Thoughts on Word and Sentence Segmentation in Thai

Wirote Aroonmanakun

Department of Linguistics; Center for Research in Speech and Language Processing
Faculty of Arts, Chulalongkorn University,
Phayathai Rd., Pathumwan 10330, Bangkok, Thailand
Tel. +66-2-218-4696, Fax.: +66-2-218-4695
e-mail : awirote@chula.ac.th, awirote@gmail.com

Abstract

This paper discusses problems of word and sentence segmentation in Thai. Disagreements on word segmentation are caused mostly from compound words. To set a standard resource and tool of word segmentation, we suggest that only simple words and true compound words should be segmented in the process of word segmentation. Other compounds can be grouped later by the same means as multiword identification in other languages. Sentence segmentation is also difficult because the boundary of sentence in Thai is fuzzy. We suggest that a discourse should be seen as a combination of clauses rather than sentences. Some discourse clues then can be used to segment these discourse units. The result from sentence segmentation module could be a sequence of segments composed of clauses, which then can be constructed into the discourse structure.

1 Introduction

Segmenting words and sentences is considered a basic task of Thai language processing. But because of the absence of explicit word/sentence markers and unclear definitions of Thai words and sentences, it is difficult to compare results from different segmentation systems, or set up standard tools and resources for Thai NLP. This paper aims to clarify the root of these problems and suggest some solutions.

In the first part, we discuss the basic idea of what is a word. Since the notion of word can refer to different things, defining words is not straight forward even in a language with explicit word boundary like English. Thus, it is not surprising to find segmentation of Thai words by different persons or systems to be different. We will argue that word segmentation should be done by segmenting minimal integrity units similar to orthographic words in English. The result will be suitable for applications that prefer short words, while other applications that prefer words as a lexeme can group these minimal integrity words into a multiword or a lexical phrase later.

In the second part, we will argue that Thai sentences cannot be seen in the same fashion as English sentences. Since clues for sentence segmentation are not well-defined, it might be more practical to segment texts into discourse segments, which are composed of clauses rather than sentences.

2 Thai word Segmentation

2.1 What is a word?

Although the notion ‘word’ is commonly used by everybody, it is not easy to define precisely what a word is. A word can be viewed from different aspects as phonological words, orthographic words, or lexical items (lexemes) (Trask 2004, Julian 2005). A phonological word is defined as a unit of pronunciation by certain phonological properties. For example, a word in English can be determined by stress. One English word will have only one main stress. But by using this criterion, an utterance like “*I’ll be*” will be analyzed as two words, while many people would prefer to analyze it as three words. An orthographic word, on the other hand, is determined from written markers such as spaces. Orthographic words in languages that have explicit word boundary markers, e.g. English, is easy to be determined. But orthographic words are not the same thing as a lexeme. Compound words like “*ice cream*”, “*pocket knife*” are two

orthographic words, while "*pocketbook*" is one orthographic word. But many people would prefer to treat these examples as one word rather than two words. Thus, a word can be seen in another aspect as the form of a lexeme. A lexeme is an abstract unit representing a mental object. A word in this sense equates to a lexical entry in the lexicon or dictionary. But if a word equates to a mental object or a concept, a space will not always mark a word boundary. Because one concept can be signified by different forms, e.g. a simple word (e.g. *molecule*), a complex word (e.g. *intramolecule*), a compound word (e.g. *photosynthesis*), a multiword (e.g. *sewing machine*), a phrasal unit (e.g. *bridges with pin-joined members*), or a set phrase (e.g. *night and day*), it is not always possible to determine word boundaries with concept-based criterion. It can be easily seen, when browsing through entries in a specialized dictionary, that many technical terms, which signifies a single concept, are units larger than a word, e.g. *plutonic rock, divergent plate boundary, critical discourse analysis*, etc.

2.2 Problems on marking word boundaries in Thai

Since Thai writing system does not have markers for word boundaries, there is no explicit orthographic word in Thai. Phonological words in Thai are also not applicable in text-based applications. At the first thought, the concept-based method might be suitable for segmenting words in Thai. Since there is no marking for orthographic forms, there would not be any confusion between orthographic word and a lexeme. If any form does signify a concept, it should be segmented as a word in Thai. But by determining word boundaries from a single concept criterion, we could fall into a trap of segmenting a string larger than one word. For example, the concept of 'proceeding' is written in English as one word, but in Thai, it is written as หนังสือรวมบทความทางวิชาการในการประชุมสัมมนา, which is clearly not a lexical item but a noun phrase composed of nine words:

หนังสือ (book) รวม (collect) บทความ (article) ทาง (about) วิชาการ (academic) ใน (in) การ (nom.) ประชุม (meet) สัมมนา (seminar).

