Semi-automatic Classification Based on ICD Code for Thai Text-Based Chief Complaint by Machine Learning Techniques

Jarunee Duangsuwan and Pawin Saeku

***Abstract*—We proposed the methods to classify the text-based chief complaint in Thai language, our native language, into the symptom code based on ICD-10. Using Thai sign and symptom descriptions from ICD-10 document is the training data to build Thai text-based corpus in domain of sign and symptom. Then the corpus has been used for tokenization of Thai text-based chief complaint (ThCC) into a particular word by using the longest matching technique and our proposed technique named two-level tokenization technique. The tokens from two techniques are evaluated by five different classifiers including decision tree classifier, K-mean neighbours classifier, radius neighbours classifier, random forest classifier, and extremely randomized tree classifier. The experimental result shows 85% accuracy for assigning ICD-10 code to Thai text-based chief complaint by using our proposed technique with decision tree classifier.**

***Index Terms*—Classification, Thai language processing, chief complaint identification, machine learning.**

I. Introduction

Thai language is an official language of Thailand, and is used as the first language by Thai people. Thai script has three main components: 44 consonants (in Thai "พยัญชนะ" pronounced " pha-yan-cha-na"), 15 vowels (in Thai "สระ" pronounced "sa-ra") that combine into at least 28 vowel forms, and 4 tone diacritics (in Thai "วรรณยุกต์" pronounced "wan-na-yuk"). Unlike English language, Thai writing system is a non-word boundary style that means no delimiter to separate a word in a sentence. Typically, the few spaces that occur in a sentence act as punctuation marks rather like commas and full stops in English. We use Thai chief complaint as a text-based input for our study. Therefore, the challenge is how to do tokenization because it is difficult for Thai language which does not use space between words. The chief complaint (CC) or presenting complaint (PC) is a concise medical statement presenting the sickness complaint from patients recorded by a physician, nurse, or health care professional. The information classified from CCs supports a decision making of syndromic surveillance. However, the limitation is the existing CC classification systems serving mainly an English language, and makes other languages difficult to employ. The result from decision making on CC is a partial of symptom diagnosis by a physician [1].

Manuscript received February16, 2018; revised April 20, 2018.

The authors are with the Department of Computer Science, Prince of Songkla University, Hat Yai, Songkhla, Thailand (e-mail: jarunee.d@psu.ac.th, tamakosan14@gmail.com).

Using ICD to show the result of CC classification is one of the standard forms. ICD, the International Statistical Classification of Diseases and Related Health Problems, is a code system of a medical classification established by World Health Organization. ICD represents a list of codes for diseases, signs and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or diseases [2]. We refer ICD-10 which is the tenth version and is the lasted revision of ICD to identify CC in our study. The CC can be classified as ICD code to identify the sickness problem without the limitations of language. Tokenization is generally considered as easier task than others in natural language processing, and one of the more uninteresting tasks for English. However, the non-segmented languages such as Chinese, Japanese, Laotian, and Thai need a specific tokenization method to deal with specific characteristics of each language. We proposed [3] which is a two-level tokenization to split each word in ThCC, and then separated word is assigned a part of speed (POS) tag according to medical domain. In this paper we apply machine learning techniques to classify tagged CC represented in terms of ICD-10 code which is used as sign and symptom identification.

II. Background

Although many researches have proposed CC classifiers, most of such classifiers do not pay more attention to a data preparation required in order to improve the classification. The research [4], [5] proposed two chief complaint data cleaners: chief complaint processor (CCP) and emergency medical text processor (EMP-P). Such processors have been designed by probabilistic-based and key-word based classifier, and shown that text-based chief complaint preprocessing can increase the performance of chief complaint classification. The ontology concept explained in [6] applied the rule-based approach to map the text-based chief complaint into symptom group which is classified later into syndrome group. This ontology concept supports that the text-based chief complaint including English text-based chief complaint needs the chief complaint preparation to increase the performance of classification.

