Language Sense and Ambiguity in Thai

Thepchai Supnithi[1], Krit Kosawat[1], Monthika Boriboon[1] and Virach Sornlertlamvanich[2]

Email: {thepchai, krit.kosawat, monthika}@nectec.or.th, virach@tcllab.org
[1]Information Research and Development Division,
NECTEC, Thailand Science Park,
[2]Thai Computational Linguistics Laboratory,
National Institute of Information and Communications Technology
NICT ASIA RESEARCH CENTER

112 Phaholyothin Road, Klong 1, Klong Luang,
Pathumthani, 12120 Thailand

Abstract

Since Thai writing system has no explicit word and sentence boundaries, language sense in Thai depends on how we segment them. Disambiguation by grammars cannot handle all problems because many exceptions occur in the language. Machine learning technique is then introduced to cope with the ambiguity problems. This technique, however, needs good corpora to learn. In this paper, the fundamental problems in Thai natural language processing are expressed and the research topics on problem solving are explored.

1. Introduction

Without explicit word and sentence boundaries in Thai writing system, we can usually segment them in different ways to derive different senses. Research in Thai text segmentation has been conducted in [12] and [21].

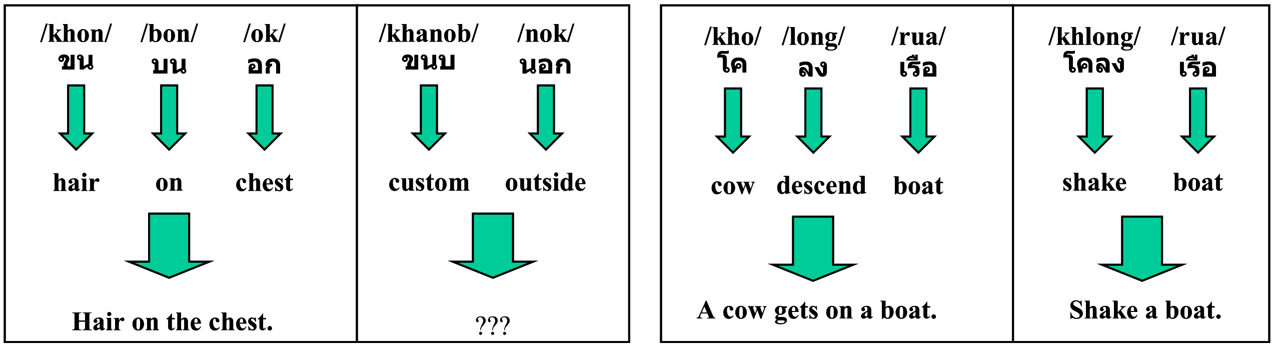
Like Chinese and Japanese, which also have no word boundary, many researches are dedicated to solve the word segmentation ambiguity [3,4,8,11]. Traditional methods are normally based on rule-based approach. Because grammatical rules cannot solve all ambiguities, machine learning technique has often been used in this area [1,5,16,20]. This technique, however, requires good corpora in order to learn and create its own rules. Our research laboratory, RDI, at NECTEC has constructed ORCHID [18] and LEX*i*TRON [7] corpora to facilitate and promote research in Thai language. In addition, RDI is active in research in Thai word segmentation [9,10,13,14,17] and Thai sentence segmentation [9, 10]. We, at present, focus in word sense disambiguation [15].

This paper surveys the research activities on language sense and ambiguity in Thailand. We will describe some problems based on word and sentence boundaries as well as some related research topics.

2. Characteristics of Thai language

There are four relevant characteristics of Thai writing style in this paper. Firstly, both single and compound words do not have explicit boundary. Secondly, a long string may be interpreted as one or more sentences as sentence ending is not clear. Thirdly, compound words and sentence identification can be controversial because some Thai compound words resemble sentences. And finally, the word sense ambiguity is explored.

2.1 Word Boundary Problem



1a) Context independent type

1b) Context dependent type

Figure 1. Example on Word boundary Problem

Word segmentation could be classified into context independent and context dependent type. In the context independent type, the meaning of the string is unambiguous; there is no need to consider contexts. For example, as shown in Fig. 1a) the string “ขนบนอก” can be segmented into three units; “ขน|บน|อก” or two units; “ขนบ|นอก”, but only the first segmentation is meaningful. In the context dependent type, several meaningful segmentations are possible. Contexts are necessary to clarify the exact meaning. For example, in Fig. 1b), the string “โคลงเรือ” can be segmented into three units; “โค|ลง|เรือ” or two units; “โคลง|เรือ”. The sequence “โค|ลง|เรือ”, which means, “A cow gets on a boat”, could be found in the sentence “พนักงาน|นำ|โค|ลง|เรือ|สาม|ตัว” (A staff brings three cows on a boat). The sequence “โคลง|เรือ”, which means, “Shake a boat”, is used in sentence “ฉัน|โคลง|เรือ” (I shake a boat).

