# THAI EDU SEGMENTATION USING CLUE MARKERS AND SYNTACTIC INFORMATION FROM SHALLOW PARSER

**AUTHAPON KONGWAN[1], SITI SAKIRA BINTI KAMARUDDIN[2],**

**FARZANA BINTI KABIR AHMAD[3]**

[123]School of Computing, College of Arts and Sciences, Universiti Utara Malaysia, Malaysia
[1]Computer Engineering Department, Faculty of Engineering, Rajamangala University of Technology Srivijaya, Thailand
E-mail: [1]authapon.k@rmutsv.ac.th, [2]sakira@uum.edu.my, [3]farzana58@uum.edu.my

## ABSTRACT

Text is one of the useful knowledge sources of a human. Each element in a text has to be analyzed to identify the piece of information and knowledge. EDU is important for NLP applications that need a smaller unit to process rather than a sentence such as text summarization, information extraction, and question answering. Therefore, EDU can be more appropriated than a sentence to extract knowledge and information from the text. This paper presents a pipeline of the process for Thai EDU segmentation from word segmentation to EDU segmentation. The shallow parser is applied to chunk a non-recursive phrase in a text to reveal partial syntactic information for EDU segmentation. And then, syntactic information is utilized to identify and reconstruct the EDU segmentation in text. From the experiment, the results show that the precision, recall, and F1 score are 0.88865, 0.91577, and 0.90200 respectively.

**Keywords:** *Word Segmentation, EDU Segmentation, Conditional Random Field, Shallow Parser, Natural Language Processing*

## 1. INTRODUCTION

Text is a very interesting source of human knowledge. The research areas in Information Extraction (IE), Knowledge Extraction (KE), and Question Answering System (QAS) need text processing to identify the interesting information or construct a knowledge-based. Natural Language Processing (NLP) tasks are the essential task to achieve that purpose, especially in Thai text [1, 2].

Word segmentation and sentence boundary are an important process for Thai NLP tasks. There are some characteristics in Thai text that makes it a complicated text to be analyzed. In Thai text, there are no space and punctuation to indicate the boundary of word and sentence. Moreover, the Thai word can have more than one meaning depending on its function or context. For example "กำลัง" can mean "power" if its function is a noun and can be "-ing" if its function is an auxiliary verb. Furthermore, this word can be segmented into 2 words "กำ (grasp)" and "ลัง (box)". Therefore, the precision of word segmentation is a crucial factor for Thai NLP tasks. Thai sentence segmentation is also a complicated process. The sentences can be written continuously in a paragraph without any punctuation or marker to separate each sentence. In some writing styles, space can be used to separate each sentence in a paragraph. However, space can be used in some part of a sentence such as 1) before and after the number "นกกระจอกเทศจะหนักประมาณ 160 กิโลกรัม (Ostrich's weight is 160 kilograms approximately)", 2) before conjunctions "และ (and)" "หรือ (or)", and 3) before the repeater sign "ๆ".

In some applications such as text summarization [3, 4], a sentence can be too large and a smaller unit is more needed to process. A minimal discourse unit from a discourse tree structure, which is called an elementary discourse unit (EDU) [5, 6, 7], is more appropriate data to process. Some other characteristics in Thai text, such as the zero anaphora, can also affect the EDU segmentation process. Zero anaphora is the omission of the subject in Thai sentences. Furthermore, the use of zero anaphora without any punctuation in Thai text can cause unclear boundaries of sentence and EDU. Moreover, there is an embedded clause, which is a relative clause, in noun phrases that make the boundary of EDU more complicated.

Syntactic information can be useful information to identify the boundary of EDU in a paragraph. However, full parsing is difficult without sentence breaking before. A non-recursive phrase in a paragraph could be easier to identify by a shallow parser [8]. Hence, this paper aims to use the syntactic information from the shallow parser to identify the boundary of Thai EDU in a paragraph. In this work, word segmentation is developed to identify the boundary of words in a paragraph and shallow parser is developed to provide the syntactic information for the EDU segmentation process. Clue word is utilized to identify the boundary of EDU as an explicit marker in the first step. Then, a syntactic pattern is applied to segment the EDU in a paragraph and the EDU reconstruction process applies rule-based to reconstruct the noun list and fragment EDU to improve the precision of the EDU segment.