The research in healthcare domain, such as [1], [7], [8], used information extraction strategic approach to collect and extract the required information from medical record. For [7], it presents clinical entities extraction using machine learning-based approach for name entities recognition (NER) task and then combining machine learning-based approach with rule-based approach to improve performance of

doi: 10.18178/ijfcc.2018.7.2.517

extraction. Furthermore, [8] integrated active learning (AL) which is a sample selection approach with supervised machine learning. Consequently, this study proposed the AL methods which can reduce amount of annotated sample in a clinical NER task for identifying the medical concepts from the clinical notes. [1] have proposed a knowledge discovery process which extracts the name entities from publications by using dictionary-based approach ,and finds a rule-based relation between those name entities by employing biotic corpus as study case.

Most studies mentioned previously have based on English so tokenization, preprocessing step before doing chief complaint classification, is simpler for the text-based processing application, but is nontrivial for Thai language because of no punctuation and inflection in Thai script at all. There has been not much research about Thai Language Processing for example; [9], [10] have proposed the corpus named Orchid. The Orchid has been built under the collaboration between Communications Research Laboratory (CRL) of Japan and National Electronics and Computer technology Center (NECTEC) of Thailand. The Orchid is about 2MB word collection of samples from the annually conference proceedings of a NECTEC. This part-of-speech (POS) tagged corpus consists of 13 categories which are subcategorized into 45 subcategories. TaLAPi or Thai Linguistically Annotated Corpus for Language Processing proposed by [11] is 1-million-word corpus of Thai. The goal of building TaLAPi is to support the implementation of a real-time Thai language processing. TaLAPi contains 35 POS tags adapted from Orchid corpus and 10 named entity categories. Additionally, this corpus covers some loan and foreign words.

III. Methods

Fig. 1 illustrates our proposed methods which are explained in detail as follows:

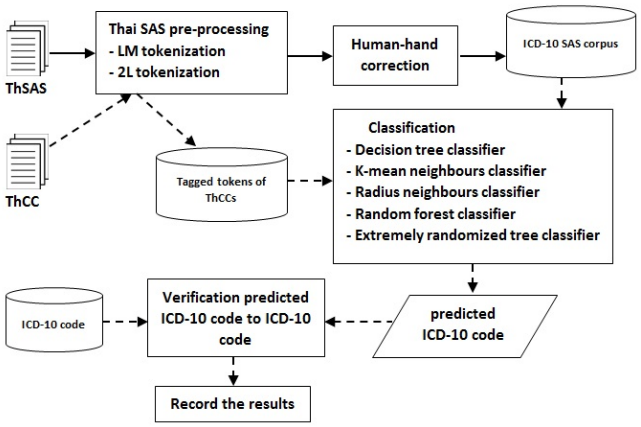

Fig. 1. The summarization of classifying methods for Thai text-based chief complaints.

A. Thai Text-Based SAS Pre-processing

The goal of pre-processing is to construct corpus from Thai text-based sign and symptom (ThSAS) which is the samples from ICD-10 document [12]. These samples used as our training dataset covering the sign and symptom description appearing in group R00 to R69 which contain 80 cases, and are tagged by 38 codes. We briefly explain this pre-processing because the elaborate details are shown by [3].

TABLE I: Example of Prepared Data Yielding from Thai Text-Based SAS Pre-processing

ICD-10 code	Text based sign and symptom		Tokens of ThSAS (T)			
	English	*Thai (ThSAS)*	*LM technique*	*No. of Tokens*	*2LT technique*	*No. of Tokens*
R04.0	Epistaxis	เลือดกำเดาไหล	เลือดกำเดา(T1) | ไหล(T2)	2	เลือด(T1) | กำเดา(T2) | ไหล(T3)	3
R06.0	Dyspnea	หายใจลำบาก	หายใจ(T1) | ลำบาก(T2)	2	หาย(T1) | ใจ(T2) | ลำ(T3) | บาก(T4)	4
R10.0	Acute abdomen	ปวดท้องเฉียบพลัน	ปวดท้อง(T1) | เฉียบพลัน(T2)	2	ปวด(T1) | ท้อง(T2) | เฉียบพลัน(T3)	3
R00.0	Tachycardia	ใจเต้นเร็ว	ใจเต้น(T1) | เร็ว(T2)	2	ใจ(T1) | เต้น(T2) | เร็ว(T3)	3
R30.9	Painful micturition	ปวดขณะถ่ายปัสสาวะ	ปวด(T1) | ขณะ(T2) | ถ่ายปัสสาวะ(T3)	3	ปวด(T1) | ขณะ(T2) | ถ่าย(T3) | ปัสสาวะ(T4)	4

