Thai Personal Named Entity Extraction without using Word Segmentation or POS Tagging

P. Sutheebanjard and W. Premchaiswadi

***Abstract*—Named Entity (NE) extraction for Thai language is a difficult and time consuming task because sentences in Thai language are composed of a series of words formed by a stream of characters. Moreover, there are no delimiters (blank space) to show word boundaries. Currently, most named entity extraction methods for Thai language are associated with word segmentation and Part of Speech (POS) tagging processes. The accuracy of named entity extraction is mostly affected the efficiency of those processes. At present, it is still lack of suitable methods for identifying the boundary of word for Thai sentence. Therefore this paper proposes the method to extract Thai personal named entity without using word segmentation or POS tagging. The proposed method is composed of 3 steps. Firstly, pre-processing, this process is used to remove non alphabet such as parenthesizes and numerical. Then, personal named entity is extracted by using contextual environment, front and rear, of personal name. Finally, post-processing, a simple rule base is employed to identify personal names. The training corpus of 900 political news articles and the test corpus of 100 political news, 100 financial news and 100 sport news articles were used in the experiments. The results showed that the F-measures in political and financial domain are 91.442% and 91.720% respectively which are nearly the same as in [5]. However, the proposed scheme used neither word segmentation nor POS tagging process that can significantly reduce the effort and speed up the process in building the training corpus.**

I. Introduction

NAMED entity (NE) extraction is the task that identifies expressions such as the names of people, place and organization. Named entity extraction task was first introduced in the 6th Message Understanding Conference (MUC-6) in 1995 as a subtask of MUC [4]. The Named entity task in MUCs consists of three subtasks for recognizing entity names (person, organization, and location name), temporal expressions (date and time), and numerical expression (monetary values and percentages). Especially on person named entity, it is an important in some kind of application like a people search engine such as pipl.com. Therefore, this paper is focus on the entity names which are categorized into person names.

Named entity (NE) extraction is one of the basic and important tasks performed in several processes such as information retrieval (IR), information extraction (IE), question answering (QA), machine translation (MT), etc. The main task of named entity recognition (NER) is to automatically extract interested named entities from text document. In English language, the capital letter at the beginning of a personal name is an important clue to spot name. Furthermore, the surrounding context can also be used to improve the accuracy as well. Unfortunately, this does not work with personal name in Thai language because there is no capital letters in Thai language. Thai language is similar to other Asian languages such as Chinese, Japanese and Korean that have no explicit word boundary delimiters (blank space) to indicate word boundaries. Because of this absence of explicit word delimiters and unclear definition of Thai words, it is difficult to use the results generated from difference Thai word segmentation systems. Aroonmanakun [1] proposed some ideas as a guideline of Thai word segmentation, but now there is still no standard for Thai word segmentation yet.

Several research studies have proposed the methods to efficiently extract personal name. Chen *et al.* [2] proposed a statistical approach to learn the characteristics of personal name in form of probabilistic regular grammars with rule-based of Chinese personal names. Then the learned models were used to identify personal names from several kinds of news. Yui-Jie Zhang and Tao Zhang [3] presented a method that based on Maximum Entropy (ME) principle for automatic Chinese personal name and location name recognition. They extracted contextual features from the training corpus and employed the Maximum Entropy principle to train the features. Finally, the trained features together with a Dynamic Word List and a simple Rule Base were used to recognize Chinese personal names and the location of names in the test corpus.

In Thai language, Chanlekha and Kawtrakul [4] proposed an approach to extract Thai multiword named entity by using combination of rule-based, dictionary-based and statistical-based models to predict boundaries of named entities. The method applied Maximum Entropy model and incorporate knowledge, which are rules and dictionary to named entity extraction system such as extracting personal names, locating personal names and organization names. Charoenpornsawat *et al.* [5] proposed an approach to identify Thai proper names partially composed of known and unknown substrings using probabilistic trigram models and the Winnow algorithm. The features used in the method were context words, collocations and part of speech (POS)

Manuscript received August 8, 2009.

P. Sutheebanjard is with the Graduate School of Information Technology, Siam University 235 Petchkasem Road, Phasi-charoen, Bangkok 10163, Thailand (e-mail: mr.phaisarn@gmail.com).

W. Premchaiswadi is with the Graduate School of Information Technology, Siam University 235 Petchkasem Road, Phasi-charoen, Bangkok 10163, Thailand (e-mail: wichian@siam.edu).