The paper is organized as follows: Section 2 presents the related works. Section 3 discusses the issues in Thai EDU segmentation. Section 4 describes the definition of Thai EDU. Section 5 presents the methodology from word segmentation to EDU segmentation. Section 6 presents the experimental results. And Section 7 is the conclusion.

## 2. RELATED WORKS

In Thai text, there are some features such as the absence of word and unclear word/sentence boundaries that makes it a complicated text to be processed [9]. The first problem in Thai text analysis is how to identify the word boundary. There is no space or any clue to identify the boundary of the word in Thai sentence. Thai sentence boundary is also unclear to identify. Thai sentences can appear as a continuous character stream in a paragraph with no explicit marker to indicate the boundary of the sentence. According to the problem of Thai sentence boundary, Thai EDU segmentation is also a complicated task. This section describes the chronicle of the research of Thai word segmentation, Thai sentence boundary detection, and Thai EDU segmentation.

### 2.1 Thai Word Segmentation

The algorithm to segment the word in a sentence in the first era is done by using the matching algorithm by dictionary-based [10, 11]. However, the only use of the matching algorithm cannot achieve satisfactory precision because the boundary of the word is very complicated to be identified with a straight matching algorithm. Machine learning are algorithms that enable systems to automatically learn from given data and are used in various research fields [12, 13, 14, 15]. Machine learning stepped in to play an important role to increase the precision of the word segmentation process. Kawtrakul and Thumkanon [16] proposed the statistical approach based on the tri-gram Markov model [17] with the dictionary to find the best path of all possibility of word segmentation path with part of speech tagging. Aroonmanakun [18] proposed a syllable-based tri-gram model and maximum collocation to solve word segmentation problem. Each syllable is merged by finding the maximum collocation between syllables and then construct the word. The Conditional Random Field (CRF) [19, 20] is introduced for Thai word segmentation with an impressive result. The possibilities of word segmentation are produced and then the CRF framework is used to select the optimal path of word segmentation with part of speech tagging [21, 22]. The deep learning technique is used by integrating two-level backoff models and character-level contexts to process the word segmentation and part of speech tagging [23]. Thai and Chinese languages are used in experiment and the results reveal that the model significantly improves the overall word segmentation. Nowadays, Thai word segmentation still is an active research area. Some machine learning can perform a good result of word segmentation. However, the precision of word segmentation can affect to the next processing. The high precision word segmentation is needed for accurate Thai Language processing.

### 2.2 Thai Sentence Boundary Detection

Thai sentence segmentation is an important Thai NLP process. There are no punctuation or explicit marker to identify the boundary of Thai sentence. In some researches, the space character is an essential element that used to indicate the sentence breaker. However, the space character can be used in many places in sentences such as before and after number, between coordinated words in lists. Mittrapiyanurak and Sornlertlamvanich [24] used the part of speech tri-gram model to compute the most probable sequence of part of speech (POS) to identify that space character is a breaking space. Charoenpornsawat and Sornlertlamvanich [25] proposed the winnow algorithm to solve the problem of Thai sentence boundary by learning words and POSs around the target space to disambiguate the sentence space breaker. After that, a maximum entropy classifier is applied to indicate the sentence breaking for the statistic machine translation (SMT) system [26]. A word labeling approach is introduced to identify the sentence boundary [27]. CRF is utilized to learn words and space to indicate the sentence breaking. The boundary detection of the Thai sentence is a difficult task due to the lack of punctuation and the informal use of space in the sentence. The processing of the piece of the sentence that is smaller than a sentence can be more useful for Thai NLP applications.

