

Thai Word Segmentation a Lexical Semantic Approach

Krisda Khankasikam

Department of Computer Engineering
King Mongkut's University of Technology Thonburi
Bangkok, Thailand, 10140
KrisdaK@gmail.com

Nuttanart Muansuwan

Department of Computer Engineering
King Mongkut's University of Technology Thonburi
Bangkok, Thailand, 10140
Nuttanart@cpe.kmutt.ac.th

Abstract

In Thai language, the word boundary is not explicitly clear, therefore, word segmentation is needed to determine word boundary in Thai sentences. Many applications of Thai Language Processing require the word segmentation. Several approaches of Thai word segmentation such as maximal matching, longest matching and n-gram model do not take semantics into consideration. This paper presents a Thai word segmentation system using semantic corpus which is composed of four steps: generating all possible candidates, proper noun consideration, semantic tagging and semantic checking. The first three steps are conducted using a dictionary. Semantic checking is carried out on the basis of corpus-based approach. Finally, we assign the semantic scores to segmented words and select the ones that contain maximum semantic scores. In order to assign semantic scores, we use a Thai proper noun database and the semantic corpus derived from ORCHID corpus. This approach is more reliable than other approaches that do not take the meaning into consideration and performs the level of accuracy at 96-99% depending on the characteristic of input and the dictionary used in the segmentation.

1 Introduction

It has long been known that word segmentation is an essential problem in natural language processing for certain Asian languages such as Chinese, Japanese, and Thai. Unlike English, Thai does not have word boundary, i.e. there is no space between words. Other than that, compounds are the serious problem for segmenting words in Thai. For example แม่ (mother) and น้ำ (water), together they form a compound แม่น้ำ which mean "river".

2 Previous Approaches

This section briefly reviews previous research on Thai word segmentation.

2.1 Rule-based Approach

The rule-based approach (Chamyapornpong, 1983) is the method used in the early development of word segmentation system. For this method, it check rule of language such as space and beginning of new paragraph to specify the word boundary. The rule-based of character specifies a probability of word segmentation in the position of that character. This method divides character into 5 groups.

1. Non-spacing character such as ั, ิ, ุ, ็, ๋ and ็.
2. Leader character such as ิ, ุ, ็, ๋ and ็.
3. Follower character such as ะ, ็ and ็.
4. The mark placed over the final consonant of a word in Thai language to indicate that it is mute character. It is ั.
5. Remain character.

The disadvantage of rule-based approach is that it does not yield the high precision and requires hand-crafted rules resource. However, this method is suitable for word wrapping.

2.2 Thai Character Cluster

In Thai, some close characters tend to be an inseparable unit, called Thai character cluster (TCCs). Unlike word segmentation that is a very difficult task, segmenting a text into TCCs is easily recognized by applying a set of rules. The method to segment a text into TCCs was proposed in (Theeramunkong, Tanhermhong, Chinnan and Sornlertlamvanich, 2000). This method needs no dictionary and can always correctly segment a text at every word boundary. As the first step of word segmentation approach, a set of rules is applied to group of close characters in a text together to form TCCs. The accuracy of this process is 100% in the sense that there is no possibility that these units are divided to two or more units, which are substrings in two or more different words. This process can be implemented without a dictionary, but uses a set of simple linguistic rules based on the types of characters. Table 1 displays the types of Thai characters. As an example rule, a front vowel and its next consonant must exist in the same unit. Table 2 shows a fragment of a text segmented into

TCCs by the proposed method and its correct word segmentation. Here, a character ‘|’ indicates a segmentation point. The corpus where characters are grouped into TCCs is called a TCC corpus.