Two main steps have been done in this pre-processing. Firstly, the 80 cases are spitted into words or tokens by open source tool named Lexto that is a longest matching (LM) tokenizer [13]. However, the outputs from Lexto do not fit to our research domain so output refinery needed by the next step. The refining step, called two-level tokenization (2LT) which is our proposed methods of tokenization for ThSAS, finds out element conflict then checks medical term matching. The 2LT applies the intersection theory in mathematics to an adapted algorithm based on Lexto. When the conflict is detected by the adapted algorithm, the conflicting token and the remaining token, the outputs of the first level, are forwarded to the second level. The adapted algorithm is used again for matching the conflicting token and the remaining token with an open source medical spelling word list named e-MedTools [14]. If the relation of both occurs from the matching then the new token is constructed base on the conflicting token and the remaining token and recorded in the corpus of sign and symptom named ICD-10 SAS corpus otherwise the conflicting token and the remaining token are eliminated from the considerations.

For the test dataset, we use the 57 sample cases from ThCC recorded by a physician or a nurse in terms of medical electronic data which are collected from [15], [16]. Such 57 cases have been processed in the same way by using the pre-processing mentioned above which yield tagged tokens of ThCC. Table I illustrates examples of training data which describes ICD-10 code, description of sign and symptom in

English and Thai Languages, tokens of ThSAS which is an output of tokenization, and numbers of tokens getting from each technique. The first row of figure explains that "R04.0" is ICD-10 code of Epistaxis which is represented in ThCC as "เลือดกำเดาไหล". ThCC "เลือดกำเดาไหล" is separated by LM into 2 tokens: "เลือดกำเดา" (T1) and "ไหล" (T2) whereas 2LT separates ThCC "เลือดกำเดาไหล" into 3 tokens: "เลือด" (T1) "กำเดา" (T2) and "ไหล" (T3). These data have been collected in order to construct corpus in next process.

TABLE II: OUTPUTS OF DATASET TRANSFORMATION

ICD-10 code	Bag of Words	
	LM technique	*2LT technique*
R04.0	Text :['ไหล', 'อดกำเดา', 'เล'] Index :[50,43,69] Count :[99,104,11]	Text :['ไหล', 'กำเดา', 'อด', 'เล'] Index : [76,2,36,62] Count: [221,81,100,12]
R06.0	Text : ['ลำบาก', 'หายใจ'] Index : [29,36] Count : [33,100]	Text : ['บาก', 'ลำ', 'หาย', 'ใจ'] Index : [17,25,32,71] Count : [231,418,333,915]
R10.0	Text : ['ยบพล', 'เฉ', 'อง', 'ปวดท'] Index : [22,59,32,18] Count : [18,111,333,99]	Text : ['ยบพล', 'เฉ', 'อง', 'ปวด'] Index : [22,52,34,18] Count : [18,1,1,130]
R00.0	Text : ['ใจเต', 'เร'] Index : [79,68] Count : [13,210]	Text : ['เร', 'เต', 'ใจ'] Index : [61,56,71] Count : [123,331,915]
R30.9	Text : ['ขณะ', 'สสาวะ', 'ายป', 'ปวด'] Index : [2,32,51,71] Count : [794,333,1, 913]	Text : ['ขณะ', 'สสาวะ', 'าย', 'ปวด'], Index : [3,28,47,18] Count : [794,1,1,913]

B. Dataset Transforming

Text classification differs from general attribute dataset because text length is various sizes. Consequently, to transform text-based data into a fixed size data is needed in order to using such data as training and testing set for classification with machine learning approach. The feature-based vectorization provided to transform text-based data into the numeric vector-based features. Therefore, we convert the training dataset via bag of words model to form the vectors representing a frequency of each word in a particular dataset. Using text feature extraction of Sklearn feature extraction module [17] transforms the training dataset to the numeric vector-based features. The output of the transformation shows in Table II. For example, for ICD-10 code "R04.0" when we used only LM technique, the output from data transformation is ['ไหล', 'อดกำเดา', 'เล'] which are represented by bag number 50, 43, and 69 with the word frequency 99, 104, and 11 respectively. The frequency means the counts of each word in a particular bag of words. Table III shows examples of ThCC which is used as test dataset. Case 1 represents ThCC "มีแก๊สในท้อง" which is separated by LM into 4 tokens including: "มี" ,"แก๊ส" ,"ใน" , and "ท้อง" , and means "feeling of gas in stomach" in English. In this case the frequency of token as seen symbol "-" from "count" line is null which means that Sklearn could not count the tokens. However applying Sklearn feature extraction module to case1 shows the different outputs because the limitation of Sklearn for Thai language leads to the unreadable Thai tokens such as "แก", "อง", "ใน". Although Sklearn gives the unreadable tokens as the results, it still works for our study because those tokens are represented by the numeric vector-based features. Unlike 2LT, the frequency of tokens "แก", "อง", and "ใน" in case 1 are 11, 30, and 97. Furthermore, such tokens have put into bag number 69, 34, and 72 in consequence but put into bag number 76, 38, and 80 in LM technique.