### 2.3 Thai EDU Segmentation

Thai EDU segmentation is needed rather than a sentence for some Thai NLP applications. Charoensuk, Sukvakree, and Kawtrakul [28] have an experiment on Thai EDU segmentation by using discourse cues and syntactic information. This work used correlative discourse marker, blank, and POS with classification rules to identify the starting and ending of EDU on an agriculture domain. Sinthupoun and Sornil [29] used a probabilistic approach to determine the EDU boundary and experiment on Thai family law corpus. Ketui, Theeramunkong, and Onsuwan [30, 31] developed a context-free grammar (CFG) rules to detect Thai EDU in Thai-NEST corpus. To develop an accurate Thai EDU segmentation, Syntactic information and explicit clue marker are very useful information to determine the boundary of Thai EDU. The EDU such as the noun list can be segmented to fragment of EDU that need more syntactic analysis to produce an accurate EDU segmentation.

## 3. ISSUES IN THAI EDU SEGMENTATION

From previous studies and our observation, there are some interesting issues in Thai EDU segmentation that we have to be concerned of. In general, the Thai language is a language structure of Subject-Verb-Object (S-V-O) likes English and many other languages. But some features in the Thai language can make a sentence more complicated and ambiguous. Those features are discussed in this section as follows:

- **Lack of Explicit EDU Boundary Marker**: In Thai text, there is no explicit marker or punctuation to indicate the ending of sentence or EDU. Thai text can be seen as a stream of continuous characters in a paragraph without any space character or punctuation. In some writing, a space character is used to indicate the ending of a sentence. However, a space character is just an option and does not appear in all sentence ending. Moreover, a space character can appear in many parts of sentence for example before and after a number, before conjunctions "และ (and)" "หรือ (or)", and before the repeater sign " ๆ".

- **Ambiguity of Word Marker**: Some words which are subordinate conjunction words can be considered to be a word marker to segment EDU for example "ซึ่ง (that)", "โดย (by)", "ดังนั้น (therefore)", "เพราะ (because)", "เพื่อ (for)". However, some words can be more than one POS depending on its function and its context.

In this issue, the accurate word segmentation and POS tagging process is an essential component to reduce the ambiguity.

- **Zero Anaphora**: Sentence in the Thai language can use a gap or omits the subject to refer back to the previous object. Then, the structure of the Thai sentence can be a Verb-Object (V-O) structure.

- **Sentential Noun Phrase**: Some noun phrases in the Thai language can be structured similar to a sentence. A verb can be a part of a noun phrase that can cause ambiguity of sentence and noun phrase. For example "ห้อง (room) นอน (sleep)" is meant to a bedroom. To overcome this issue, the accurate noun phrase chunker or shallow parser can be a key role to disambiguate sentential noun phrases and sentences.

- **Relative Clause in Noun Phrase**: Noun phrase in the Thai language can be embedded with a relative clause. Sentence with a relative clause will complicate the task to identify the boundary of EDU. Syntactic information can be useful to indicate the boundary of a relative clause.

## 4. DEFINITION OF THAI EDU

EDU is the minimal discourse unit from discourse structure. A sentence can consist of several EDUs. In our work, we define our EDU as follows.

- **Simple EDU**: EDU with a simple S-V-O structure. This EDU can consist of a subject, a verb, and an optional object. A preposition is possible to embed in this EDU structure also.

- **Zero Anaphoric EDU**: EDU with the omission of the subject can be a V-O structure. Same as a simple EDU, a preposition can also embed in this EDU structure.

- **Relative Clause EDU**: A relative clause, that is embedded in a noun phrase, is considered to be a separated EDU from its main structure. The function of a relative clause is a noun modification and its structure is similar to zero anaphoric EDU.

- **Noun List EDU**: Noun list EDU is part of a sentence that describes a list or an example of a noun group. This EDU mostly starts with word "ได้แก่ (for instance), เช่น (such as)".

## 5. METHODOLOGY

In this work, we use a data source from the Thai Wikipedia webpage[1]. The selected pages are downloaded to a database and then fed to the EDU segmentation process. The overview of the EDU segmentation process is depicted in Figure 1.