978-1-4244-4139-6/09/$25.00 ©2009 IEEE

tag, as well as heuristics information from dictionary and POS to generate named entity candidates to solve named entity boundary problem. The accuracy presented in the system was 92.17%.

Previous works in Thai named entity can be divided into 2 groups as follows:

1) Using word segmentation such as: [4].

2) Using word segmentation and POS tagging such as: [5]–[7].

The corpus used in all experiments was manually segmented into words and the POS was tagged by linguists which are time consuming and require a lot of linguist's skill. The results from their experiments therefore depend on the efficiency of word segmentation and POS tagging.

This paper proposes a method to reduce time and effort in building training corpus and eliminate the using of word segmentation and POS tagging problems by using plain text as the input of the system. The proposed method extracts personal named entity by applied contextual environment of Thai personal name to compare against the word list. The word list is a list of words that could possibly have the same contextual as Thai personal name, and then apply a simple rule base to recognize Thai personal name from plain text.

II. The Introduction to Thai Language

A. Characteristics of Thai Language

Thai words are multi-syllabic words which stringing together could form a new word as shown in equation (1) and (2). Since Thai language has no word delimiters, Thai morphological processing is mainly to recognize word boundaries instead of recognizing a lexical form from a surface form as in English.

Let *C* be a sequence of characters

$$C = c_1 c_2 c_3 ... c_n ; n >= 1 \quad (1)$$

Let *W* be a sequence of words

$$W = w_1 w_2 w_3 ... w_m ; m >= 1 \quad (2)$$

Where $w_i = c_{i1} c_{i2} ... c_{ir} ; i >= 1, r >= 2$

Since Thai sentences are composed of sequence of words formed by a stream of characters, i.e., $c_1c_2c_3…c_n$ with out explicit delimiters. The word boundary in "$c_1c_2c_3c_4c_5$" pattern as shown below can have two ambiguous forms: One is "c_1c_2" and "$c_3c_4c_5$". Another one is "$c_1c_2c_3$" and "c_4c_5" [8]

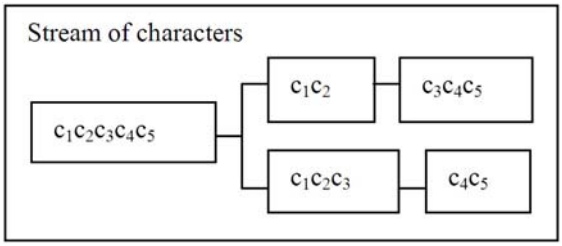

Fig. 1. Word boundary ambiguity.

In figure 1, if characters were grouped differently, the meaning of words will be different too. For example, given a string "ตากลม", there are two possible segmentation sequences: ตา-กลม 'round eye' and ตาก-ลม 'to be exposed to the wind'. Another example as in [1] is:

--[ตู้[เสื้อผ้า[สีขาว]]] 'closet for white clothes'.

--[[ตู้[เสื้อผ้า]]สีขาว] 'clothes closet that is white'

The above examples indicate that the results of named entity extraction method which uses word segmented as input have an influence on the correctness of word segmentation.

B. Thai Personal Name's Characteristic

Although, a single stream of characters in Thai language has no delimiters (blank space) to show word boundary. However, Thai personal names usually start with title that can be used as clue [4].

In general, native Thai personal names are composed of title (with or without space) + first name + one space + last name. And the foreign name composed of title (with or without space) + first name + one space + middle name + one space + last name as shown in figure 2 and table I. In table I, Thai name in row number 1 is composed of first name and last name, while in row number 2 is composed of first name and last name with 2 spaces inside. Row number 3 show a foreign name that composed of first name, middle name, and last name. Therefore, there is no simple rule to determine the boundary of personal name in Thai language.

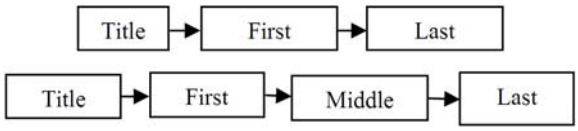

Fig. 2. Regular grammar for personal name written in Thai language.

TABLE I
EXAMPLE OF PERSONAL NAME WRITTEN IN THAI LANGUAGE.