Type of Thai Characters	Member
Consonant	ก ข ฃ ค ฅ ฉ ง จ ฉ ช ซ ฌ ญ ฎ ฏ ฐ ฑ ฒ ณ ด ต ถ ท ธ น บ ป ผ ฝ พ ฟ ภ ม ย ร ฤ ล ฬ ว ศ ษ ส ห พ อ ฮ
Upper vowel	อี อี อึ อึ อี อี อี อี
Lower vowel	อุ อุ
Front vowel	แ อ ไ อ
Rear vowel	า ำ ำ ะ ะ ำ ำ

Table 1: Types of Thai characters

TCCs	การ เก็บ ภาษี ประเทศ ไทย และ ประเทศ
Correct	การ เก็บภาษีประเทศไทยและประเทศ

Table 2: An example of TCCs vs. correct segmentation

2.3 Longest Matching

This method is used in the early work of Thai word segmentation (Poonwarawan, 1986). This method scans an input sentence from left to right, and selects the longest match with a dictionary entry at each point. In case that the selected match cannot lead the algorithm to find the rest of the words in the sentence, the algorithm will backtrack to find the next longest one and continue finding the rest and so on. It is obvious that this algorithm will fail to find the correct the segmentation in many cases because of its greedy characteristic. For example “ไปหาหมเหสี” (go to see the queen) will be incorrectly segmented as ไป (go) หา (carry) เห (deviate) สี (color), while the correct one that cannot be found by the algorithm is ไป (go) หา (see) หมเหสี (Queen).

2.4 Maximum Matching

The maximum matching (Sornlertlamvanich, 1993) method was proposed to solve the problem of the longest matching method which is described above. This method first generates all possible segmentations for a sentence and then select the one that contain the fewest words, which can be done efficiently by using dynamic programming technique. Because the method actually finds real maximum matching instead of using local greedy heuristics to guess, it always outperforms the

longest matching method. Nevertheless, when the alternatives have the same number of words, the algorithm cannot determine the best candidate and some other heuristics have to be applied. The heuristic often used is again the greedy one to prefer the longest matching at each point. For the example, ตา (expose) ลม (wind) is preferred to ตา (eye) กลม (round).

2.5 Feature-based Approach

The feature-based approaches (Meknavin, Charenpornsawat and Kijisirikul, 1997) have been developed for solving ambiguity in word segmentation. In this approach, the system generates all possible segmentation for a string, which has segmentation ambiguity. The problem is that how to select the best segmentation from the set of candidates. At this point, this research applies and compares two learning techniques, called RIPPER and Winnow. RIPPER algorithm is a propositional learning algorithm that constructs a set of rules while Winnow algorithm is a weighted-majority learning algorithm that learns a network, where each node in the network is called a specialist. Each specialist looks at a particular value of an attribute of the target concept, and will vote for a value of the target concept based on its specialty; i.e., based on a value of the attribute it examines. The global algorithm combines the votes from all specialists and makes decision. This approach is based on dictionary.

2.6 Statistic-based approach

This approach is based on the word context consideration. The information of the neighbor words is used to decide on the word boundary (Aroonmanakul, 2002) and (Krawtrakul, Thumkanon, Poovorawan and Suktarachan, 1997). There are two points to be resolved for this approach, which are the context width and the applied statistical method. The part of context width is concerned the wider the more complex. The part of statistical method is concerned, the hidden first-order Markov has always been applied. However, this method greatly depends on the corpus for its training. In case one method is applicable for the political corpus, it cannot be applied to literal ones, for example. In addition, there are some words of high probability but of syntactical functions only.

The methods mentioned above for word segmentation in Thai do not take the semantics of Thai language into consideration which can potentially improve the word segmentation.

3 Lexical Semantics: Word Hierarchy

Word hierarchy is classifying words by meaning hierarchy. Each word in the sentence such as noun, verb and adjective are divided to group of meaning that is called “A Kind Of” (AKO). A kind of (AKO) is information that uses to consider word meaning in Thai language to form of group. AKO is beneficial because it can be used to analyze the sentence’s meaning and reduce ambiguity in the sentence. For example “ตากลับบ้าน”, there is an ambiguity of the noun “ตา” between “grandfather” and “eye”. To fix the problem, the computer must use nearby words to identify meaning of “ตา” which is a word “กลับ”. The word “กลับ” is a verb whose subject of “กลับ” should be a living thing. Therefore, the word “ตา” should have the meaning “grandfather”. In our research, we have 74 sub categories of AKO number. The example of the separation of meaning hierarchy is shown below (Tantiswetratch, Yamket, Choksuwanich, Chanchareon, Boriboon and Tannin, 1993).