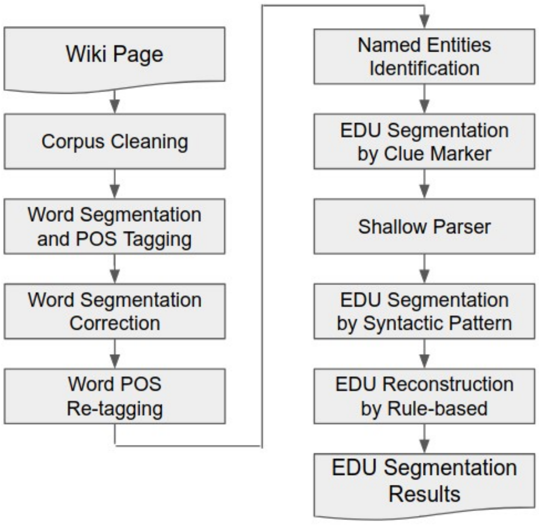

Wiki Page
Corpus Cleaning
Word Segmentation and POS Tagging
Word Segmentation Correction
Word POS Re-tagging
Named Entities Identification
EDU Segmentation by Clue Marker
Shallow Parser
EDU Segmentation by Syntactic Pattern
EDU Reconstruction by Rule-based
EDU Segmentation Results

*Figure 1: Overview of the EDU Segmentation Process*

The EDU Segmentation process consists of corpus cleaning, word segmentation, and POS tagging, word segmentation correction, word POS re-tagging, named entities identification, EDU segmentation by clue markers, shallow parser, EDU segmentation by syntactic pattern and EDU

[1] https://th.wikipedia.org/

reconstruction by rule-based. All sub-process will be described in this section.

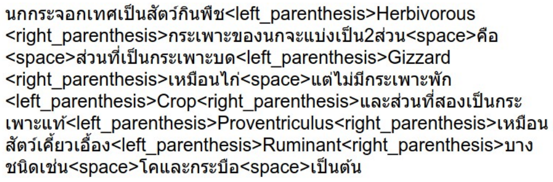
### 5.1 Corpus Cleaning

Our corpus is constructed from a selected Thai Wikipedia webpages. The webpage is cleaned by removing HTML tags and unused information. Then, the output is only a raw text from the webpage. All symbols and space are converted into a symbol tag. For example a space character is converted to "<space>", a left parenthesis is converted to "<left_parenthesis>". Table 1 shows examples of symbol tags.

*Table 1: Examples of Symbol Tags*

| Symbol Name | Symbol Character | Symbol Tag |
|---|---|---|
| Colon | : | <colon> |
| Semi Colon | ; | <semi_colon> |
| Comma | , | <comma> |
| Left Square Bracket | [ | <left_square_bracket> |
| Right Square Bracket | ] | <right_square_bracket> |
| Left Parenthesis | ( | <left_parenthesis> |
| Right Parenthesis | ) | <right_parenthesis> |
| Space | | <space> |

In this process, some space characters, that are insignificant, are removed such as space before and after a number, a space before repeater sign. This cleaned corpus is the data source for training and experiment in the next process. Figure 2 shows an example of a cleaned corpus with symbol tags.



นกกระจอกเทศเป็นสัตว์กินพืช<left_parenthesis>Herbivorous
<right_parenthesis>กระเพาะของนกจะแบ่งเป็น2ส่วน<space>คือ
<space>ส่วนที่เป็นกระเพาะบด<left_parenthesis>Gizzard
<right_parenthesis>เหมือนไก่<space>แต่ไม่มีกระเพาะพัก
<left_parenthesis>Crop<right_parenthesis>และส่วนที่สองเป็นกระ
เพาะแท้<left_parenthesis>Proventriculus<right_parenthesis>เหมือน
สัตว์เคี้ยวเอื้อง<left_parenthesis>Ruminant<right_parenthesis>บาง
ชนิดเช่น<space>โคและกระบือ<space>เป็นต้น

*Figure 2: Example of A Cleaned Corpus with Symbol Tags*

### 5.2 Word Segmentation and POS Tagging

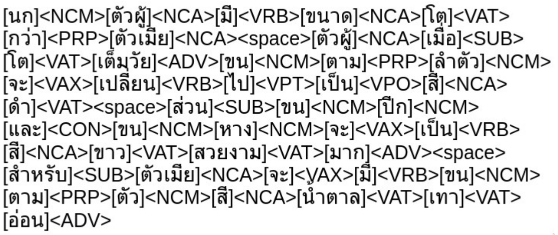
The CRF algorithm will play an important role in word segmentation and POS tagging process in this work. Tagged corpus, dictionary, and POS features are developed to train the CRF model. Examples of the POS tagging label, that is used in this work, is shown in Table 2.