No	Thai name	First name	Middle name	Last name
1	พระนาย สุวรรณรัฐ	พระนาย		สุวรรณรัฐ
2	จีระสิทธิ์ กุญชร ณ อยุธยา	จีระสิทธิ์		กุญชร ณ อยุธยา
3	จอร์จ ดับเบิ้ลยู บุช	จอร์จ	ดับเบิ้ลยู	บุช

Unfortunately, sometimes the title clue is not obtained correctly because the title in Thai language can be subset of other words such as "นาย" (Mr.) is subset of "ทนาย"(counsel), "นายกรัฐมนตรี"(Prime minister) and "นาย พระนาย สุวรรณรัฐ"(the first "นาย" is Mr. but the second "นาย" is subset of the first name).

Usually, the first name in Thai language does not have any space inside, but last name may be possible to contain one or more space(s) as shown in table I. This means that the space can not be used to identify the boundary of Thai last name.

III. THAI PERSONAL NAMED ENTITY EXTRACTION

From the observation of the domain corpus of 1000 Thai political news, we found that the content of Thai political news usually refers to a person by using first name, nick name or even full name (with or without title + first name + one space + last name). In addition, Thai political news always refers to a person with full name at least once in each news article.

The proposed system collects the front context and rear context features of full name and using this feature to specify full personal name in Thai. The proposed system is divided into 2 modules: front and rear context extraction modules and Thai personal named entity extraction module.

A. Front and Rear context extraction module

Thai personal name dataset is created manually from 900 political news articles. This dataset consists of 1487 Thai personal names. Then these named entities are used to create a list of front context and rear context as shown in figure 3. The range of front context and rear context is varying from 5 characters to 20 characters as shown in figure 4.

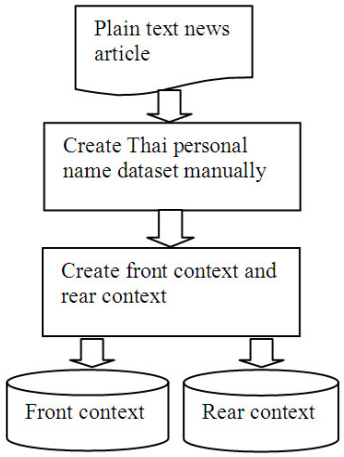

Fig. 3. Front and rear context extraction module.

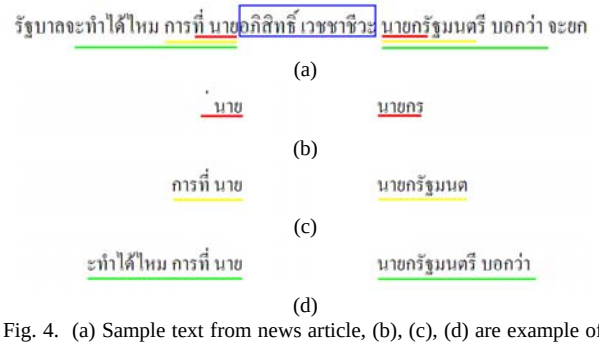

Fig. 4. (a) Sample text from news article, (b), (c), (d) are example of front and rear context extracted from (a) with 5, 10, and 20 character respectively.

Figure 4 (a) shows a sample text from news article with the personal named entity surrounding with blue color, front and rear context underline with another color. Figure 4 (b), (c) and (d) show front context and rear context extracted from figure 4 (a) with 5, 10 and 20 characters respectly.

B. Thai personal named entity extraction module

This module composed of 3 steps (as shown in figure 5). Firstly, pre-processing step, this process is used to remove numeric, non alphabet such as parenthesizes and escaped characters such as \n (new line character). Secondly, extraction process, it is used to extract personal name. Finally, post-processing, the process is used to remove non personal name by applying rules that full personal names always have such as at lease one space between first name and last name, and also remove non personal name by comparing against non personal name word list (192 words of verb). Word list was created from verbs that usually follow the first name such as กล่าว 'say', ชี้แจง 'to explain', and ตรวจสอบ 'check'.

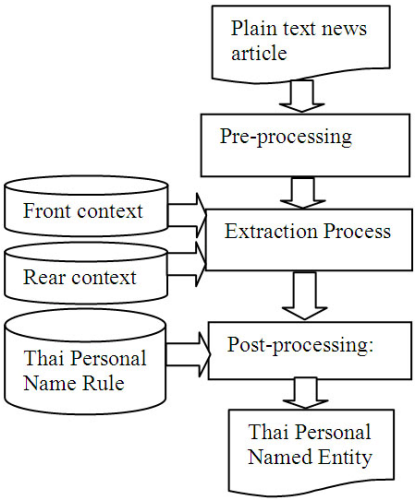

Fig. 5. Thai personal named entity extraction module.