2.2 Sentence Boundary Problem

Thai sentences are generally long and complicated. Both grammatical subjects and objects are often omitted. Space serves not only as a common hint for sentence ending but also used between phrases, proper nouns, and some specific words. To create automatic sentence segmentation, we thus need to identify sentence boundary.

The string, as shown in Fig. 2, could be segmented into three patterns: short and acceptable, subject-based and topic-based pattern. The first pattern is based on the shortest possibility by segmenting three strings between the spaces. In the scheme, the subjects of the sentences are usually omitted, for example, in the second sentence of pattern 1. The problem is that omitted units may be interpreted differently and subsequently, ambiguity occurs. The “subject-based patterns” requires that each sentence start with a grammatical subject. As in the figure, the number of sentences is reduced to two as a result. The “topic-based pattern” is constructed by detecting the most important keyword in the paragraph. It usually gets a longest sentence.

Example sentence
ผู้สูงอายุจำนวนมากยังมีคุณค่าในสังคม สามารถทำประโยชน์ให้แก่ประเทศชาติได้ แต่กิจกรรมต่างๆ ก็จะลดน้อยลงไปตามอายุและสภาพร่างกาย

Pattern 1 (Short and acceptable pattern)
1) ผู้สูงอายุจำนวนมากยังมีคุณค่าในสังคม
(Many old ages are valued member of society.)
2) สามารถทำประโยชน์ให้แก่ประเทศชาติได้
(Could give advantage to the country.)
3) แต่กิจกรรมต่างๆ ก็จะลดน้อยลงไปตามอายุและสภาพร่างกาย
(But their activities would be decreased according to their ages and physical conditions)

Pattern 2 (Subject based pattern)
1) ผู้สูงอายุจำนวนมากยังมีคุณค่าในสังคม สามารถทำประโยชน์ให้แก่ประเทศชาติได้
(Many old ages are valued member of society **and** could give advantage to the country.)
2) แต่กิจกรรมต่างๆ ก็จะลดน้อยลงไปตามอายุและสภาพร่างกาย
(But their activities would be decreased according to their ages and physical conditions.)

Pattern 3 (Topic based pattern)
1) ผู้สูงอายุจำนวนมากยังมีคุณค่าในสังคม สามารถทำประโยชน์ให้แก่ประเทศชาติได้ แต่กิจกรรมต่างๆ ก็จะลดน้อยลงไปตามอายุและสภาพร่างกาย
(Many old ages are valued member of society **and** could give advantage to the country, **although** their activities would be decreased according to their ages and physical conditions.)

Figure 2. An Example of Patterns on sentence boundary problem

Sentence boundary also causes ambiguity. For example, the string "**ยานี้ดีกินแล้วแข็งแรงไม่มีโรคภัยเบียดเบียน**" can be separated into the following ways:

(1) Three sub-sentences: "**ยานี้ดี|กินแล้วแข็งแรง|ไม่มีโรคภัยเบียดเบียน**"

(2) Four sub-sentences: "**ยานี้ดี|กินแล้วแข็ง|แรงไม่มี|โรคภัยเบียดเบียน**"

The sentence in (1) means "This medicine is good | After you drink, you will become stronger | No illness will disturb you." The sentence in (2) means "This medicine is good | After you drink, you will become stiff | Your body will have no power | Illness will disturb you." The former segmentation shows the meaning of the drug's beneficial sense, but the latter shows the meaning of the drug' harmful sense.

2.3 Word and Sentence Disambiguation Problem

In Thai, some compound words have the same form as sentence structure; as a result, there may be an ambiguity between word and sentence as shown in the following example.

The compound word "**หม้อหุงข้าว** – rice cooker" is composed of three units; "**หม้อ**-pot (noun)", "**หุง**-cook (verb)" and "**ข้าว**-rice (noun)." The word can alternatively be viewed as a sentence "A pot cooks rice," with the structure "Subject+Verb+Object."