1 Concrete:

11 Subject:

111 Person: such as friends and parent.

112 Organization: such as committee.

12 Concrete place:

121 Region: such as province.

122 Nature:

1221 Topography: such as mountain.

2 Abstract:

21 Abstract matter:

211 Activity:

2111 Action: such as walk.

212 Phenomenon:

2121 Event: such as destiny.

4 Building the Semantic Corpus

The semantic corpus is a corpus which contained semantic information to identify the meaning of each word in a corpus. The meaning of each word is in form of AKO (A Kind Of) number. The words that have the same AKO number are considered to have the same meaning.

In our research, we apply Thai ORCHID corpus (Charoenporn, Sornlertlamvanich and Isaraha, 1997) which can be viewed as a syntactic-semantic corpus. To augment ORCHID corpus to be a finally semantic corpus, AKO information is added. This can be done by using an electronic dictionary which contains the word categories and subcategories (AKO) in the format compatible with ORCHID corpus. We will then assign AKO from dictionary to ORCHID corpus too. Our semantic corpus contained 431,338 words and

137,244 sentences. Figure 1 shows an example of a semantic corpus derived from ORCHID.

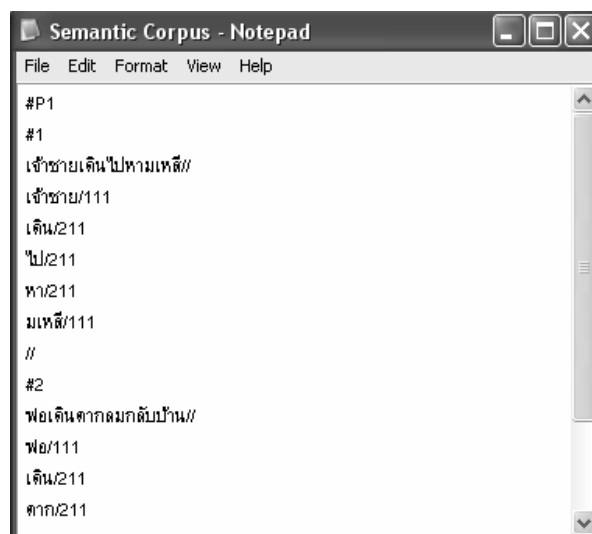


Figure1: Semantic Corpus

5 Thai Word Segmentation using Semantic Corpus

Thai word segmentation using semantic corpus is composed of four steps: generating all possible candidates, proper noun consideration, semantic tagging and semantic checking.

5.1 Generating all Possible Candidates

At first we apply forward maximal matching algorithm and backward maximal matching algorithm using dictionary to generate all possible segmentation. Suppose the input is “หม่อมเจ้าชาตรีเฉลิมฉลองวันเกิด”, this step will give 16 possible segmentations as shown below

1. หม่อมเจ้า- ชาตรี- เฉลิมฉลอง- วันเกิด
2. หม่อมเจ้า- ชาตรี- เฉลิมฉลอง- วัน- เกิด
3. หม่อมเจ้า- ชาตรี- เฉลิม- ฉลอง- วันเกิด
4. หม่อมเจ้า- ชาตรี- เฉลิม- ฉลอง- วัน- เกิด
5. หม่อมเจ้า- ชา- ตรี- เฉลิมฉลอง- วันเกิด
6. หม่อมเจ้า- ชา- ตรี- เฉลิมฉลอง- วัน- เกิด
7. หม่อมเจ้า- ชา- ตรี- เฉลิม- ฉลอง- วันเกิด
8. หม่อมเจ้า- ชา- ตรี- เฉลิม- ฉลอง- วัน- เกิด
9. หม่อม- เจ้า- ชาตรี- เฉลิมฉลอง- วันเกิด
10. หม่อม- เจ้า- ชาตรี- เฉลิมฉลอง- วัน- เกิด
11. หม่อม- เจ้า- ชาตรี- เฉลิม- ฉลอง- วันเกิด
12. หม่อม- เจ้า- ชาตรี- เฉลิม- ฉลอง- วัน- เกิด
13. หม่อม- เจ้า- ชา- ตรี- เฉลิมฉลอง- วันเกิด
14. หม่อม- เจ้า- ชา- ตรี- เฉลิมฉลอง- วัน- เกิด
15. หม่อม- เจ้า- ชา- ตรี- เฉลิม- ฉลอง- วันเกิด
16. หม่อม- เจ้า- ชา- ตรี- เฉลิม- ฉลอง- วัน- เกิด