IV. EXPERIMENTAL RESULTS

The training corpus of 900 political news articles and the test corpus of 100 political news, 100 financial news and 100 sport news articles were used in the experiments. Detail of the training and test data show in table II and table III.

The measurement used for this experimental setting was Precision (P), Recall(R), and F-measure (F). They were computed for evaluating the performance of the proposed method by using equation (3), (4), and (5) respectively.

$$P = \frac{TP}{TP + FP} \tag{3}$$

$$R = \frac{TP}{TP + FN} \tag{4}$$

$$F = \frac{2PR}{P + R} \tag{5}$$

Whereas TP (True positives) are examples correctly labeled as positives. FP (False positives) refers to negative examples incorrectly labeled as positive. TN (True negatives) corresponds to negatives correctly labeled as negative. Finally, FN (false negatives) refers to positive examples incorrectly labeled as negative [9], as shown in table IV.

. TABLE II
CORPUS STATISTICS

	Training Corpus	Test Corpus		
	Political	Political	Financial	Sport
No. of news article	900	100	100	100
Avg. character per news	2999	3818	2172	1972
No. of unique Thai personal named entity	1487	325	147	331
No. of times that Thai personal named entity appeared in corpus	5357	783	199	664
Start Date	2006-03-14	2006-11-07	2008-11-28	2009-03-15
End Date	2006-11-06	2006-11-27	2008-12-22	2009-07-26

TABLE III
PERSONAL NAME STATISTICS.

	In Training Corpus		Not in Training Corpus		Total	
	Unique	Appear	Unique	Appear	Unique	Appear
Political	191	613	134	170	325	783
Financial	13	26	134	173	147	199
Sport	12	17	319	647	331	664

TABLE IV
CONFUSION MATRIX

	Actual positive	Actual negative
Predicted positive	TP	FP
Predicted negative	FN	TN

The average performance of the proposed method that composed of the length of front and rear context vary between 5 and 20 characters are shown in table V, VI, VII and figure 6, 7, 8.

TABLE V
METHOD PERFORMANCE IN POLITICAL DOMAIN

Context Range	P(%)	R(%)	F(%)
5 characters	77.361	89.794	83.115
6 characters	83.704	89.415	86.466
7 characters	91.190	**91.696**	**91.442**
8 characters	92.263	87.830	89.992
9 characters	94.535	84.949	89.486
10 characters	95.569	82.668	88.652
11 characters	95.978	80.643	87.645
12 characters	95.850	76.289	84.958
13 characters	95.282	73.344	82.886
14 characters	95.542	71.096	81.526
15 characters	95.697	68.491	79.840
16 characters	95.926	65.667	77.964
17 characters	96.302	62.727	75.971
18 characters	**96.631**	60.990	74.781
19 characters	95.887	56.910	71.427
20 characters	94.604	54.696	69.316

TABLE VI
METHOD PERFORMANCE IN FINANCIAL DOMAIN

Context Range	P(%)	R(%)	F(%)
5 characters	88.059	**90.745**	89.381
6 characters	88.014	88.022	88.018
7 characters	**92.784**	90.681	**91.720**
8 characters	89.060	84.164	86.543
9 characters	85.071	79.952	82.432
10 characters	81.649	72.576	76.846
11 characters	75.621	64.073	69.369
12 characters	71.188	58.452	64.195
13 characters	67.465	51.715	58.549
14 characters	60.106	42.974	50.116
15 characters	53.191	37.318	43.863
16 characters	50.000	32.609	39.474
17 characters	47.872	30.659	37.379
18 characters	35.638	20.091	25.696
19 characters	31.383	17.042	22.089
20 characters	26.596	14.754	18.980

TABLE VII
METHOD PERFORMANCE IN SPORT DOMAIN

Context Range	P(%)	R(%)	F(%)
5 characters	**37.424**	**21.459**	**27.277**
6 characters	34.038	18.807	24.227
7 characters	30.368	15.900	20.872
8 characters	26.759	13.667	18.093
9 characters	26.759	13.593	18.028
10 characters	19.543	10.164	13.373
11 characters	15.935	8.711	11.265
12 characters	14.902	7.567	10.037
13 characters	14.386	7.052	9.464
14 characters	13.355	6.765	8.981
15 characters	12.324	6.508	8.518
16 characters	12.324	6.250	8.294
17 characters	11.293	6.078	7.903
18 characters	8.201	3.827	5.219
19 characters	8.201	3.827	5.219
20 characters	8.201	3.827	5.219