*Table 2: Examples of The POS Tagging Label*

| POS Label | Description | Example Words |
|---|---|---|
| NCM | Common Noun | นก ลิง เศรษฐกิจ |
| NCA | Noun and Classifier for Attribute, Kind and Group | สี อายุ ประเภท ชนิด |
| NNA | Noun for Amount | สิบ ร้อย พัน หมื่น |
| VRB | Transitive Verb | กิน มอง รัก |
| VRI | Intransitive Verb | ยิ้ม วิ่ง เดิน |
| VAT | Attribute Verb | ใหญ่ หนัก สวย |
| FVN | Prefix to transform Noun and Verb to be Noun | การ |
| NBO | Ordinal Number Word | แรก สุดท้าย ต่อไป |

After that, word segmentation correction is developed to increase the precision of word segmentation and then POS tagging is redetermined to fix some incorrect POS. All processes are described as follows.

- **Manual Tagged Corpus for Training**: Corpus with manual tagging is needed to train the CRF model. POS tags will be added to the corpus by manual for training the CRF model. In this process, a word boundary is determined and the POS tag is defined to each word in the corpus. Square Bracket is used to identify the boundary of the word and followed by the POS tag in "less than" and "greater than" symbol. Free text with manual tagged will be the source of the CRF training model. Figure 3 shows an example of a manually tagged corpus.



```
[นก]<NCM>[ตัวผู้]<NCA>[มี]<VRB>[ขนาด]<NCA>[โต]<VAT>
[กว่า]<PRP>[ตัวเมีย]<NCA><space>[ตัวผู้]<NCA>[เมื่อ]<SUB>
[โต]<VAT>[เต็มวัย]<ADV>[ขน]<NCM>[ตาม]<PRP>[ลำตัว]<NCM>
[จะ]<VAX>[เปลี่ยน]<VRB>[ไป]<VPT>[เป็น]<VPO>[สี]<NCA>
[ดำ]<VAT><space>[ส่วน]<SUB>[ขน]<NCM>[ปีก]<NCM>
[และ]<CON>[ขน]<NCM>[หาง]<NCM>[จะ]<VAX>[เป็น]<VRB>
[สี]<NCA>[ขาว]<VAT>[สวยงาม]<VAT>[มาก]<ADV><space>
[สำหรับ]<SUB>[ตัวเมีย]<NCA>[จะ]<VAX>[มี]<VRB>[ขน]<NCM>
[ตาม]<PRP>[ตัว]<NCM>[สี]<NCA>[น้ำตาล]<VAT>[เทา]<VAT>
[อ่อน]<ADV>
```

*Figure 3: Example of A Manually Tagged Corpus*

- **Dictionary Development**: After finished the corpus tagging process, a dictionary is developed by gathering words from the tagged corpus and also its POS. Dictionary is an essential resource to generate all possibility segment sequences of word segmentation for word segmentation process. Table 3 shows examples of a dictionary.

*Table 3: Examples of A Dictionary*

| Word | POS |
|---|---|
| นกพิราบ | NCM |
| อยู่ | VRB VPO VPT VPA VPR |
| อย่าง | FAV CLS |
| บาง | DQE VAT |
| สวยงาม | VAT ADV |
| กำลัง | VAX NCM |
| ให้ | VPO VRB VPR VPT VPA |

- **CRF Training**: POS feature is needed for the CRF model to train and also to determine the best POS tag in word segment sequence. POS feature is the sequence pattern of POS of words that appear in corpus data. POS features are extracted from the tagged corpus and then collected to the database. After that, the CRF model is trained by using data from the tagged corpus and POS feature database to adjust the weight of the POS feature.