5.2 Proper Noun Consideration

In this step, proper noun consideration is considered the result from first step by comparing the word unit with proper noun database. If a resulting word from the first step matches a word in proper noun database, it will assign the low priority of proper noun. If there are two next word match with proper noun database it will generate the new candidate by merging two next words and assign the normal priority of proper noun. If there are more than two of next words match with the word in proper noun database it will generate the new candidate from the first step by merging the group of next word and assign the high priority of proper noun.

For example, “ชาตรีเฉลิม” is a name in proper noun database. Proper noun consideration will generate new candidate segmentation by merging ชาตรี-เฉลิม to ชาตรีเฉลิม, ชา – ตรี – เฉลิม to ชาตรีเฉลิม and assign the normal, high priority of proper noun to ชาตรีเฉลิม. The proper noun consideration step gives the following results.

1. หม่อมเจ้า – ชาตรี – เฉลิมฉลอง – วันเกิด
2. หม่อมเจ้า – ชาตรี – เฉลิมฉลอง – วัน – เกิด
3. หม่อมเจ้า – ชาตรี – เฉลิม – ฉลอง – วันเกิด
4. หม่อมเจ้า – ชาตรีเฉลิม (normal priority) – ฉลอง – วันเกิด
5. หม่อมเจ้า – ชาตรี – เฉลิม – ฉลอง – วัน – เกิด
6. หม่อมเจ้า – ชาตรีเฉลิม (normal priority) – ฉลอง – วัน – เกิด
7. หม่อมเจ้า – ชา – ตรี – เฉลิมฉลอง – วันเกิด
8. หม่อมเจ้า – ชา – ตรี – เฉลิมฉลอง – วัน – เกิด
9. หม่อมเจ้า – ชา – ตรี – เฉลิม – ฉลอง – วันเกิด
10. หม่อมเจ้า – ชาตรีเฉลิม (high priority) – ฉลอง – วันเกิด
11. หม่อมเจ้า – ชา – ตรี – เฉลิม – ฉลอง – วัน – เกิด
12. หม่อมเจ้า – ชาตรีเฉลิม (high priority) – ฉลอง – วัน – เกิด
13. หม่อม – เจ้า – ชาตรี – เฉลิมฉลอง – วันเกิด
14. หม่อม – เจ้า – ชาตรี – เฉลิมฉลอง – วัน – เกิด
15. หม่อม – เจ้า – ชาตรี – เฉลิม – ฉลอง – วันเกิด
16. หม่อม – เจ้า – ชาตรีเฉลิม (normal priority) – ฉลอง – วันเกิด
17. หม่อม – เจ้า – ชาตรี – เฉลิม – ฉลอง – วัน – เกิด
18. หม่อม – เจ้า – ชาตรีเฉลิม (normal priority) – ฉลอง – วัน – เกิด
19. หม่อม – เจ้า – ชา – ตรี – เฉลิมฉลอง – วันเกิด
20. หม่อม – เจ้า – ชา – ตรี – เฉลิมฉลอง – วัน – เกิด
21. หม่อม – เจ้า – ชา – ตรี – เฉลิม – ฉลอง – วันเกิด
22. หม่อม – เจ้า – ชาตรีเฉลิม (high priority) – ฉลอง – วันเกิด

23. หม่อม – เจ้า – ชา – ตรี – เฉลิม – ฉลอง – วัน – เกิด
24. หม่อม – เจ้า – ชาตรีเฉลิม (high priority) – ฉลอง – วัน – เกิด

5.3 Semantic Tagging

In this step, each segmented word is tagged with an AKO number. From the above example “หม่อมเจ้าชาตรีเฉลิมฉลองวันเกิด”, the semantic tagging step gives the following results.

