Collocation and Thai Word Segmentation

Wirote Aroonmanakun

Department of Linguistics
Faculty of Arts, Chulalongkorn University
Phyathai Rd., Bangkok, Thailand, 10330
Tel. +66-2-218-4696
Fax.: +66-2-218-4695
e-mail: wirote.a@chula.ac.th

Abstract

This paper presents another approach of Thai word segmentation, which is composed of two processes : syllable segmentation and syllable merging. Syllable segmentation is done on the basis of trigram statistics. Syllable merging is done on the basis of collocation between syllables. We argue that many of word segmentation ambiguities can be resolved at the level of syllable segmentation. Since a syllable is a more well-defined unit and more consistent in analysis than a word, this approach is more reliable than other approaches that use a word-segmented corpus. This approach can perform well at the level of accuracy 81-98% depending on the dictionary used in the segmentation.

1 Background

Thai word segmentation is a basic and essential issue for processing the Thai language. Since 1981, many approaches have been proposed to handle this task. Like many languages that do not have explicit word boundary, such as Chinese, word segmentation is viewed as a problem of inserting word boundaries or word separators into the input sentence. Word segmentation is difficult because, usually, there is more than one way for inserting word separators. Examples as in Figure 1 are usually used to illustrate the difficulty of Thai word segmentation.

a. ตากลม : ตา_กลม or ตาก_ลม
b. โคลงเรือ : โค_ลง_เรือ or โคลง_เรือ
c. ขนบนอก : ขน_บน_อก or ขนบ_นอก
d. มากว่า : มา_กว่า or มาก_ว่า
e. หลวงตามหาบัว : หลวงตา_มหาบัว or หลวง_ตาม_หา_บัว

Figure 1. Ambiguity of word segmentation

Although many approaches have been implemented for word segmentation, the resolution is not satisfying yet. When a Thai word segmentation program, which is based on a trigram model and a learning algorithm (Charoenpornsawat 1998), is applied on real texts from a newspaper, though the program is claimed to segment Thai words correctly more than 90%, incorrect segmentations are easily found. For example, some words, e.g. คนใช้ (servant), การเมือง (politics), เลือกตั้ง (elect), ขี้เถ้า (ash), are incorrectly segmented into two words as คน (man) ใช้ (use), การ (nom.) เมือง (town), เลือก (choose) ตั้ง (set up), ขี้ (excrement) เถ้า (ash), respectively. Some words, e.g. ผันผวน, are incorrectly segmented as meaningless words, ผัน_ผ_วน.

This paper discusses why Thai word segmentation is difficult. We will briefly review previous approaches and their underlying assumptions on the Thai language. We will point out the drawback of the trigram approach that is trained on a word-segmented corpus. Then, we will argue that most of the ambiguities on segmentation can be viewed as a problem of syllable segmentation rather than a problem of word segmentation. Thus, we will propose another word segmentation approach based on a syllable-based trigram model and maximum collocation. In short, word segmentation can be viewed as processes of segmenting syllables and merging syllables rather than as inserting word boundary. Segmenting syllables is done by

applying a trigram model of syllables. Merging syllables is done on the basis of collocation strength between syllables. The best segmentation is the one with maximum collocation strength.

2 Previous research

This section briefly reviews previous research on Thai word segmentation. We discuss the basic assumptions underlying each method and its implications on the Thai language. The rule-based approach (Thairatananond 1981, Chamyapompong 1983) is excluded because it is used for syllable segmentation rather than word segmentation. Most approaches use a dictionary as the basis, but segment texts by applying different strategies, such as longest matching (Poowarawan 1986), maximum matching[1] (Sornlertlamvanich 1993). For approaches that are corpus-based, the dictionary is already implicit, such as a trigram model (Kawtrakul et al. 1997), and a feature-based segmentation (Meknavin et al. 1997). Some do not use the dictionary to avoid the problem resulted from unknown words. (Theeramunkong et al. 2000)

