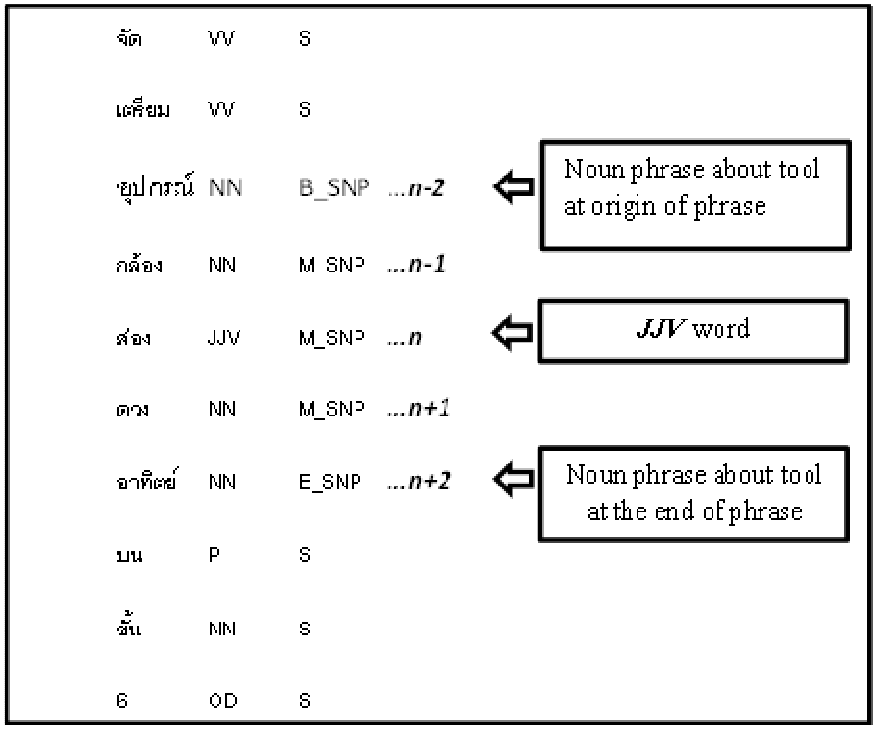

Figure 2. Method of tagging technique in sentential noun phrase.

Figure 2 present step of tagging using five main nouns. The first step of tagging sentential noun phrase is searching for "JJV" and then begin find boundary of phrase. Researchers start with the beginning of phrase. The main noun of sentential noun phrase always located at the origin of phrase. After researcher found the beginning of the phrase, they will find the end of a noun phrase by observing the type of words that contain NN, NR, JJV, JJA.

3.2 Recognition system of sentential noun phrase

Recognition system was developed for tagging sentential noun phrase. The sentences that contain verbs in Thai text corpus (BEST2012) were marked by abbreviation "JJV". After, these sentences were retrieved from corpus. These data bring to pre-processing, which including Word segmentation, removal token (e.g. white space and tab). After that the data obtained from the pre-processing bring to set feature and tagging sentential noun phrases. Then, data that have already tagging bring to training using conditional random field model. After, finished training process, researcher will get model. This model will be tested accuracy with testing data. Figure 3 present flow chart of recognition system.

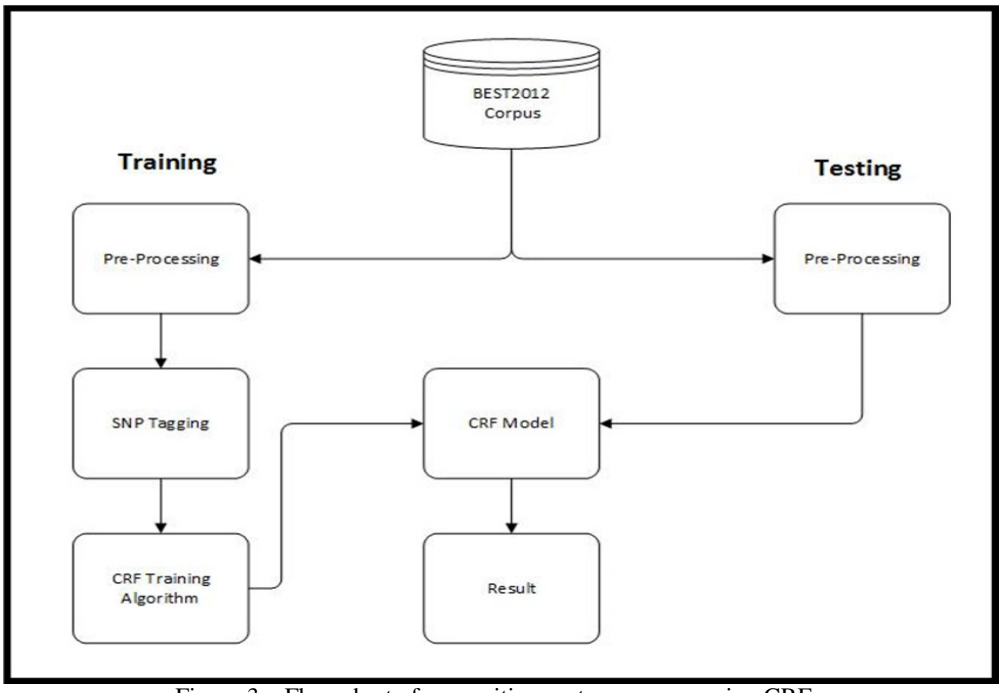

Figure 3. Flow chart of recognition system process using CRFs

4. EXPERIMENT AND RESULT

In this research, Researchers have taken samples of the sentential noun phrase from the archives BEST2012 of 900 words and used 10-Fold Cross-Validation to testing for avoid result that based on any sets of training data. Data set were divided into 10 sections. Nine sections used for training and another one section were used for testing that show calculation in Table 5. After that, researcher changed the data used to train and test for 10 times. Finally, the results were evaluated precision, completeness (Recall) and the F-measure.