TABLE III: EXAMPLES OF THAI TEXT-BASED CHIEF COMPLAINTS USING LONGEST MACTHING TECHNIQUE

Case	ThCC	LM technique		2LT technique	
		Tokens	*Bag of words*	*Tokens*	*Bag of words*
1	มีแก๊สในท้อง	มี(T1)| แก๊ส(T2)| ใน(T3)| ท้อง(T4)	Text: ['แก', 'อง', 'ใน'] Index: [76,38,80] Count: -	มี(T1)| แก๊ส(T2)| ใน(T3)| ท้อง(T4)	Text: ['แก', 'อง', 'ใน'] Index: [69,34,72] Count: [11,30,97]
2	เคลื่อนที่ลำบาก	เคลื่อนที่(T1)| ลำบาก(T2)	Text: ['อนท', 'เคล', 'ลำบาก'] Index: [44,57,29] Count: [1,1,33]	เคลื่อนที่(T1)| ลำ(T2)| บาก(T3)	Text: ['อนท', 'เคล', 'บาก', 'ลำ'] Index: [36,49,17,25] Count: [1,1,231,418]
3	เจ็บหน้าอกขณะหายใจ	เจ็บ(T1)| หน้าอก(T2)| ขณะ(T3)| หายใจ(T4)	Text: ['ขณะ', 'าอก', 'หน', 'เจ', 'หายใจ'] Index: [2,55,35,58,36] Count: [801,1,12,61,100]	เจ็บ(T1)| หน้าอก(T2)| ขณะ(T3)| หาย(T4)| ใจ(T5)	Text: ['ขณะ', 'าอก', 'หน', 'เจ', 'หาย', 'ใจ'] Index: [3,47,30,50,31,70] Count: [991,26,1,99,501,899]
4	มีอาการปวดที่ศีรษะ	มี(T1)| อาการ(T2)| ปวด(T3)| ที่(T4)| ศีรษะ(T5)	Text: ['รษะ', 'อาการ', 'ปวด'] Index: [27,51,17] Count: [1,554,231]	มี(T1)| อาการ(T2)| ปวด(T3)| ที่(T4)| ศีรษะ(T5)	Text: ['รษะ', 'อาการ', 'ปวด'] Index: [24,43,18] Count: [1,104,130]
5	เป็นไข้จามและในเสลดมีเลือด	เป็นไข้(T1)|จาม(T2)| และ(T3)| ใน(T4)| เสลด(T5)| มี(T6)| เลือด(T7)	Text: ['นไข', 'และ', 'ใน', 'จาม', 'เสลด', 'อด', 'เป', 'เล'] Index: - Count: -	เป็น(T1)| ไข้(T2)|จาม(T3)| และ(T4)| ใน(T5)| เสลด(T6)| มี(T7)| เลือด(T8)	Text: ['ไข', 'และ', 'ใน', 'จาม', 'เสลด', 'เป', 'อด', 'เล'] Index: [72,69,71,9,66,57,35,61] Count: [97,11,915,321,61,1,15, 123]

C. Identify ICD-10 Code to Sign and Symptom from ThCC

In real life a ThCC may refer to more than one symptom hence the training and testing dataset can be classified into more than one ICD-10 code. Consequently, the machine learning needed to find the best ICD-10 code for ThCC. Unlike ThCC, ICD-10 SAS is shorter and cleaner data than ThCC composing with common words, modifiers, etc. so ICD-10 SAS can be assigned only one ICD-10 code. We

apply classifiers from Sklearn, providing 5 classifiers: decision tree classifier called CART or Classification & Regression Trees methodology, K-mean neighbours classifier, radius neighbours classifier, random forest classifier, and extremely randomized tree classifier, to identify ICD-10 code for ThCC. The results from each classifier also have been used to evaluate 2LT our proposed method for separating tokens in ThCC.