Since words are a linguistic unit that are larger than a morpheme (a minimal meaningful unit in a language) but smaller that a phrase, they should have integrity in terms of forms and meaning. Chaicharoen (2002), thus, used the uninterruptablility of a word as one criterion for determining a Thai word. Its integrity in form and meaning makes it unlikely to be intervened or separated by any linguistic unit without changing its meaning, and its meaning is far from the combination of meanings from its parts. For example, for the word ตู้เย็น-'refrigerator', it is not possible to insert any words in between ตู้-'closet' and เย็น-'cold' without changing its meaning, and it does not refer to a closet that is cold. But for the word ตู้เสื้อผ้า-'clothes closet', it can be analyzed as two words because its meaning is not much different from the sum of meanings from its parts: ตู้-'closet' and เสื้อผ้า-'clothes'. In addition, it is possible to have a phrase like ตู้เสื้อผ้าสีขาว, whose structure can be ambiguous as follow:

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

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

This suggests that the words ตู้ and เสื้อผ้า are not yet tightly combined. It is acceptable to be analyzed as two words. However, integrity of form and meaning is a subjective criterion because it is based on interpretation. Some may analyze คนขายของ as composed of three words 'man-sell-thing', and then set up a rule stating that any phrase with the pattern คน+v+n is not a single word. But for the word คนล้วงกระเป๋า-'pickpocket', many would prefer to analyze it as one word rather than three words, คน-ล้วง-กระเป๋า-'man-put (his hand) in-pocket'. This problem results from the degree of compounding. For a compound that is loosely combined e.g. ตู้เสื้อผ้า-'clothes closet', it is easier for most people to agree with the analysis as two words. For a compound that is tightly combined e.g. ตู้เย็น-'refrigerator', it is quite difficult to see it as two words ตู้-'closet' and เย็น-'cold'. But for a compound that is neither loose nor tight, different person could analyze it differently. Therefore, the results of segmenting Thai words even by human can be different (Aroonmanakun 2002). And in some examples, without a context, words could be ambiguous. The string คนขับรถ in some contexts, e.g. **คนขับรถ**นั่งคอยอยู่ในรถ-'the (man who works as a) driver is waiting in the car', should be viewed as one word referring

to a driver, but in some contexts, e.g. **คนขับรถ**ผ่านแยกนี้ไม่มากนัก-'not many people drive through this intersection', it should be viewed as three words refers to any persons who are driving.

2.3 Marking Thai words as a minimal integrity unit

Since word segmentation is the pre-processing for other language processing tasks. Whether the result of word segmentation is good or not is not self-determined. Different applications may prefer different kinds of segmentation. For example, in English, information retrieval systems can use orthographic words determined by spaces for indexing. Multiword indexing method for information retrieval, either statistical or syntactical, does not show much improvement in the performance than single word indexing method (Mitra et al. 1997). On the other hand, machine translation systems would prefer to have word segmented as a lexeme so that it would match an entry in the dictionary. Orthographic words as marked by spaces are not adequate for some applications. Multiword units have to be identified to get the correct lexemes. In fact, research focused on identifying multiword units has been studied for sometimes, as evidenced from the ACL workshop on Multi-word-expressions in 2003, 2004, and 2006.

Based on these facts, we think that segmenting words in Thai does not need to have results that are always a lexeme. It could yield a unit that is similar to orthographic words in English. Then, multiword units can be identified later if needed by further applications.

Thai word segmentation programs should give us results of two kinds: simple words and true compound words. Simple words are words with one morpheme, e.g. เสื้อ-'clothes', สะพาน-'bridge', นาฬิกา-'clock', etc. True compound words are words composed of two or more morphemes, and its meaning is significantly different from the sum of meanings from its parts, as seen below:

แม่น้ำ-'river' ≠ แม่-'mother' + น้ำ-'water'
ยินดี-'glad' ≠ ยิน-'hear' + ดี-'good'
หายใจ-'breath' ≠ หาย-'lost' + ใจ-'heart'

But for a compound that its meaning is not much different from the combination of its part, it should be segmented as multiple words. For example, the following words could be segmented as two or more words.