These approaches reflect different assumptions about the Thai language. Longest matching and maximum matching approaches share the same view that compound words are preferred over simple words, when they are applicable. But the maximum matching prefers the overall number of words to be minimum. While these assumptions are not yet proved, they seem to be mostly correct since the performance of the maximum matching is claimed to be higher than 90% correct. For the corpus-based approach, the trigram statistics should play an important role in resolving segmentation ambiguities. However, the result relies heavily on the training corpus, which is manually word-segmented. This approach could suffer from a lack of clear definition of Thai words and inconsistency of segmentation in the training corpus. In a simple experiment, in which five subjects were asked to manually segment words on a sample text of 3,070 syllables, the result indicates that agreement on word boundaries is not perfect. The Fleiss' kappa co-efficiency (Rietveld and van Hout 1993:221-222) indicates the degree of agreement at 0.75. (The value for perfect agreement is 1, while the value for agreement by chance is 0). The formula of Fleiss's kappa is presented below:

$$k = \frac{Po - Pe}{1 - Pe}$$

with:
Po = proportion of agreeing pairs of judgments
Pe = proportion of agreeing pairs on the basis of chance

These proportions are calculated as follows:

$$Po = \frac{\sum n_{ij}^2 - Nk}{Nk(k-1)}$$

$$Pe = \sum_{j=1}^{v} P_j^2$$

$$Pj = \frac{\sum_{i=1}^{N} n_{ij}}{Nk}$$

The symbols used are:
N = number of judged objects, or the number of syllables in the text.
k = number of subjects, which is 5 in this case.
v = number of categories, which is 2 because subjects decide whether each syllable boundary is a word boundary.
n_{ij} = number of subjects who assign object i to category j.

When compared the segmentation results between each pair of subjects, the average precision, recall, and balanced F-Measure (2 * P * R / (P + R))[2] (Rijsbergen, 1979), as shown in Table 1, confirm that the agreement is not perfect. The average F-measure is only about 82%. The result suggests that word boundary may not be always intuitively determined. For a word-segmented corpus to be useful, the corpus has to be prepared by few people who are trained with the same operational criteria for word segmentation.

[1] The term 'maximum matching' may be used in the same meaning as 'longest matching' in other papers, such as Palmer (1997).

[2] Precision is the number of identical words when comparing the segmentation of subject_i with that of subject_j. Recall is counted in the opposite direction.

	Precision	Recall	F-Measure
S1-S2	86.35	84.12	85.22
S1-S3	84.77	78.29	81.40
S1-S4	83.30	88.04	85.61
S1-S5	82.04	84.74	83.37
S2-S3	93.16	88.32	90.68
S2-S4	80.54	87.38	83.82
S2-S5	77.59	82.28	79.87
S3-S4	72.62	83.12	77.52
S3-S5	73.24	81.93	77.34
S4-S5	80.14	78.33	79.22
Average	81.38	83.66	82.41

Table 1. Comparing word segmentation between each pair of subjects

3 Problems on Word Segmentation

The lack of clear definition of Thai words could cause a problem for word segmentation. Without agreement on word segmentation, a training corpus will not be very useful because a corpus that is word-segmented by different persons is not compatible or sometimes in conflict. In fact, even segmentation is carried out by the same person, it could be inconsistent. Thus, we should discuss first at the definition of Thai words. Though, in linguistics, a word is defined as a linguistic unit composing of one or more morphemes, Thai grammar books usually view a word as a composition of syllables and distinguish two types of word as follows:

1. Simple words: A simple word can have one or more syllables. In a multi-syllable word, each syllable may have a meaning, but the meaning of the word is not related to the meaning of any syllable. Examples of these simple words are นอน (sleep), อ่าน (read), สะ-พาน[3] (bridge), นา-ฬิ-กา (clock) etc.

2. Compound words: A compound word is composed of two or more simple words. The meaning of the word may not be the total composition of the meaning of its parts, though it can be related to the meaning of its parts. For example, แม่-น้ำ (river) is composed of แม่ (mother) and น้ำ (water); though the meaning of แม่-น้ำ is not 'the mother of water', it is related to water. The meaning of some compound words could be different from the meaning of its part, such as ยิน-ดี (glad) is composed of ยิน (hear) and ดี (good), หาย-ใจ (breathe) is composed of หาย (lost) and ใจ (heart). Some compound words are created by conjoining two simple words that are quite similar in meaning, such as ดู (look)-แล (see), สวย (pretty)-งาม (beautiful). Some are created by conjoining the same word, such as แดง (red) ๆ (symbol for duplication) ดำ (black) ๆ.