Consider the string "**หม้อหุงข้าวสวยดี**" as a more complex example, there are two substrings "**หม้อหุงข้าว**" and "**ข้าวสวย**." Both substrings are possibly words and sentences simultaneously. This example can be interpreted as follows:

a. English: A rice cooker is beautiful.
Thai: หม้อหุงข้าว|สวย|ดี.
Sentence Pattern: Subject+Verb+Modifier
b. English: A pot is good for cooking steamed rice.
Thai: หม้อ|หุง|ข้าวสวย|ดี
Sentence Pattern: Subject+Verb+Object+Modifier
c. English: A pot can cook rice that looks delicious.
Thai: หม้อ|หุง|ข้าว|สวย|ดี.
Sentence Pattern: Subject+Verb+Object+Modifier+Modifier

The example shows all possible senses. It is clearly that all senses can be extracted from the sentence.

2.4 Word Sense Disambiguation Problem

Word Sense Disambiguation is a technique to assign appropriate meaning (or sense) to a given word in contexts. Meanings of words can be explicitly clarified by contexts. Figure 3 shows some meaning of word “กิน” as found in a corpus. It is defined in six meaning.
Quality of disambiguation depends on two issues: the richness of word meaning in the corpus and the accuracy on matching the appropriate meaning to target word in a sentence.

กิน (gin)

Sense1)	**take off**
Sense2)	**use; take; spend**
Sense3)	**corrupt; be venal; be embezzle**
Sense4)	**eat; consume; take; have**
Sense5)	**get (get a salary)**
Sense6)	**partial of some vocabularies such as “รุ้งกินน้ำ”(rainbow)”**

Figure 3. Word sense for “กิน”

2.5 Problem Complexity

It is evident that each of the characteristics of Thai explained in 2.1-2.4 is complexed on its own and it is necessary to make a clear understanding to provide any solution. To add to the scale of complexity, any combination of the characteristics is common.
Fig. 4 shows an example of the problem complexity issue. In Fig. 4a, a string “น้ำตก” can be divided into two units; “น้ำ (water)” and “ตก (to fall).” The meaning of the string is “Water falls.” In Fig. 4b, the string “น้ำตก (waterfall)” is considered as one unit. As another possible interpretation, the string “น้ำตก”, in Fig. 4c, refers to a Thai dish “spicy chopped-meat salad.” In conclusion, the string in Fig. 4a is represented in a sentence strucutre. The string in Fig 4b and 4c is represented in

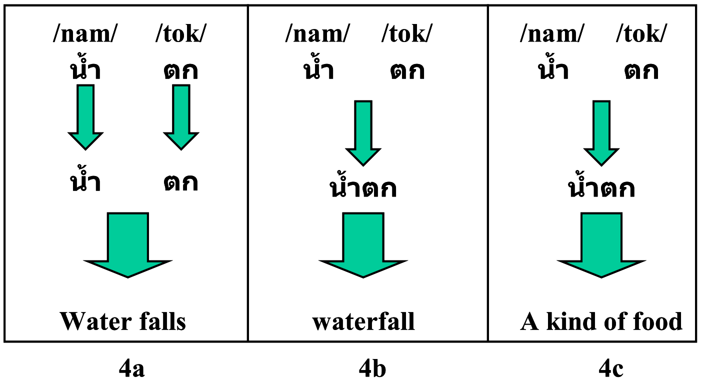


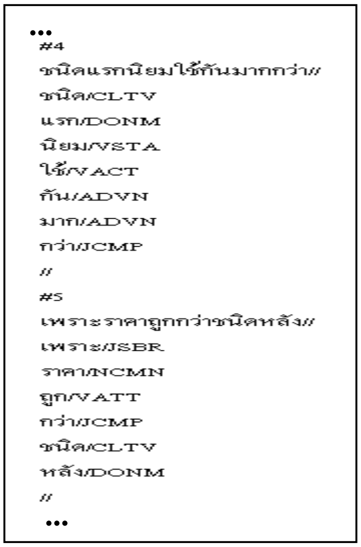
Figure 4. An example of word “waterfall”

compound word, and show a homograph relation. It is obvious that combined boundary and disambiguation is a much more difficult problem.

3. Corpora in Thai Language

To deal with mentioned characteristics of Thai Language by using machine learning, corpora are main resources. We now publicly provide ORCHID Corpus and LEXiTRON for the research purpose.