1. หม่อมเจ้า_2367 / king’s grandson / Title
ชาตรี_111/ warrior / Person
เฉลิมฉลอง_2111 / celebrate / Action
วันเกิด_232 / birthday / Time
2. หม่อมเจ้า_2367 / king’s grandson / Title
ชาตรี_111/ warrior / Person
เฉลิมฉลอง_2111 / celebrate / Action
วัน_232 / day / Time
เกิด_2121 / occur / Event
3. หม่อมเจ้า_2367 / king’s grandson / Title
ชาตรี_111 / warrior / Person
เฉลิม_2111 / glorify / Action
ฉลอง_2111 / celebrate / Action
วันเกิด_232 / birthday / Time
4. หม่อมเจ้า_2367 / king’s grandson / Title
ชาตรีเฉลิม (normal priority) _111_666 / Person
proper name Chatreechalm
ฉลอง_2111 / celebrate / Action
วันเกิด_232 / birthday / Time
5. หม่อมเจ้า_2367 / king’s grandson / Title
ชาตรี_111 / warrior / Person
เฉลิม_2111 / glorify / Action
ฉลอง_2111 / celebrate / Action
วัน_232 / day / Time
เกิด_2121 / occur / Event
6. หม่อมเจ้า_2367 / king’s grandson
ชาตรีเฉลิม (normal priority) _111_666 / Person
proper name Chatreechalm
ฉลอง_2111 / celebrate / Action
วัน_232 / day / Time
เกิด_2121 / occur / Event

7. หม่อมเจ้า_2367 / king's grandson / Title
ชา_1322 / tea / Finished product
ตรี_2365 / three / Number
เฉลิมฉลอง_2111 / celebrate / Action
วันเกิด_232 / birthday / Time
8. หม่อมเจ้า_2367 / king's grandson / Title
ชา_1322 / tea / Finished product
ตรี_2365 / three / Number
เฉลิมฉลอง_2111 / celebrate / Action
วัน_232 / day / Time
เกิด_2121 / occur / Event
9. หม่อมเจ้า_2367 / king's grandson / Title
ชา_1322 / tea / Finished product
ตรี_2365 / three / Number
เฉลิม_2111 / glorify / Action
ฉลอง_2111 / celebrate / Action
วันเกิด_232 / birthday / Time
10. หม่อมเจ้า_2367 / king's grandson / Title
ชาติเฉลิม (high priority)_111_666 / Person
proper name Chatreechalerm
ฉลอง_2111 / celebrate / Action
วันเกิด_232 / birthday / Time
11. หม่อมเจ้า_2367 / king's grandson / Title
ชา_1322 / tea / Finished product
ตรี_2365 / three / Number
เฉลิม_2111 / glorify / Action
ฉลอง_2111 / celebrate / Action
วัน_232 / day / Time
เกิด_2121 / occur / Event
12. หม่อมเจ้า_2367 / king's grandson / Title
ชาติเฉลิม (high priority)_111_666 / Person
proper name Chatreechalerm
ฉลอง_2111 / celebrate / Action
วัน_232 / day / Time
เกิด_2121 / occur / Event
13. หม่อม_2367 / king's grandson's wife / Title
เจ้า_2364 / royalty / Type
ชาติตรี_111 / warrior / Person
เฉลิมฉลอง_2111 / celebrate / Action
วันเกิด_232 / birthday / Time
14. หม่อม_2367 / king's grandson's wife / Title
เจ้า_2364 / royalty / Type
ชาติตรี_111 / warrior / Person
เฉลิมฉลอง_2111 / celebrate / Action
วัน_232 / day / Time
เกิด_2121 / occur / Event
15. หม่อม_2367 / king's grandson's wife / Title
เจ้า_2364 / royalty / Type
ชาติตรี_111 / warrior / Person
เฉลิม_2111 / glorify / Action
ฉลอง_2111 / celebrate / Action
วันเกิด_232 / birthday / Time
16. หม่อม_2367 / king's grandson's wife / Title
เจ้า_2364 / royalty / Type
ชาติเฉลิม (normal priority)_111_666 / Person
proper name Chatreechalerm
ฉลอง_2111 / celebrate / Action
วันเกิด_232 / birthday / Time
17. หม่อม_2367 / king's grandson's wife / Title
เจ้า_2364 / royalty / Type
ชาติตรี_111 / warrior / Person
เฉลิม_2111 / glorify / Action
ฉลอง_2111 / celebrate / Action
วัน_232 / day / Time
เกิด_2121 / occur / Event
18. หม่อม_2367 / king's grandson's wife / Title
เจ้า_2364 / royalty / Type
ชาติเฉลิม (normal priority)_111_666 / Person
proper name Chatreechalerm
ฉลอง_2111 / celebrate / Action
วัน_232 / day / Time
เกิด_2121 / occur / Event
19. หม่อม_2367 / king's grandson's wife / Title
เจ้า_2364 / royalty / Type
ชา_1322 / tea / Finished product
ตรี_2365 / three / Number
เฉลิมฉลอง_2111 / celebrate / Action
วันเกิด_232 / birthday / Time