Although the criteria above seem to be clear, when looking at the real data, it is not always easy to determine the number of words in a given input. For example, should หม้อ-หุง-ข้าว (rice cooker) be analyzed as one compound word, or three simple words หม้อ (pot), หุง (cook), and ข้าว (rice)? If we analyze หม้อ-หุง-ข้าว as a single word by assuming it denotes a single referent, should we analyze หม้อ-หุง-ข้าว-ไฟ-ฟ้า (electric rice cooker) as a single word too, since it denotes a single referent. How about หนัง-สือ-รวม-บท-ความ-ทาง-วิ-ชา-การ-ใน-การ-ประ-ชุม-สัม-มนา, which is assigned as a translation equivalent of the word 'proceeding'? Should this string be a single word? Or should it be analyzed as composed of nine words: หนัง-สือ (book) รวม (collect) บท-ความ (article) ทาง (about) วิ-ชา-การ (academic) ใน (in) การ (nom.) ประ-ชุม (meet) สัม-มนา (seminar)?

This unclear-cut semantic criterion could be a reason why word segmentations performed by different persons, or even by the same person, can be inconsistent. Chaicharoen (2002), thus, proposes to use the uninterruptablility of a word as one criterion for determining a Thai word. A sequence of syllables is considered a word if its meaning is changed when inserting some other syllables in between. For example, หม้อ-หุง-ข้าว (rice cooker) is considered a word because when inserting ที่-ใช้-สำ-หรับ (that-is-used-for) between หม้อ (pot) and หุง-ข้าว (cook-rice) (หม้อ-ที่-ใช้-สำ-หรับ-หุง-ข้าว), the meaning is not the same. But เครื่อง-พิมพ์-ดีด-ไฟ-ฟ้า (electric-type-writer) are considered two words, เครื่อง-พิมพ์-ดีด (type-writer) and ไฟ-ฟ้า (electric), because when inserting ที่-ใช้ (that-use), the meaning of เครื่อง-พิมพ์-ดีด-ที่-ใช้-ไฟ-ฟ้า is not much different from the original one. This criterion reflects the internal cohesion of a word. Of course, we cannot apply an insertion test directly for word segmentation programs, but we may use the idea of internal cohesion as a clue for word segmentation. In this study, we use collocation strength between syllables for measuring this internal cohesion.

[3] In this paper, the symbol - is used for segmenting syllables, while _ is used for segmenting word, though, in actual text, there is no boundary markers.

4 Segmentation As Two Processes

Previous approaches on Thai word segmentation usually view the segmentation problem as the resolution of word boundary ambiguities, as shown in Figure 1. However, we think that many segmentation ambiguities can be resolved by just performing syllable segmentation. Since a syllable is a more well-defined unit than a word, it is easier and more consistent to build a syllable-segmented corpus. Therefore, we view Thai word segmentation as composing of two processes.[4] The first one is to do syllable segmentation, which could be done by applying a trigram model trained with a syllable-segmented corpus. This process should resolve many segmentation ambiguities, at least in those classic examples in Figure 1. The next process is to group syllables into words. The latter is more difficult than the first one. In this study, we use the idea of collocation to measure internal cohesion of a word. Collocation here refers to co-occurrence of syllables observed from the training corpus. It can be measured by many statistical methods, such as mutual information, chi-square, Dunning's log-likelihood, etc. (Manning and Schutze 1999) But in this study, to reflect the idea of internal cohesion, we use the ratio of the chance of finding two syllables together to the chance of finding other syllables in between the two syllables. This is discussed in section 6.