IV. Experimental Results

We used different classifiers including: (1) CART, (2) K-mean neighbours classifier, (3) radius neighbours classifier, (4) random forest classifier, and (5) extremely randomized tree classifier to categorize the separated words of ThCC by two tokenizing techniques : LM and 2LT. As shown by Table IV and V, we found that by using our proposed 2LT with different classifiers the accuracy of ICD-10 code identification is better than LM. For example, ThCC represented by case1 in table IV is classified as "R07.4 and R14", "null" , "null" , "null" , "R07.4 and R14" using the classifier (1), (2), (3), (4), (5) respectively. This result means classifiers (1) and (5) categorize the chief complaint into ICD-10 code R07.4 and code R14 but the correct code is R14 represented in "Expected ICD-10 code" column. Although classifiers (2), (3), and (4) cannot find out the code represented by "null", those classifiers give the correct codes as the results by using 2LT technique.

TABLE IV: Results of Classification Using LM Technique

Case	*Expected ICD-10 code*	**LM technique**				
		(1)	*(2)*	*(3)*	*(4)*	*(5)*
1	R14	R07.4 R14	null	null	R14	R07.0 R14
2	R26.3	R26.3 R30.0	null	null	R30.0	R26.3 R30.0
3	R07.1	R07.1 R07.4 R30.9	null	null	null	R07.1 R30.9
4	R51	R25.1	null	null	null	R25.1
5	R50.9, R06.7, R04.2	R06.7 R07.0 R31	R04.2	R04.2	R31	R06.7 R31

TABLE V: Results of Classification Using 2LT Technique

Case	*Expected ICD-10 code*	**2LT technique**				
		(1)	*(2)*	*(3)*	*(4)*	*(5)*
1	R14	R14	R14	R14	R14	R14
2	R26.3	R26.3 R30.0	null	null	R26.3 R30.0	R30.0
3	R07.1	R07.1 R30.9	null	null	R07.1 R07.4	R07.1
4	R51	R25.1 R51	null	null	R51	R25.1 R51
5	R50.9, R06.7, R04.2	R06.7 R31 R50.9	R04.2	R04.2	null	R04.2

In our experiment we have calculated precision, recall, and F1 for different classifiers. Table VI shows that CART is the classifier which is suitable for our test dataset which results into 85% precision, 71%, recall and 76% F1-measure. Although the recall and F1-measure are not good, we infers that as high as 85% of ThCC tagged by ICD-10 SAS corpus can been classified by CART. The precision is expected as the average quality of the two-level tokenization and the classifiers.

TABLE VI: Results from Five Classifiers

Classifier	**2LT technique**		
	precise	***recall***	***F1***
CART	0.85	0.71	0.76
K-mean Neighbors	0.10	1.0	0.18
Radius Neighbors	0.10	1.0	0.18
RandomForest	0.38	0.79	0.52
Extremely randomized tree classifier	0.66	0.67	0.63

V. Conclusion

The aim of this paper is applying our propose methods of tokenization to Thai text-based chief complaint which is the test dataset in order to determine an appropriate standard code of sign and symptom named ICD-10. Our method starts with separating chief complaint in Thai text into a word or token by using longest matching technique at the first step, and then using two-level tokenization technique to improve the outputs. Improvement the output is a sub-process of 2LT that the words or the tokens representing the keywords in sign and symptom domain are tagged by POS from ICD-10 SAS corpus. After that manual correction is needed to recheck the tagged words before transforming such tagged words into the numeric vector-based features. We have applied the five different classifiers to identify code in ICD-10 code for the tagged words.

We have shown that CART outperformed the others so we implies that applying our proposed two-level tokenization is suitable for CART to find out the appropriate ICD-10 code assigning to ThCC.