20. หม่อม_2367 / king's grandson's wife / Title
 เจ้า_2364 / royalty / Type
 ชา_1322 / tea / Finished product
 ครี_2365 / three / Number
 เฉลิมฉลอง_2111 / celebrate / Action
 วัน_232 / day / Time
 เกิด_2121 / occur / Event
21. หม่อม_2367 / king's grandson's wife / Title
 เจ้า_2364 / royalty / Type
 ชา_1322 / tea / Finished product
 ครี_2365 / three / Number
 เฉลิม_2111 / glorify / Action
 ฉลอง_2111 / celebrate / Action
 วันเกิด_232 / birthday / Time
22. หม่อม_2367 / king's grandson's wife / Title
 เจ้า_2364 / royalty / Type
 ชาตรีเฉลิม (high priority) _111_666 / Person
 proper name Chatreechalerm
 ฉลอง_2111 / celebrate / Action
 วันเกิด_232 / birthday / Event
23. หม่อม_2367 / king's grandson's wife / Title
 เจ้า_2364 / royalty / Type
 ชา_1322 / tea / Finished product
 ครี_2365 / three / Number
 เฉลิม_2111 / glorify / Action
 ฉลอง_2111 / celebrate / Action
 วัน_232 / day / Time
 เกิด_2121 / occur / Event
24. หม่อม_2367 / king's grandson's wife / Title
 เจ้า_2364 / royalty / Type
 ชาตรีเฉลิม (high priority) _111_666 / Person
 proper name Chatreechalerm
 ฉลอง_2111 / celebrate / Action
 วัน_232 / day / Time
 เกิด_2121 / occur / Event

Among these results, if the semantic patterns are the same, we will select the ones with higher priority of proper noun. The results will be reduced down to 20 as shown below:

1. หม่อมเจ้า- ชาตรี- เฉลิมฉลอง- วันเกิด
(2367 - 111 - 2111 - 232)
2. หม่อมเจ้า- ชาตรี- เฉลิมฉลอง- วัน- เกิด
(2367 - 111 - 2111 - 232 - 2121)
3. หม่อมเจ้า- ชาตรี- เฉลิม- ฉลอง- วันเกิด
(2367 - 111 - 2111 - 2111 - 232)
4. หม่อมเจ้า- ชาตรี- เฉลิม- ฉลอง- วัน- เกิด
(2367 - 111 - 2111 - 2111 - 232 - 2121)
5. หม่อมเจ้า- ชา- ตรี- เฉลิมฉลอง- วันเกิด
(2367 - 1322 - 2365 - 2111 - 232)
6. หม่อมเจ้า- ชา- ตรี- เฉลิมฉลอง- วัน- เกิด
(2367 - 1322 - 2365 - 2111 - 232 - 2121)
7. หม่อมเจ้า- ชา- ตรี- เฉลิม- ฉลอง- วันเกิด
(2367 - 1322 - 2365 - 2111 - 2111 - 232)
8. หม่อมเจ้า- ชาตรีเฉลิม (high priority)- ฉลอง- วันเกิด
(2367 - 111/666 - 2111 - 232)
9. หม่อมเจ้า- ชา- ตรี- เฉลิม- ฉลอง- วัน- เกิด
(2367 - 2121 - 2365 - 2111 - 2111 - 232 - 2121)
10. หม่อมเจ้า- ชาตรีเฉลิม (high priority)- ฉลอง- วัน- เกิด
(2367 - 111/666 - 2111 - 232 - 2121)
11. หม่อม- เจ้า- ชาตรี- เฉลิมฉลอง- วันเกิด
(2367 - 2364 - 111 - 2111 - 232)
12. หม่อม- เจ้า- ชาตรี- เฉลิมฉลอง- วัน- เกิด
(2367 - 2364 - 111 - 2111 - 232 - 2121)
13. หม่อม- เจ้า- ชาตรี- เฉลิม- ฉลอง- วันเกิด
(2367 - 2364 - 111 - 2111 - 2111 - 232)
14. หม่อม- เจ้า- ชาตรี- เฉลิม- ฉลอง- วัน- เกิด
(2367 - 2364 - 111 - 2111 - 2111 - 232 - 2121)
15. หม่อม- เจ้า- ชา- ตรี- เฉลิมฉลอง- วันเกิด
(2367 - 2364 - 2121 - 2365 - 2111 - 232)
16. หม่อม- เจ้า- ชา- ตรี- เฉลิมฉลอง- วัน- เกิด
(2367 - 2364 - 2121 - 2365 - 2111 - 232 - 2121)
17. หม่อม- เจ้า- ชา- ตรี- เฉลิม- ฉลอง- วันเกิด
(2367 - 2364 - 2121 - 2365 - 2111 - 2111 - 232)
18. หม่อม- เจ้า- ชาตรีเฉลิม (high priority)- ฉลอง- วันเกิด
(2367 - 2364 - 111/666 - 2111 - 2121)
19. หม่อม- เจ้า- ชา- ตรี- เฉลิม- ฉลอง- วัน- เกิด
(2367 - 2364 - 2121 - 2365 - 2111 - 2111 - 232 - 2121)
20. หม่อม- เจ้า- ชาตรีเฉลิม (high priority)- ฉลอง- วัน- เกิด
(2367 - 2364 - 111/666 - 2111 - 232 - 2121)

5.4 Semantic Checking

In the last step, the semantic scores are calculated by counting the occurrence frequency of semantic pattern in semantic corpus. The results that contain the maximum of semantic scores and highest priority of proper noun will be selected. For the above example, the semantic scores and priority of proper noun ordering from maximum to minimum are as shown below.

Candidates Number	Semantic Scores	Priority of Proper Noun
8	1,273	1 High
1	1,146	-
10	415	1 High
2	314	-
18	26	1 High
12	5	-
20	0	1 High
3,4,5,6,7,9,11,13, 14,15,16,17, 19	0	-

Table3: The results of Thai word segmentation using semantic corpus

In this case, the candidate number 8 “หอมเจ้า – ชาตรีเฉลิม (high priority) – ฉลอง – วันเกิด” is selected because it contains the maximum semantic scores and highest priority of proper noun. The semantic scores are 12,735. It means that, in the semantic corpus, the total sentences which have the same semantic pattern of 2367 – 111/666 – 2111 – 232 are 12,735 sentences.

6 Experiment Results

We test our algorithm with 3,657,845 words and the resulted segmented words will be evaluated by linguists for the accuracy. So far, the resulted segmented words are accurate both in terms of segmentation and the right meaning assigned as in the case of “ชาตรีเฉลิม” which we could segment correctly as King’s grandson named Chatreechalerm celebrated his birthday, i.e. we solved the ambiguity of the string “ชาตรีเฉลิม”. The average percentage of correctness is 97.34%. The result of Thai word segmentation using the lexical semantic approach is shown below.