5 Syllable Segmentation

Thai syllables here are referred to written syllables only. Typically, a syllable is composed of vowel forms, initial consonants, and final consonants. In some syllables, vowel forms are omitted, like กด. Some syllables use more than one character for vowel forms, such as เสียง, เสือ, etc. Some have two initial consonants such as, กว้าง, ฉลาด, เขลา, etc. Some have more than one character for final consonant, such as จักร, ฉัตร, etc. Some syllables are unique, such as ก็, ณ, etc. Nevertheless, we can define all syllable patterns, and the number of patterns are finite. In this study, we define about 200 syllable patterns for matching an input string. For example, เCRTาะ, เCีRTอะ, เCRTY are syllable patterns in which X, C, R, Y, T stands for a different group of characters. The results after matching these syllable patterns are usually ambiguous. For example, in segmenting an input string like กรรมการกรมพลศึกษารอยกร่าง, 36 possibilities of segmentation were found. But when trigram statistics of syllables is applied, this sentence is segmented correctly as กรรม-การ-กรม-พล-ศึก-ษา-รอ-ยก-ร่าง.

In this study, a training corpus of 553,372 syllables from a newspaper is manually syllable-segmented. Witten-Bell discounting is used for smoothing (Chen and Goodman, 1998). Viterbi algorithm is used for determining the best segmentation. When tested on another corpus of 30,498 syllables, 52 errors of segmentation were found. Thus, the program can segment syllables correctly up to 99.8%. Of these 52 errors, 22 are proper names and foreign words written in Thai.

6 Syllable Merging

In this step, we assume that every syllable boundary is a potential word boundary. In a sentence that is syllable-segmented, this process will determine which boundaries can be deleted. Those that are left are regarded as word boundaries. The output then is a sentence that is word-segmented. The first design is to use collocation strength between syllables to merge syllables. The assumption is that if a word contains two or more syllables, those syllables will always co-occur. Thus, the probability of co-occurrence should be highly greater than by chance. Collocation strength between two syllables that are parts of a word should higher than collocation between two syllables that are not a part of word. For example, in a phrase เปิด_หน้า-ต่าง, which consists of two words เปิด (open) and หน้า-ต่าง (window), the collocation between หน้า and ต่าง should be higher than that of เปิด and หน้า. However, since the value of collocation strength between two syllables in any circumstances is always constant, it is insufficient to determine word boundary by considering only collocation at the boundary. For example, even ข้อ-ต่อ could be a two-syllable word, as in (2a), but it could also be a part of a multi-syllable word as in (2b) and (2c), or be two different words as in (2d), or be one word and a syllable of another word as in (2e) and (2f). Clearly, we cannot use the collocation

[4] Sawamiphakdi (1990) also did word segmentation in two steps : building syllables by rules and merging them by dictionary look-up. However, she did not use statistical method for resolving segmentation ambiguities.

a. ฟิต-เฟล็กซ์-เลือก-ทำ-ธุร-กิจ-เกี่ยว-กับ-**ข้อ-ต่อ**-และ-สาย-อ่อน-ทุก-ชนิด — **ข้อ-ต่อ** (joint)
b. ที่-นาย-ปิ่น-กล่าว-มา- เป็น-เพียง-ข้อ-อ้าง-หรือ-**ข้อ-ต่อ**-สู้-หนึ่ง-เท่า-นั้น — **ข้อ-ต่อ**-สู้ (argument)
c. ออส-เตร-เลีย อาจ-หยิบ-ยก-เรื่อง-นี้-ขึ้น-มา-เป็น-**ข้อ-ต่อ**-รอง-ปิด-ตลาด-ใน-บาง-สิน-ค้า-ของ-ไทย — **ข้อ-ต่อ**-รอง (bargaining point)
d. มอบ-นโย-บาย-5-**ข้อ-ต่อ**-ดร.-สม-บูรณ์ — **ข้อ** (clss.) **ต่อ** (to)
e. ความ-เสี่ยง-ที่-สูง-**ข้อ-ต่อ**-มา-คือ-การ-เปลี่ยน-แปลง-ของ-อัตรา-ดอก-เบี้ย — **ข้อ** (clss.) **ต่อ**-มา (next)
f. หาก-คณะ-กรรม-การ-ปล่อย-ให้-ผู้-รับ-เหมา-แข็ง-**ข้อ-ต่อ**-รา-คา-ประ-มูล — แข็ง-**ข้อ** (defy) **ต่อ** (to)

Figure 2. Examples of ข้อ-ต่อ in different contexts

between ข้อ and ต่อ to determine whether that boundary is a word boundary.