There are some limitations of this study. Firstly, our work needs manual correction in some process so this is a challenge to overcome in future work by developing a method for the automatic correction. Secondly, the samples used as our training dataset covering the sign and symptom description appearing in group R00 to R69. In our future work we plan to extend the group of training dataset to enhance ICD-10 SAS corpus. Finally, we have implemented our system by Python which is not support Thai language then we need more study in this aspect.

References

[1] M. Song, W. C. Kim, D. Lee, G. E. Heo, and K. Y. Kang, "Entity and relation extraction for public knowledge discovery," *Journal of Biomedical Informatics*, vol. 57, pp. 320-332, 2015.

[2] ICD-10. (October 31, 2017). Wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/ICD-10

[3] P. Saeku and J. Duangsuwan, "Signs and symptoms tagging for Thai chief complaints based on ICD-10," in *Proc. Int. Conf. on Algorithms, Computing and Systems*, 2017, pp. 44-49.

[4] J. Dara, J. N. Dowling, D. Travers, G. F. Cooper, and W. W. Chapman, "Evaluation of preprocessing techniques for chief complaint classification," *Journal of Biomedical Informatics*, vol. 41, no. 4, pp. 613-623, 2008.

[5] D. Travers and S. Haas, "Evaluation of emergency medical text processor, a system for cleaning chief complaint text data," *Academic Emergency Medicine*, vol. 11, no. 11, pp. 1170-1176, 2004.

[6] H.-M. Lu, D. Zeng, L. Trujillo, K. Komatsu, and H. Chen, "Ontology-enhanced automatic chief complaint classification for syndromic surveillance," *Journal of Biomedical Informatics*, vol. 41, no. 2, pp. 340-356, 2008.
[7] M. Jiang, Y. Chen, M. Liu, S. T. Rosenbloom, S. Mani, J. C. Denny, and H. Xu, "A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 601-606, 2011.
[8] Y. Chen, T. A. Lasko, Q. Mei, J. C. Denny, and H. Xu, "A study of active learning methods for named entity recognition in clinical text," *Journal of Biomedical Informatics*, vol. 58, pp. 11-18, 2015.
[9] V. Sornlertlamvanich, T. Charoenporn, and H. Isahara, "ORCHID: Thai part-of-speech tagged corpus," National Electronics and Computer Technology Center Technical Report, pp. 5-19, 1997.
[10] V. Sornlertlamvanich, N. Takahashi, and H. Isahara, "Building a Thai part-of-speech tagged corpus (ORCHID)," *Journal of the Acoustical Society of Japan (E)*, vol. 20, no. 3, pp. 189-198, 1999.
[11] A. Aw, S. M. Aljunied, N. Lertcheva, and S. Kalunsima, "TaLAPi — A Thai linguistically annotated corpus for language processing," *LREC*, pp. 125-132, 2014.
[12] ICD-10-TM. (2016). Thai Health Coding Centre. [Online]. Available: http://thcc.or.th/ICD-10TM/
[13] Thai Lexeme Tokenizer. The National Electronics and Computer Technology Center (NECTEC). [Online]. Available: http://www.sansarn.com/lexto/
[14] Raj, "Free Medical Spell Checker for Microsoft Word, Custom Dictionary," RAJ & CO, August 3, 2009.
[15] Chief Complaint. (2011). [Online]. Available: https://www.gotoknow.org/posts/402169
[16] Patient Interviewing & History Taking. (2013). [Online]. Available: http://thainurseclub.blogspot.com
[17] scikit-learn. (February 1, 2010). [Online]. Available: http://scikit-learn.org/stable/

Jarunee Duangsuwan was born on May, 1975 in Hat Yai, Thailand. She received a B.S. degree in computer science from Chiang Mai University, Thailand and a M.S. from Prince of Songkla University, Thailand, and her Ph.D. in computer science from the University of Reading, UK in 2012. She is currently a lecturer of the Department of Computer Science at Prince of Songkla University where she has been since 2007. Her research interests in artificial intelligence especially innovative methods based on natural language processing and machine learning technique for smart residence and smart healthcare.

Pawin Saeku was born at Hat Yai, Songkla, Thailand, on June 4th, 1989. He graduated the bachelor degree at Computer Science Department Prince of Songkla University (PSU). He is currently a master degree student under supervision of Dr. Jarunee Duangsuwan at the Department of Computer Science at Prince of Songkla University. He is interested in natural language processing (NLP). Currently he is doing the research on Thai medical text manipulation.