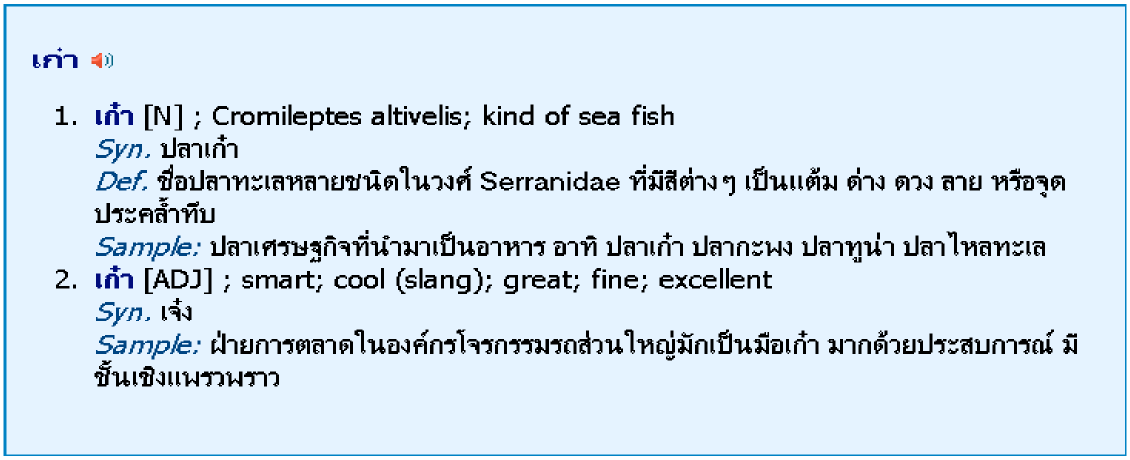
ORCHID Corpus [18] is designed for fundamental research on Machine Learning such as word and sentence boundary disambiguation and others. Texts in the corpus are split into paragraphs, sentences and words. Part-of-speech is assigned to each word. As an example, a typical set of data in ORCHID Corpus is shown in Fig. 5. The tag "#?num" shows the sentence index. A word in sentence is represented in "word/POS" form. The marker "//" is placed at end of sentence. ORCHID corpus currently collects over 400,000 words. ORCHID is useful to a number of researches related to word boundary and sentence boundary [9,10,14,17].



```
…
#4
ชนิดแรกนิยมใช้กันมากกว่า//
ชนิด/CLTV
แรก/DONM
นิยม/VSTA
ใช้/VACT
กัน/ADVN
มาก/ADVN
กว่า/JCMP
//
#5
เพราะราคาถูกกว่าชนิดหลัง//
เพราะ/JSBR
ราคา/NCMN
ถูก/VATT
กว่า/JCMP
ชนิด/CLTV
หลัง/DONM
//
…
```

Figure 5. An example of data in ORCHID corpus

LEXiTRON [7] is a public English-Thai dictionary. Its contents have been retrieved from large collection of text - such as news, literatures, essays, technical documents, and so on. LEXiTRON shows many found word senses nonexistent in traditional dictionaries. For example, the string "เก๋า" in traditional dictionary means "grouper" while the meanings in LEXiTRON are "a kind of sea fish" and "smart"as shown in Fig. 6. LEXiTRON is also applied to researches related to word sense disambiguation [15].



เก๋า

1. เก๋า [N] ; Cromileptes altivelis; kind of sea fish
   Syn. ปลาเก๋า
   Def. ชื่อปลาทะเลหลายชนิดในวงศ์ Serranidae ที่มีสีต่างๆ เป็นแต้ม ด่าง ดวง ลาย หรือจุดประคล้ำทึบ
   Sample: ปลาเศรษฐกิจที่นำมาเป็นอาหาร อาทิ ปลาเก๋า ปลากะพง ปลาทูน่า ปลาไหลทะเล
2. เก๋า [ADJ] ; smart; cool (slang); great; fine; excellent
   Syn. เจ๋ง
   Sample: ฝ่ายการตลาดในองค์กรโจรกรรมรถส่วนใหญ่มักเป็นมือเก๋า มากด้วยประสบการณ์ มีชั้นเชิงแพรวพราว

Figure 6. An example of data in LEX*i*TRON

4. Machine Learning Researches on Sense Disambiguation

At present many research projects utilize our corpora to deal with problems shown in section 2. Next, we illustrate the machine learning approach for ambiguity problem in Thai.