หมอฟัน-'dentist' ≈ หมอ-'specialist'+ฟัน-'dental'
กระเป๋าเดินทาง-'luggage' ≈กระเป๋า-'bag'+เดินทาง-'travel'
เครื่องตัดหญ้า-'lawnmower' ≈ เครื่อง-'machine'+ตัด-'cut'+หญ้า-'grass'

Beside the meaning criterion, syntactic criterion may be used to verify whether that word is a true compound. A compound that can be syntactically separated should not be marked as one word. For example, the word เตรียมใจ-'prepare one's mind' (เตรียม+ใจ) should not be analyzed as one word, even its structure is similar to the words ถอนใจ-'sigh', ชอบใจ-'like'. This is because the latter part, ใจ-'mind', in เตรียมใจ could be syntactically separated and combined with other words as a constituent, such as เตรียมใจของผมไว้-'prepare my mind for'; ใจของผม-'heart-of-me' is a noun phrase that is an argument of เตรียม-'prepare'. But we cannot do the same with ถอนใจ-'sigh', or ชอบใจ-'like'.

To set the standard of word segmentation, we think that word segmentation system should segment compound words that are not tightly bound as multi-words. This approach can lessen disagreement on Thai word segmentation. By applying this minimalist approach, long compound words, which are the major sources of word segmentation disagreements, would be segmented as multiple words. Strings that are ambiguous like คนขับรถ discussed earlier would be segmented as multiple words. To determine whether this example, คนขับรถ, is one lexeme or three lexemes is not the task of word segmentation program. If we think that disambiguation in these cases would require syntactic or semantic information, it is acceptable for word segmentation module to leave the problem there.

In sum, since no initial boundary is marked in Thai, when we apply concept-based for marking word boundaries in Thai, we would fall into a trap of marking unit larger than a word. Thus, breaking words as a minimal integrity unit is a more suitable solution. It is suitable for applications that rely on surface forms like information retrieval. For other applications that rely on lexemes like machine translation, these minimal integrity words can be grouped in the similar way English orthographic words are

treated in the case of multiword units.

3 Thai sentence segmentation

3.1 What is a sentence?

Like word segmentation, detecting sentence boundary is generally assumed to be a basic task for language processing. In a language in which sentence marker is explicit like English, the markers can be ambiguous for a machine. Thus, several approaches have been proposed for detecting sentence boundary in English, such as Palmer (1994), Grefenstette and Tapanainen (1994), Walker et al. (2001), Xu et al. (2005). In Thai, the problem is even worse since sentence boundary is not explicit. Previous research on Thai sentence segmentation, e.g. Mittrapiyanuruk and Sornlertlamvanich (2000), Charoenpornsawat and Sornlertlamvanich (2001), focused on disambiguating whether a space is a sentence marker or not. However, no clear definition of a Thai sentence was provided.

To understand what a sentence is, we will start by looking at languages which explicit sentence markers like English. Generally, a sentence in English can be categorized into different types as follows:[1]

- A simple sentence is a sentence containing one main clause and no subordinate clause.
- A complex sentence is a sentence which has at least one main clause and at least one subordinate clause, e.g. *The man whom you see is my brother.*
- A compound sentence is a sentence composed of two or more coordinate clauses, e.g. *John likes hamburgers, but Mary prefers hot dogs.*
- A matrix sentence is a sentence in which a clause has been embedded as a constituent, e.g. After eating the raw fish, *the dog died. The dog* that ate the raw fish *died*.

From these definitions, we can see that a sentence is a combination of clauses. When one clause depends on another, it is called a subordinate clause. When two clauses are not dependent on one another and conjoined either with or without a conjunction, it is a compound sentence. But when a clause is embedded as a constituent in a sentence, that sentence is called a matrix sentence. In actual texts, a sentence can be a mixed type. It can be a combination of clauses called a compound-complex sentence, as seen below.