CRF is the probability of the label sequence that globally normalizes to avoid the label-bias problem and also provides the flexibility to use non-independent features [19]. Let *l* be a label of POS sequence, *s* be a word segmentation, *f* be a feature function, and $\lambda$ be the weight of feature function. Then, we can define the score function *score(l|s)* to determine the score of POS sequence to the sentence as shown in Equation 1.

$$score(l|s) = \sum_{j=1}^{m}\sum_{i=1}^{n}\lambda_j f_j(s,i,l_i) \qquad (1)$$

The conditional probability distribution *p(l|s)* for a linear-chain CRF can be defined as shown in Equation 2.

To estimate the weight of feature function $\lambda$ for training data, Equation 3 can be defined where $\alpha$ is a learning rate.

The features of learning for POS tagging are the POS sequence pattern that is extracted from the tagged corpus. There are 11 types of POS sequence pattern combination in this work that is shown in Table 4. The notation *V* is the POS in the position that is determined, *X* is the POS in the previous position that is determined, *Y* is the POS in next position that is determined, *W* is a surface word, and then, *none* is the position that is not determined.

$$p(l|s) = \frac{\exp\left[score(l|s)\right]}{\sum_{l'}\exp\left[score(l'|s)\right]} = \frac{\exp\left[\sum_{j=1}^{m}\sum_{i=1}^{n}\lambda_j f_j(s,i,l_i)\right]}{\sum_{l'}\exp\left[\sum_{j=1}^{m}\sum_{i=1}^{n}\lambda_j f_j(s,i,l'_i)\right]} \qquad (2)$$

$$\lambda_i = \lambda_i + \alpha\left[\sum_{j=1}^{m} f_i(s,j,l_j) - \sum_{l'} p(l'|s)\sum_{j=1}^{m} f_i(s,j,l'_j)\right] \qquad (3)$$

*Table 4: POS Pattern Rules*

| Feature Type | POS Pattern |
|---|---|
| 1 | $none : X_1 : V : none : none$ |
| 2 | $X_2 : X_1 : V : none : none$ |
| 3 | $none : none : V : Y_1 : none$ |
| 4 | $none : none : V : Y_1 : Y_2$ |
| 5 | $none : X_1 : V : Y_1 : none$ |
| 6 | $X_2 : X_1 : V : Y_1 : none$ |
| 7 | $none : X_1 : V : Y_1 : Y_2$ |
| 8 | $X_2 : X_1 : V : Y_1 : Y_2$ |
| 9 | $none : X_1 : V : Y_1 : none\#W$ |
| 10 | $none : X_1 : V : none : none\#W$ |
| 11 | $none : none : V : Y_1 : none\#W$ |

- **Word Segmentation Process**: After finished training, the CRF model is ready to determine word segmentation and POS tagging. Raw text after the cleaning process is the input for the word segmentation process. By using a dictionary, all possibilities of word segmentation sequence are generated. Due to the time complexity, the word segmentation sequence will be generated limit to only 4 words each time. CRF is used to determine the best POS tag for each sequence by selecting the best probability. The best word segmentation sequence is selected by the best score. The first word in the best word segmentation sequence will be selected to the answer sequence. After that, the next word segmentation sequence will be generated and re-operated again until the last word is determined.

- **Word Segmentation Correction**: Words in the Thai language can be composed of more than one other word and become a new meaning. For example "พื้นที่ (area)" is composed of 2 words are "พื้น (floor)" and "ที่ (at)". Word segmentation by the CRF model tends to segment words into more chunks. Dictionary can be a useful resource to correct word segmentation [32]. To correct word segmentation, we apply POS pattern matching together with a dictionary. POS patterns are used to match and find the word in Dictionary with the POS target. After that, words with POS pattern is replaced by word with POS target from a dictionary. Table 5 shows the examples of the POS pattern for word correction.

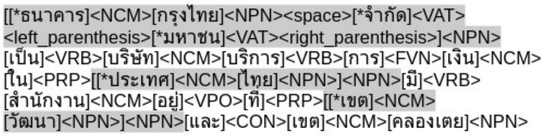
*Table 5: Examples of The POS Pattern for Word Correction*

| POS Pattern | POS Target Word |
|---|---|
| CON NCM | VAC |
| VAC NCM | VAC |
| NCM PRP | NCM |
| NCM PRL | NCM |

- **Word POS Re-tagging**: After the word correction process, there are some mistaken POS tagging that could be fixed. All POS tagging will be cleared and reconsidered by the POS re-tagging process. This process can improve the precision of POS tagging in the word sequence.