Input	Number of Word	Correct Word	Correctness Percentage
Proceeding of conference	1,503,980	1,454,048	96.68%
Fairy tale	1,138,212	1,108,732	97.41%
Text book of computer	892,368	875,859	98.15%
Thai Encyclopedia for children	85,133	84,060	98.74%
Newspaper	25,798	25,571	99.12%
Tradition article	12,354	12,308	99.63%
Total	3,657,845	3,560,578	-

Table 4: The result of Thai word segmentation using lexical semantic approach

We found that most of incorrect segmented words are proper nouns. For example, for “หลวงตามหาบัว” (the reverend grandfather monk “Maha Bua”) our algorithm can segment into หลวง (title of government officer in the old time), ตามหา (look for someone), บัว (lotus) but the correct segmentation should be หลวงตา (the reverend grandfather monk) มหาบัว (Maha Bua—proper name). This is because our proper noun database does not contain the word “มหาบัว”. The percentages of proper noun from incorrect result are shown below.

Input	Incorrect Word	Proper Noun	Percentage
Proceeding of conference	49,932	36,291	72.68%
Fairy tale	29,480	18,004	61.07%
Text book of computer	16,509	12,142	73.55%
Thai Encyclopedia for children	1,073	738	68.78%
Newspaper	227	186	81.93%
Nakhonsawan Tradition	46	38	99.63%

Table5: The percentage of proper noun from incorrect result

7 Conclusion

This paper describes Thai word segmentation method that takes the semantics of words (lexical semantics) in sentences into consideration which can reduce the ambiguity better than the approaches that do not consider the meaning.

References

- Aroonmanakul. W. 2002. *Collocation and Thai Word Segmentation*. In proceeding of SNLP-Oriental COCOSDA.
- Aroonmanakul. W. 2002. *Corpus Linguistics*. Chulalongkorn University, Bangkok.
- Chamyapornpong. S. 1983. *A Thai Syllable Separation Algorithm*. Master thesis, Asian Institute of Technology, Bangkok.
- Charoenporn. T., Sornlertlamvanich. V. and Isaraha. H. 1997. *Building A Large Thai Text Corpus-Part-Of-Speech Tagged Corpus: ORCHID*. NECTEC, Bangkok.
- 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.
- Meknavin. S., Charenpornsawat. P. and Kijisirikul. B. 1997. *Feature-based Thai Words Segmentation*. NLPRS, Incorporating SNLP.
- Poonwarawan. Y. 1986. *Dictionary-based Thai Syllable Separation*. In proceeding of the 9th Electrical Engineering Conference.
- Promchan. P. 2002. *Performance Comparison of Thai Word Separation Algorithms*. Chulalongkorn University, Bangkok.
- Sornlertlamvanich. V. 1993. *Word Segmentation for Thai in a Machine Translation System*. NECTEC, Bangkok.
- Sornlertlamvanich. V., Potipiti. T., Wutiwiwatchai. C. and Mittrapiyanurak. P. 2001. *The State of Art in Thai Language Processing*. NECTEC, Bangkok.
- Sornlertlamvanich. V. and Charenpornsawat. P. 2002. *Automatic Sentence Break Disambiguation for Thai*. NECTEC, Bangkok.
- Tantiswetratch. N., Yamket. K., Choksuwanich. T., Chanchareon. K., Boriboon. M. and Tannin. N. 1993. *The project to develop the dictionary for machine translation*. NECTEC, Bangkok.
- Theeramunkong. T., Usanavasin. S., Machomsomboon. T. and Opanant. B. 2002. *Thai word Segmentation without a Dictionary by using Decision Trees*. The fourth Symposium on Natural Language Processing.
- Theeramunkong. T., Sornlertlamvanich. V., Tanhermhong. T. and Chinnan. W. 2000. *Character-Cluster Based Thai Information Retrieval*. Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages, pages 75-80, Hong Kong.