Table 5. Preparing data for training and testing classify to five categories

	Training	Testing	Total
1. Noun phrase about career	225	25	250
2. Noun phrases about medical profession	108	12	120
3. Noun phrase about tool	135	15	150
4. Noun phrases about building or place	225	25	250
5. Noun phrases about time and event	117	13	130
Total	810	90	900

There are results of training from recognition system of sentential noun phrase in Figure 4.

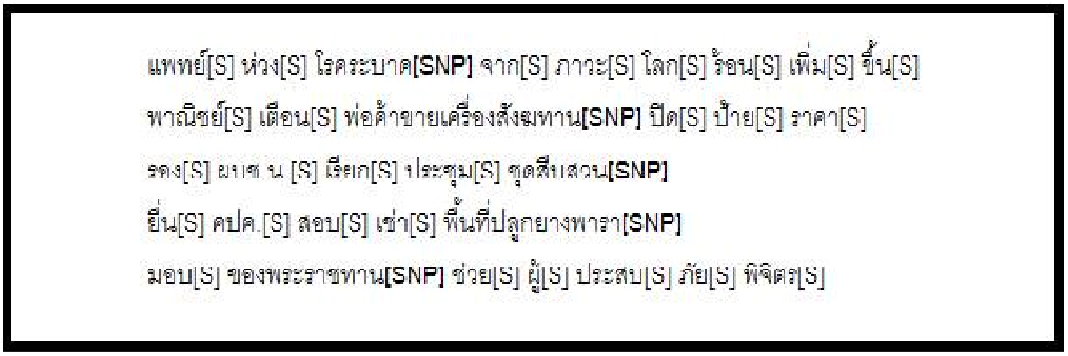

Figure 4. Example of tagging technique

This study evaluated the performance of the system by precision , completeness (Recall) and F-measure that present in Table 6 at below,

Table 6. The accuracy of results

	Precision (%)	Recall (%)	F- measure (%)
1. Noun phrase about career	81.77	79.10	80.41
2. Noun phrases about medical profession	80.93	79.25	80.08
3. Noun phrase about tool	81.54	80.67	81.10
4. Noun phrases about building or place	77.18	76.00	76.58
5. Noun phrases about time and event	69.15	66.92	68.01
Total	78.61	76.56	77.57

5. CONCLUSIONS AND FUTURE DIRECTION

For this research conducted study using conditional random fields model (CRF) in order to recognize sentential noun phrases. The results showed that the corrected data of noun phrase was detected more than 78.61 % based on this technique. However, this study has several limitations as follows; numbers in noun phrase, conjunction in noun phrase and spaces between noun phrases, etc. Therefore, the mixed technique (rule-based and machine learning) should be applied in further studies for more accuracy results and resolve these limitations.

REFERENCES

[1] Pornprasertsakul, A. and V. Wuwongse. (1989). The Suitability of Generalized Phrase Structure Grammar to Describe Thai Syntax. In Asian Institute of Technology, editor.

[2] Kawtrakul, A. and P. Bookwan. (2004). Parsing Thai Using Augmented State Transducer and Lexical Functional Grammar. In Proceedings of NCSEC2004.Thailand: Haad Yai.

[3] Kawtrakul, A., P. Bookwan, V. Satyamas, and P. Varasrai. (2004). Featurebased Finite-State Automaton: Enhancement of Phrase Chunker for Lexical Functional Grammar. Bangkok: Kasetsart University.
[4] Lafferty, J., McCallum, A. and Pereira, F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001), pages 282-289
[5] Sha F. and Pereira F. 2003 .Shallow Parsing with Conditional Random Fields. Proceedings of HLT-NAACL (2003), pages 213-220. Edmonton, Canada.
[6] Kruengkrai, C., Sornlertlamvanich, V., and Isahara, H. (2006). A Conditional Random Field Framework for Thai Morphological Analysis. In Proceeding of the Fifth International Conference on Language Resources and Evaluation. Genoa
[7] Haruechaiyasak, C., Kongyoung, S., and Dailey, M. N. (2008). A Comparative Study on Thai Word Segmentation Approach. In Proceedings of ECTI-CON. Krabi.
[8] Supnithi, T., Onman , C., Porkaew P. et al. (2010). A Supervised Learning based Chunking in Thai using Categorial Grammar. Proceedings of the Eighth Workshop on Asian Language Resouces.
[9] Sutton, C. and McCallum, A. (2007). An Introduction to Conditional Random Fields for Relational Learning. MIT Press.
[10] Warawudhi R. (2006). Thai Noun Phrase Analysis for Natural Language Processing. Bangkok; Kasetsart University

Authors

Phuridech Phopiphat received a B.Sc degree from Thammasat University, Thailand. Currently, He is Master student in Computer Science in Thammasat Universi ty and works as System Engineer in National Science and Technology Development Agency (NSTDA). His research interests include Artificial Intelligent, Handwritten Recognition, and Machine Learning.

Rachada Kongakchandra received the Ph.D. degree in electrical and computer engineering from King Mongkut's University of Technology Thonburi, Thailand. Currently, she works as Assistant Professor at the Computer Science department, Faculty of Science and Technology, Thammasat University. Her research interests include Artificial Intelligent, Natural Language Processing, Semantic Processing, and Machine Learning.