Michael, who has been working on collaborative songwriting through the internet, thinks that the medium shows great promise, but Norah is not so sure about the quality that such an endeavor can produce.[2]

3.2 Marking sentence boundary in Thai

Since Thai have no explicit sentence boundary, it might be useful to look for clues of Thai sentence boundary from a parallel corpus of English and Thai texts. We may first hypothesize that a Thai segment aligned with an English sentence is a sentence, and then verify this hypothesis by asking Thai native speakers to segment sentences in the Thai texts. If the segmentation is not different from the alignment, we can hold that the aligned Thai segments can be treated as Thai sentences, and then postulate any rules of Thai sentence segmentation from the data.

To verify this idea, a small experiment was conducted. Twelve Thai native speakers were asked to segment sentences in one page of Thai translated texts without the English source texts. In addition, to confirm that their understanding of a sentence is consistent with any Thai texts, one page of Thai source text is also segmented by the same subjects.

The results show that sentence segmentation by different persons could be different. To see the overall picture, agreements on segmentation are classified into four types based on the number of agreements: slight (1-3), fair (4-6), moderate (7-9), and substantial (10-12). The correspondence between segmentation agreements and the aligned English sentence is shown in Table 1.

The result shows that most of English sentence breaks are also aligned with the Thai sentence breaks identified by the majority of subjects. Nevertheless, there are some positions where the majority viewed as a sentence break in Thai, but they are not a sentence break in the English source text. On the other hand, there are

[1] The definitions are taken from the Summer Institute of Linguistics's glossary of linguistics terms

[2] Example from http://www.englishrules.com/writing/2005/grammatical-sentence-types.php

a few positions where an English sentence is marked, but only a fair agreement of Thai break is found on those positions.

Table 1: Segmentation on Thai translated texts[3]

	Eng break	~ Eng break	Total
Substantial	13	2	15
Moderate	4	3	7
Fair	3	10	13
Slight	0	24	24
	20	39	59

When looking at the result of segmentation on Thai translated texts and source texts, we also found similar pattern of agreement, as seen in Table 2. Though many positions are marked as the sentence break by the majority, a lot of positions are marked only by a few subjects. These slight agreements are marked by various subjects. They are not the result of segmentation from some specific subjects. This suggests that the notion of Thai sentence is fuzzy.

Table 2: Comparing segmentation

	Translated	Source
Substantial	15	15
Moderate	7	13
Fair	13	11
Slight	24	27

However, the preliminary result reveals some interesting findings. Beside the known fact that a space does not always signal a sentence boundary, the findings below are observed from this small experiment.

1. When the topic shift occurs, most subjects see it as the beginning of a new sentence.
2. If the topic does continue but the same overt noun phrase or a pronoun is used, most subjects marked it the beginning of sentence.
3. Most subjects see a clause with a zero subject as a coordination of the previous clause.
4. When some phrases, e.g. และต่อมา-'and then', ตลอดระยะเวลาดังกล่าวนี้-'throughout this period', ในสมัยนี้-'in this period', ในระยะแรก-'in the first phrase', are used as a discourse marker, substantial agreement is found on that position.
5. Coordination or subordination may be identified by the presence of a conjunction, but it is not always the case. Only a few subjects see conjunctions like เพราะ-'because', แต่-'but', จึง-'thus', as the signal of sentence beginning.
6. Embedded clauses marked by ที่-'that' or ซึ่ง-'that' are analyzed as a part of the sentence by most subjects.
7. A few subjects seem to mark boundary in front of clauses preceded by a space despite the use of any conjunction.

We can conclude that substantial agreements are found at the beginning of a discourse segment in the following cases: the topic shift occurs; the topic continues with an overt noun phrase or a pronoun. These are clues for sentence segmentation because, according to Aroonmanakun (1997), in Thai, the most focused entity used to continue the topic is a zero pronoun. When an overt noun phrase or a pronoun is used as the subject, it is likely that a new topic is introduced. The use of conjunction cannot be used as a sentence marker because the clause preceded by the conjunction can be either the beginning or the continuation of a sentence. But the use of some discourse markers can signal the beginning of a sentence.

However, while marking sentence boundary may be useful for processing English, but it is questionable for Thai. What is a sentence is an invention of the writing system. Since a sentence structure is a combination of clauses and the exact boundary of sentence in Thai is fuzzy, it might be better to use clauses rather than sentences as the basic syntactic units. After clauses are identified, relations between clauses then should be determined to create the structure of clause combination. After all, sentences in English are also the combination of clauses, which finally will be combined into a discourse structure.