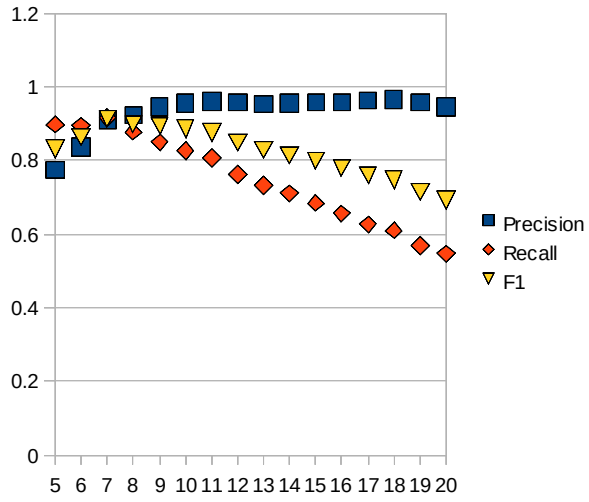

Fig. 6. Methodology performance in political domain.

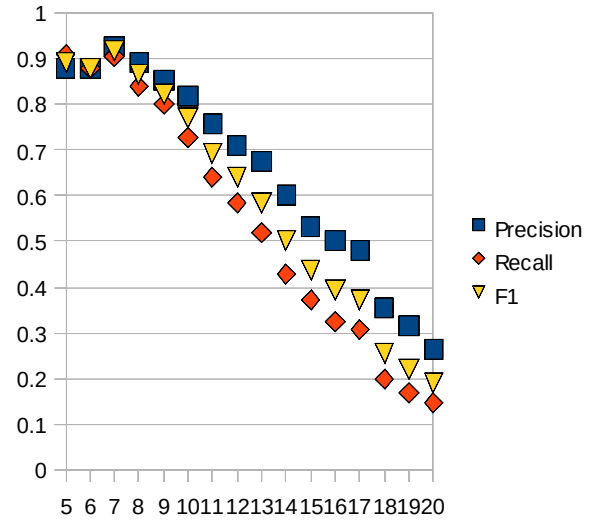

Fig. 7. Methodology performance in financial domain.

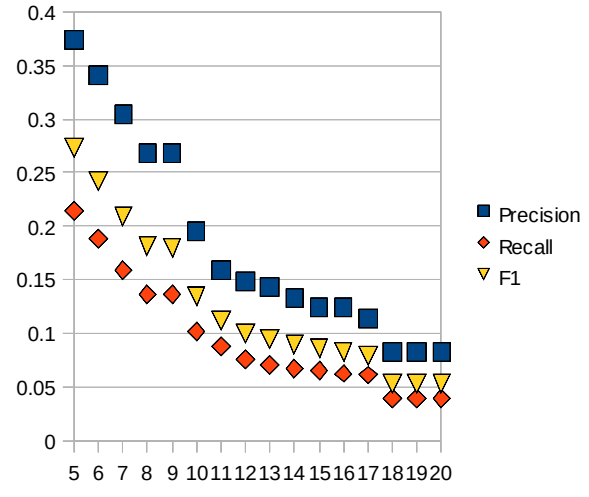

Fig. 8. Methodology performance in sport domain.

A. *Discussion of the Result*

The results of political domain in table V show that the contextual of 18 characters in length achieve the highest precision and the length of 7 characters achieves the highest recall and F-score. The results of financial domain in table VI show that the contextual of length 5 characters achieve the highest recall and the length of 7 characters achieves the highest precision and F-score. Therefore the length of 7 characters is recommended as the optimum length for extracting Thai personal named entity.

The results also show that the data training in political domain can be used to extract personal named entity both in political domain and financial domain.

However, the results of sport domain in table VII show that contextual of all characters in different context range (5-20 characters) yield the low precision, recall and F-score. Therefore the data training in political domain cannot be used to extract personal named entity in sport domain.

B. *Discussion of the Error*

The error occurred in the experiment can be classified as two types. The first type is the system that cannot extract the personal named entity as shown in table VIII. The second type is the system that extracts the incorrect personal named entity as shown in table IX.