Examples as in Figure 2 do not only indicate the insufficiency of considering only syllable collocation, but also raise a question of how to determine a word? Two syllables like ข้อ-ต่อ can be either one word ข้อ-ต่อ, or two words ข้อ_ต่อ. We know whether a sequence of syllables is a word because it refers to something. In other words, we have the lexical knowledge of that word. Thus, in this study, we will use a dictionary for determining whether a sequence of syllables could be a word. Without a dictionary, the program might have to check all possible sequences of syllables. Given that a word can consist of one or more syllables, in a sequence of n syllables, there could be 2^{n-1} possible sequences of segmentation.

Therefore, in this process, a dictionary look-up is used to match all possible words from the sequence of syllables. The result after matching the input syllables with the dictionary could be ambiguous. To determine the best resolution, though we cannot use collocation between syllables directly, we may use the overall collocation strength in the sentence. It is assumed here that there is collocation strength between syllables at every syllable boundary. This strength is a force that binds syllables into a word. On the other hand, there could be a driving force that prevents one syllable from combining to another. For example, in a sequence of syllables ...a-b-c-d-e..., in which b-c-d forms a word, there are forces between b-c and c-d that combine them together, but there are also forces between a-b and d-e that prevent b from combining to c, d from combing to c. Thus, the overall collocation strength of an input sentence is defined as the sum of collocation within a word minus the collocation strength between words.

$$St = \sum_{i=1}^{n} F_{w_i} - \sum_{i=1}^{n-1} D_{w_i,w_{i+1}}$$

$$F_{w_i} = \sum_{j=1}^{k-1} C_{s_j,s_{j+1}} \text{ such that } w_i = s_1 s_2 \ldots . s_k$$

$$D_{w_i,w_{i+1}} = C_{s_j,s_{j+1}}$$

$$\text{such that } s_j \text{ is the last syllable of } w_i$$

$$s_{j+1} \text{ is the first syllable of } w_{i+1}$$

The best segmentation is the one with the maximum collocation strength. In addition, we also tested two variations of this model. The first one is to do subtraction of collocation D only when the pair of syllables could be a part of another word. For example, in a sequences ...a-b-c-d-e,..., in which b-c-d forms a word, Da,b will be subtracted only if a-b could be an ending of some words. This variation will be called MaxColl-B. (The first model is named MaxColl-A) Another variation is to ignore driving force D, or not subtracting anything. It will be called MaxColl-C.

For the collocation strength between syllables, since we hold the idea of internal cohesion for determining a word, we use the ratio of p(x,y) to q(x,y), where p(x,y) is the probability of finding syllables x and y together, and q(x,y) is the probability of finding any syllable in between x and y (x-ANY-y), or the probability for x and y to be separated by another syllable. The collocation between syllables x-y then is defined as below:

$$\log \frac{p(x,y)}{q(x,y)} = \log \frac{p(x)p(y \mid x)}{q(x)q(y \mid x)} = \log \frac{p(y \mid x)}{q(y \mid x)}$$
$$= \log \frac{Count(x,y) / Count(x)}{Count(x,Any,Y) / Count(x)}$$
$$= \log \frac{Count(x,y)}{Count(x,Any,y)}$$

7 The Results

Four segmentation algorithms, MaxColl-A, MaxColl-B, MaxColl-C, and MaxMatch (maximum matching), are tested on the testing corpus of 30,498 syllables. The results from these algorithms are compared with the manual

segmentation done by the author. In the process of syllable segmentation, the results are 100% correct. But in the process of syllable merging, the result of each algorithm is different. In a perfect situation, where every word in the testing corpus is included in the dictionary, MaxColl-A can segment words correctly with the F-measure at 96.76%, as shown in Table 2. However, in the same setting, when compared with the results from MaxColl-B, MaxColl-C, and MaxMatch, the result from MaxMatch seems to be the best.[5]

	Precision	Recall	F-ms
Max Coll-A	96.36	97.16	96.76
	19271/19998	19271/19835	Er:37
Max Coll-B	97.97	97.66	97.81
	19371/19773	19371/19835	Er:31
Max Coll-C	98.02	97.71	97.86
	19380/19772	19380/19835	Er:28
Max Match	98.56	97.39	97.97
	19317/19600	19317/19835	Er:47

Table 2: Results of the word segmentation.