4.1.1 Researches on Word boundary

Word boundary is a common research topic in many Asian languages, such as Japanese, Chinese, and Thai. In Thai, word boundary detection is crucial because it is a prerequisite in various applications such as Machine Translation, Text Summarization, Question and Answering, data mining, speech synthesis, optical character recognition, and so on. Poowarawan applied the longest matching technique in his work on word segmentation [21]. This technique, however, could not find the segmentation in many cases because of its greedy characteristics. Sornlertlamvanich applied maximal matching techniques to solve the problem of the longest matching technique [17]. This algorithm first generates all possible segmentations for a sentence and then selects the one with the fewest words. The algorithm, however, could not determine the best candidate if alternatives had the same number of words. Kawtakul et al. applied a probabilistic technique for word boundary problem [2]. Statistical techniques are applied to compute the most likely segmentation. Meknavin applied feature-based methods [13,14]. He selects features and combines various kinds of them to solve word boundary problem. The accuracy of feature-based method, proposed by our research laboratory, was up to 99% for context independent and 95% for context dependent type.

From the experiment described above, the 95-99% accuracy seems sufficient for automatic word boundary. Since words in Thai are the smallest linguistics unit that can represent meaning. The accuracy of other applications that use word segmentation algorithm is highly depend the above accuracy. An increasing of only 0.01% for word boundary problem is still essential for Thai language processing.

4.1.2 Research on Sentence boundary

Thai writing system uses a space as a hint of sentence break. It can appear in the following senses [6].

1) A space is used to break between sentences.
2) A space is used to break between phrases or clauses within a sentence.
3) A space is used to break between sentences in a cohesive group of sentences.
4) A space is used to break before and after numerals.
5) A space is used to break between coordinate words in lists.
6) A space is used to break between the first name and family name.
7) A space is used to break before and after some special orthographic symbols and punctuation marks.

Sungkornsarun presented the method of splitting Thai sentences from paragraph [12]. He segmented a paragraph into morpheme-based structure. Then he estimated the number of sentences by main verbs. The conjunctions of sentences were marked to be the sentence boundary and identified by the syntactic analysis of Thai sentences. The accuracy of this approach was 81.18%. Mitrapiyanurak and Sornlertlamvanich proposed part of speech tri-gram model to identify sentence boundary [9]. The accuracy of this approach was about 85%. Charoenpornsawat and Sornlertlamvanich proposed feature-based algorithm to identify sentences in a paragraph by detecting the appropriate sentence breaking spaces [10]. The algorithm considers contexts around a space in order to determine whether the space is a sentence break or not. The accuracy of this approach, proposed by our research laboratory, was about 89%. It is obvious then that little research work on sentence boundary in Thai exist. All the approaches concentrate only on the "short and acceptable" pattern. Accuracy is too poor to apply in real applications. It is thus necessary to continue the research in the field.

4.1.3 Researches on Word Sense Disambiguation

Table 1. SVM experiment for a word "กิน"

Word (Position)	Accuracy	POS (Position)	Accuracy
Word (1)	91.1%	POS (1)	79.9%
Word (2)	89.2%	POS(2)	77.8%
Word (3)	87.1%	POS(3)	75.2%
Word (4)	86.2%	POS(4)	75.9%
Word (5)	85.1%	POS(5)	74.5%

Word Sense Disambiguation is one of the most important open topics in NLP. A lot of researches have been conducted in English, Japanese and Chinese. There are very few researches on Word Sense Disambiguation in Thai. One of the most successful current lines of research is applying Machine Learning algorithms to create statistical models in order to perform Word Sense Disambiguation. Kanokrattananukul applied Decision Lists method in Thai Word Sense Disambiguation [19]. The accuracy of classification was about 80%. In our research laboratory, Modhiran et al, illustrated three machine learning techniques, SNOW, Naïve Bayes and Support Vector Machine to perform Word Sense Disambiguation [15]. The accuracies are 73%, 81% and 84%, respectively. Each technique, however, highly depends on the selected words. For example, "กิน (kin)" as shown in Table 1, gives an accurate result up to 91%.

5. Conclusions and Future Work

In this paper, we describe the challenges of Thai language: word and sentence boundaries and word sense disambiguation. All of them are recognized as a fundamental issue in many applications, such as machine translation and question and answering system. Currently, computer-supported research on Thai language processing gradually increases, both from linguistic and computer-based viewpoints. Without the high-quality corpora, these researches cannot get good results. Designing qualitative corpora to serve not only grammatical structure but also linguistic information is necessary. Our corpora are constructed from real texts, which reflect present-day language usage behavior. It is, however, crucial to construct larger corpora to serve language sense processing. We plan to construct corpora to support automatic sentence boundary and word sense disambiguation. Moreover, applying machine learning to learn the behavior of language using from these corpora is our considering topics.