But identifying clauses is not as easy as identify main verbs. Thai have a serial verb construction and an embedding clause, and the structure of a compound can be similar to the structure of a clause. Unless we have a parser that could resolve structural ambiguities during the process, a paragraph should be segmented into a sequence of segments by looking for the occurrences of a space, a discourse marker, and the discourse topic (the first noun phrase at the

[3] The number excludes the end of each paragraph, which is always a sentence boundary in the data.

beginning of the segment). The result would be a sequence of segments, which can be either a clause or a combination of clauses. Then, each segment will be parsed by a dependency parser to identify the head of each segment. After that, all segments should be combined into a discourse structure by considering the relations between the head of each segment.

4 Conclusion

In this paper, some ideas of word and sentence segmentation are discussed. A preliminary analysis of sentence segmentation is concluded from the result of a small sentence segmentation experiment. Further studies on sentence segmentation should be experimented with more subjects, and then look for clues for segmenting clauses or sentences. A prototype of sentence segmentation should also be implemented.

Acknowledgement

The author would like to thank all members in the working group of Thai word segmentation standard, especially staffs from NECTEC, NAIST, SIIT, and TCL, for the fruitful discussions on word segmentation standard, which are seeds of word segmentation idea in this paper.

References

Aroonmanakun, W. 2002. Collocation and Thai Word Segmentation. In *Proceedings of the Fifth Symposium on Natural Language Processing & The Fifth Oriental COCOSDA Workshop*, Hua Hin, Thailand, pp. 68-75.

Aroonmanakun, W. 1997. Referent Resolution for Zero Pronouns in Thai. In Abramson, Arthur (ed.) *Southest Asian Linguistic Studies in Honour of Vichin Panupong,* pp. 11-24. Bangkok: Chulalongkorn University Press.

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

Charoenpornsawat, P. and V. Sornlertlamvanich. 2001. Automatic Sentence Break Disambiguation for Thai. In *Proceedings of ICCPOL2001*, Korea, pp. 231-235.

Fagan, J. L. 1989. The effectiveness of a nonsyntactic approach to automatic phrase indexing for document retrieval. *Journal of the American Society for Information Science*, 40:115-132.

Grefenstette, G. and P. Tapanainen. 1994. What is a word, What is a sentence? Problems of Tokennization. In *the 3rd Conference on Computational Lexicography and Text Research. COMPLEX'94*, Budapest, July 7-10, 1994.

Julian, M. 2005. Words. In *Encyclopedia of Language and Linguistics*. 2nd edition. pp. 617-624. Elsevier.

Mittrapiyanuruk, P. and V. Sornlertlamvanich. 2000. The Automatic Thai Sentence Extraction. *The Fourth Symposium on Natural Language Processing (SNLP2000)*, Chiang Mai, Thailand, pp 23-28.

Mitra, M., C. Buckley, A. Singhal, and C. Cardie. 1997. An Analysis of Statistical And Syntactic Phrases. In *Proceedings of RIAO97, Computer-Assisted Information Searching on the Internet*, Montreal, Canada, pp. 200–214

Palmer, David D. 1994. *Satz - an adaptive sentence segmentation system.* M.S. Thesis and UC-Berkeley Technical Report UCB/CSD 94-846.

Trask, Larry. 2004. *What is a word?* Working Papers in Linguistics and English Language. Department of Linguistics and English Language, University of Sussex.

Vechtomova, Olga. 2005. The Role of Multi-word Units in Interactive Information Retrieval. In D.E. Losada and J.M. Fernández-Luna (eds.): *ECIR 2005, LNCS 3408*, pp. 403 – 420. Springer-Verlag Berlin Heidelberg.

Walker, Daniel J., David E. Clements, Maki Darwin and Jan W. 2001. Amtrup: Sentence Boundary Detection: A Comparison of Paradigms for Improving MT Quality (R). *Proceeding of MT Summit VIII,* pp. 18-22.

Xu, Jia, Richard Zens, and Hermann Ney. 2005. Sentence segmentation using IBM word alignment model 1. In *The 10th EAMT conference Practical applications of machine translation*, Budapest, pp. 280-287.