MaxMatch has the best score because it produces the lowest number of words (19,600). That explains why its precision rate is high. But if we look at the recall rate, we will see that MaxColl-B and MaxColl-C produce more correct words than MaxMatch. However, the number of accuracy alone may not be the best indicator. Since the accuracy is measured against the author's manual segmentation and the manual segmentation is not always perfectly consistent. The mismatch between the manual segmentation and the segmentation from the algorithm does not necessary indicate that the algorithm's result is incorrect. For example, the program always segment กลุ่มตัวอย่าง as one word, but in the manual segmentation, even it is carried out with care and rechecked for its consistency, sometimes this string is segmented as one word, sometimes as two words. Therefore, we should compare the performances of these algorithms by examining which one produces less severe errors of segmentation.

By severe errors, we refer to segmentation that results in a wrong word. The meaning of the sentence then is incorrect. Examples in Figure 3 illustrate severe errors of segmentation from MaxMatch. In (3a), ที่มา should be segmented as two words, ที่ (comp.) and มา (come), rather than one word ที่มา (source). In (3b), มากว่า should be segmented as มา (asp.) กว่า (over), not มาก (more) ว่า (say). In (3c), ทางการเมือง should be segmented as ทาง (prep.) การเมือง (politics), not ทางการ (official) เมือง (city).

a. อดีต_รัฐมนตรี_**ที่มา**_อยู่_พรรค_ไทยรักไทย_ใน_ปัจจุบัน
b.ตัดสิน_เรื่อง_การ_ใช้_ถ้อยคำ_ใน_แดน_มะกะโรนี_**มาก**_**ว่า**_400 _ปี
c. ด้วย_การ_แข่งขัน_เข้าสู่_การ_มี_ตำแหน่ง_**ทางการ_เมือง**

Figure 3. Examples of severe errors

In terms of severe errors produced from the algorithm, MaxMatch produced 47 errors while MaxColl-A, MaxColl-B, and MaxColl-C produced 37, 31, 28 errors respectively. Thus, MaxColl-C should be the best segmentation algorithm by this criterion.

In a non-perfect situation where there are unknown words, the performance of all algorithms dropped as shown in Table 3. The dictionary used in this setting is derived from word list of another corpus. In this setting, there are 883 unknown lexical words from the total of 3,082 lexical words (in the testing corpus), or 29% of the lexicon. The F-measure indicates that MaxColl-B and MaxColl-C perform equally well or better than MaxMatch.

However, in terms of severe errors, MaxMatch, MaxColl-A, MaxColl-B, and MaxColl-C produced 64, 64, 52, and 51 errors respectively. Again, the result suggests that MaxColl-C is the best segmentation algorithm. Therefore, the best maximum collocation model for word segmentation should do only the summation of collocation strength of each word. The formula then is changed as below:

$$St = \sum_{i=1}^{n} F_{w_i}$$

$$F_{w_i} = \sum_{j=1}^{k-1} C_{s_j, s_{j+1}} \text{ such that } w_i = s_1 s_2 \ldots . s_k$$

[5] The accuracy is very high because, like all other research, it is evaluated against the researchers' analysis. By using the dictionary prepared by the researcher, the program shares the same idea of lexical units in the analysis. Thus, unlike the results shown in Table 1, problems from disagreement of lexical units are minimum.

	Precision	Recall	F-ms
Max Coll-A	75.39	85.72	80.23
	17003/22553	17003/19835	Er:64
Max Coll-B	76.56	86.07	81.03
	17071/22298	17071/19835	Er:52
Max Coll-C	76.56	86.06	81.03
	17070/22295	17070/19835	Er:51
Max Match	85.13	77.27	81.01
	15326/18003	15326/19835	Er:64