Acknowledgment

Special thanks to Mr. Teerapong Modhiran, a part-time staff and graduate student of Computer Engineering Department, King Mongkut's Institute of Technology Ladkrabang, for his value assistance in experiments.

References

[1] Andrew J. Carlson, Chad M. Cumby, Jeff L.Rosen, Dan Roth. "Snow User Guide," Cognitive Computation Group, Computer Science Department University of Illinois, Urbana/Champaign.

[2] Asanee Kawtrakul, Supapas Kumtanode, Thitima Jamjanya and Chanvit Jewriyavech. A Lexibase model for Writing Production Assistance System. In Proceedings of the Symposium on Natural Language Processing in Thailand (1995).

[3] Dai, Xiang-Ling, 1992. Chinese Morphology and its Interface with the Syntax. Ph.D. thesis, Ohio State University.

[4] Dan Roth, "Learning to Resolve Natural Language Ambiguities: A Unified Approach," Department of Computer Science University of Illinois at Urbana-Champaign Urbana, IL 61801.

[5]Jaynes E. T. (1979). "Where do we Stand on Maximum Entropy?." In Levine, R.D and Tribus, M. (Eds.), The Maximum Entropy Formalism, p15. M.I.T Press.

[6] Nantana Danvivathan. The Thai Writing System, Forum Phoneticum 39, Helmut Buske Verlag Hamburg, (1987).

[7] http://lexitron.nectec.or.th/

[8] Matsumoto, Y.; Kurohashi, S.; Myoki, Y.; et al. (1991) User's Guide for the JUMAN system - A User-Extensible Morphological Analyzer for Japanese, Nagao Laboratory, Kyoto University

[9] Pradit Mitrapiyanurak and Virach Sornlertlamvanich. The Automatic Thai Sentence Extraction. Proceedings of the Symposium on Natural Language Processing in Thailand (2000).

[10]Paisarn Charoenpornsawat and Virach Sornlertlamvanich. Sentence Break Disambiguation for Thai, In Proceeding of the 19th International Conference on Computer Processing of Oriental Languages, (2001).

[11] Sproat, R. and C. Shin, "A Statistical Method for Finding Word Boundaries in Chinese Text," Computer Processing of Chinese and Oriental Languages, vol. 4, no. 4, pp.336–351, 1991

[12] Sungkornsarun Longchupole. Thai Syntactical Analysis system by Method of Splitting Sentences from Paragraph for Machine Translation. Master Thesis. King Mongkut 's institute of technology Ladkrabang (in Thai) (1995).

[13] Suraphan Meknavin. Towards 99.99% Accuracy of Thai Word Segmentation. Oral Presentation at of the Symposium on Natural Language Processing in Thailand (1995).

[14]Suraphan Meknavin, Paisarn Charoenpornsawat, and Boonserm Kijsirikul. Feature-based Thai Word Segmentation. In Proceedings of the Natural Language Processing Pacific Rim Symposium (NLPRS'97), Phuket, Thailand (1997).

[15]Teerapong Modhiran Boontee Kruatrachue and Thepchai Supnithi. Comparison Thai Word-Sense Disambiguation Method, 2004 International Conference on Control, Automation and Systems. (to appear) (2004).

[16] Thorsten Joachims, "Text Categorization with support vector machine," In proc.Of European Conference on Machine Learning (ECML), 1998.

[17]Virach Sornlertlamvanich. Word Segmentation for Thai in a Machine Translation System (in Thai) (1993).

[18] Virach Sornlertlamvanich, Thatsanee Charoenporn and Hitoshi Isahara. ORCHID: Thai Part-Of-Speech Tagged Corpus, Technical Report Orchid TR-NECTEC-1997-001, National Electronics and Computer Technology Center, Thailand, pp. 5-19, 1997

[19] Wipharuk Kanokrattananukul. Word Sense Disambiguation in Thai Using Decision List Collocation, Master Thesis, Chulalongkorn University (2001).

[20] Yirong Shen and Jing Jiang. "Improving the Performance of Naïve Bayes for Text Classification," CS224N Spring, 2003

[21]Yuen Poowarawan. Dictionary-based Thai Syllable Separation. In Proceedings of the Ninth Electronics Engineering Conference (1986).