TABLE VIII
THE ERRORS FOR NOT BEING ABLE TO EXTRACT PERSONAL NAMED ENTITY

Case	Front Context length of 7 characters	Name	Rear Context length of 7 characters
1	ะที่นาย	ฟิล ฟลินน์	นักวิเค
2	ย่างนาย	เกียรติ สิทธิอมร	น่าจะเห
3	บาท นาย	ธนพงษ์ พงษ์วัฒนา	20 ล้าน
4	กุล นาย	เกียรติ สิทธิอมร	นายสรรเ

The errors as shown in table VIII occur because the rear contexts are unknown. The characters of rear context in the case of 1, 2, 3 and 4 are part of the position in (1), general word such as verb or adjective in (2), numeric in (3) and title (Mr.) followed by other person names in (4) respectively. The errors in case one and two could be fixed by using more training data sets. The errors in cases three and four cannot be fixed by the proposed algorithm. Therefore the further study is necessary for fixing in these two cases.

TABLE IX
THE ERRORS FOR EXTRACTING INCORRECT PERSONAL NAMED ENTITY

Case	Front Context length of 7 characters	Name	Rear Context length of 7 characters
1	ั้น นาย	วิเชฐ ตันติวานิช รอง	ผู้จัดก
2	กจากนาย	โอฬารจะเชิญประชุม กรอ.	เป็นการ
3	ฐมนตรี	จากคนนอกในทีมเศรษฐกิจ ผม	เชื่อว่

The errors shown in table IX can be classified into two cases. The first case (case one), there is some characters append from the personal named entity. The second case (case two and three), the system gets the non personal named entity. The first case can be fixed by using more training data sets. The second case can be fixed by adding non personal named entity word list (explained in section III). In this case, the character "กรอ." and "ผม" should be added into non personal named entity word list.

V. CONCLUSION

In this study, front and rear context were introduced to automatically extract Thai personal named entity from plain text of political, financial and sport news articles. The experimental results show that the training corpus in political domain can be used to extract personal named entity both in political domain itself and financial domain. However it cannot be use to extract personal named entity in sport domain because the front and rear context of personal named entity in the political and financial domains are similar but totally different forms that of in sport domains.

From the political and financial domains, the optimum length for front and rear context is 7 characters. The F-measure of the proposed method in political and financial domain is 91.442% and 91.720 respectively. Although the accuracy of our system is somewhat slightly lower than the existing method (92.17%, [5]). However, the proposed method can be used to simplify the implementation process of finding personal named entity. In addition, the method neither uses the features of word segmentation nor POS tagging, so it greatly reduce time and effort used in building the training corpus significantly. Therefore it also eliminates the effects on the efficiency of using word segmentation and POS tagging on named entity finding.

REFERENCES

[1] W. Aroonmanakun, "Thoughts on Word and Sentence Segmentation in Thai," in International Symposium on Natural Language Processing (SNLP), Thailand, 2007, pp. 85-90.

[2] Z. Chen, L. Wenyin, and F. Zhang, "A New Statistical Approach to Personal Name Extraction," in Proc. ICML, 2002, pp.67-74.

[3] Y. Zhang, T. Zhang, "Me-Based Chinese Person Name and Location Name Recognition Model," in Machine Learning and Cybernetics, 2007.

[4] H. Chanlekha, A. Kawtrakul, "Thai named entity extraction by incorporating maximum entropy model with simple heuristic information," in Proceedings of the 1st International Joint Conference on Natural Language Processing, 2004.

[5] P.Charoenpornsawat, B. Kijsirikul, and S. Meknavin, "Feature-based Proper Name Identification in Thai," in Proc. Of National Computer Science and Engineering Conference: NCSEC'98, Thailand, 1998.

[6] H. Chanlekha, A. Kawtrakul, P. Varasrai and I. Mulasas, "Statistical and Heuristic Rule Based Model for Thai Named Entity Recognition," in SNLP2002, Thailand, 2002.

[7] B. Kijsirikul, "Comparing Winnow and RIPPER in Thai Named-Entity Identification", in Proceedings of the Natural Language Processing Pacific Rim Symposium 1999(NLPRS'99), Bejing, China, 1999.

[8] A. Kawtrakul, M. Suktarachan, P. Varasai and H. Chanlekha, "A State of the Art of Thai Language Resources and Thai Language Behavior Analysis and Modeling," in Coling 2002 post-conference workshops: the 3rd Workshop on Asian Language Resources and International Standardization, Taipei, Taiwan, 2002.

[9] J. Davis and M. Goadrich, "The Relationship Between Precision-Recall and ROC Curves," Proceedings of the 23rd international conference on Machine learning, p.233-240, June 25-29, 2006, Pittsburgh, Pennsylvania