Table 3 : Results when there are unknown words

8 Conclusion

Though the maximum collocation approach does not clearly out-perform the maximum matching approach, segmentation resulted from the latter relies heavily on the words listed in the dictionary. When the maximum matching approach is used, there is a preference of compound words over simple words in the dictionary. For examples, when the compound word, ที่-มา (source), is included in the dictionary, the maximum matching will always view this string, ที่มา , as one word rather than two words, ที่ (comp.) and มา (come). But there is no such preference when the maximum collocation approach has been used. Whether this string is one or two words depends on the overall collocation strength of the input sentence. Thus, in this approach, the dictionary plays a role as defining what can be a word. How well the algorithm performs depends on the exhaustiveness of the dictionary. When there are many unknown words, the segmentation results may not be satisfying. However, we think that a dictionary is still a necessary component for defining a word. To cope with the problem of unknown words, a further study should focus on extending this approach to determine unknown words. In case an unknown word may exist, since the output from syllable segmentation is a sequence of syllables, we can use some statistical methods to determine potential words from this syllable sequence.

In addition, in this study we simplify the ratio of prob(x-y) to prob(x-ANY-y) by considering only one syllable in between x and y. It is possible to consider a wider scope of syllables in between x and y when calculating prob(x-ANY-y). In fact, if there is no restriction on the scope, prob(x-ANY-y) will equal to prob (x)*prob(y). The log ratio of prob(x-y) to prob (x)*prob(y) then is the same as the mutual information. Thus, it is possible to improve the performance of the algorithm by using other statistical method for measuring collocation strength between syllables. In our preliminary tests when applying log-likelihood, mutual information, and chi-square, for collocation strength in MaxColl-C, we found no improvement when using log-likelihood, a little improvement when using mutual information, or chi-squares. Since using mutual information is computationally least expensive here, mutual information might be the best method for capturing collocation strength at the moment. However, which statistical method is best for the approach is still an area for further study.

References

Chaicharoen, N. 2002. *Computerized Integrated Word Segmentation And Part-Of-Specch Tagging Of Thai*. Master Thesis, Faculty of Arts, Chulalongkorn University. (in Thai)

Chamyapornpong, S. 1983. *A Thai Syllable Seperation Algorithm.* Master thesis, Asian Institute of Technology.

Charoenpornsawat, P. 1998. *Feature-Based Thai Word Segmentation.* Master Thesis, Depratment of Engineering, Chulalongkorn University.

Chen, Stanley F., and Goodman, Joshua. 1998. *An Empirical Study of Smoothing techniques for Language Modeling.* TR-10-98. Harvard University.

Fleiss, J.L. 1971. Measuring Nominal Scale Agreement Among Many Raters. *Psychological Bulletin* 76. 378-382.

Krawtrakul, A. Thumkanon, C., Poovorawan, Y., and Suktarachan, M. 1997. *Automatic Thai Unknown Word Recognition.* In Proceedings of the natural language Processing Pacific Rim Symposium 1997 (NLPRS'1997).

Manning, Christopher and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing.* Cambridge: MIT Press.

Meknavin, S., Charenpornsawat, P., and kijsirikul, B. 1997. *Fetaure-based Thai Words Segmentation*, NLPRS, Incorporating SNLP-97.

Palmer, David D. 1997. *A Trainable Rule-based Algorithm for Word Segmentation*.

Poowarawan, Y. 1986. *Dictionary-based Thai Syllable Separation*, In Proceeding of the Ninth Electronics Engineering Conference.

Rietveld, Toni and Roeland van Hout. 1993. *Statistical Techniques for the Study of Language and Language Behaviour.* Berlin, new York: Mouton de gruyter.

Rijsbergen, C.J. Van. 1979. *Information Retrieval.* Butterworths: London.

Sawamiphakdi, Duangkaew. 1990. *Building a Thai Grammar Analyzer Software under the UNIX System.* Bangkok: Thammasart University Press. (in Thai).

Sornlertlamvanich, V. 1993. W*ord Segmentation for Thai in a Machine Translation System* NECTEC. (in Thai).

Thairatananond, Y. 1981. *Towards the Design of a Thai Text Syllable Analyzer.* Master thesis, Asian Institute of Technology.

Theeramunkong, T., Usanavasin, S., Machomsomboon, T., and Opasanont, B. 2000. *Thai Word Segmentation without a Dictionary by Using Decision Trees.* The fourth Symposium on Natural Language Processing 2000.