### 5.3 Named Entities Identification

The template matching technique is used to identify the named entities in the text. POS tag and surface word are the component of a matching template for extracting named entities. Corpus is tagged manually to create the matching template. POS tag is the main component of the template and surface word is an optional component. Figure 4 shows an example of the corpus with a manual named entity tagging.



```
[[*ธนาคาร]<NCM>[กรุงไทย]<NPN><space>[*จำกัด]<VAT>
<left_parenthesis>[*มหาชน]<VAT><right_parenthesis>]<NPN>
[เป็น]<VRB>[บริษัท]<NCM>[บริการ]<VRB>[การ]<FVN>[เงิน]<NCM>
[ใน]<PRP>[[*ประเทศ]<NCM>[ไทย]<NPN>]<NPN>[มี]<VRB>
[สำนักงาน]<NCM>[อยู่]<VPO>[ที่]<PRP>[[*เขต]<NCM>
[วัฒนา]<NPN>]<NPN>[และ]<CON>[เขต]<NCM>[คลองเตย]<NPN>
```

*Figure 4: Example of The Corpus with A Manual Named Entity Tagging*

In Figure 4, the surface word with * symbol means to indicate that the surface word is also used in matching. After that, the matching template is created by extracting it from the tagged corpus. Table 6 shows the examples of the named entity matching template.

*Table 6: Examples of The Named Entity Matching Template*

| Pattern | Tag |
|---|---|
| [เขต]<NCM>[*]<NPN> | NPN |
| [ธนาคาร]<NCM>[แห่ง]<PRP>[ประเทศ]<NCM>[*]<NPN> | NPN |
| [ธนาคาร]<NCM>[*]<NCM><space>[จำกัด]<VAT> | NPN |
| [บริษัท]<NCM><space>[*]<NPN>[*]<NCM><space>[จำกัด]<VAT> | NPN |

In Table 6, the * symbol in matching template means to indicate that it is matching with any surface word. The matching algorithm works with one to one matching on words in the whole line of text. The size of the template is an important factor for matching. The longest template will be activated first and then the next longest.

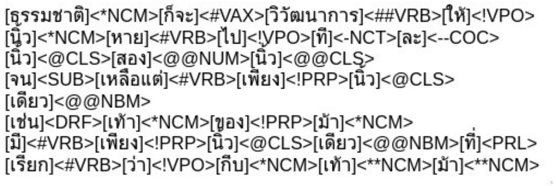
### 5.4 EDU Segmentation by Clue Markers

EDU segmentation in Thai text can be partially done by using a clue marker word to break the whole line of text to a piece of EDU. Some clue words with its POS tag are used to identify the origin of EDU and then segment into a smaller discourse unit. Words and POS tag patterns such as space, subordinate, and conjunction with verbs are examples of clue marker patterns.

### 5.5 Shallow Parser

A shallow parser is a process to identify the non-recursive of various phrase types [8]. The various phrase in a sentence can be chunked and is useful to be a precursor to a full parser or information extraction. In this work, CRF is applied to identify a non-recursive phrase.

- **Manual Tagged Corpus for Training**: Corpus with manual chunked tagging is used for training the CRF model. There are 11 types of phrases, that are chunked, consist of head noun, verbal noun, adjective, transitive verb, intransitive verb, adverb, preposition, amount, time, determiner, and classifier phrase. Some symbols are used to tag into a corpus to indicate a type of phrase in the text. Figure 5 shows an example of a corpus with a phrase tagged.



```
[ธูรรมชาติ]<*NCM>[ก็จะ]<#VAX>[วิวัฒนาการ]<##VRB>[ให้]<!VPO>
[นิ้ว]<*NCM>[หาย]<#VRB>[ไป]<!VPO>[ที]<-NCT>[ละ]<--COC>
[นิ้ว]<@CLS>[สอง]<@@NUM>[นิ้ว]<@@CLS>
[จน]<SUB>[เหลือแต่]<#VRB>[เพียง]<!PRP>[นิ้ว]<@CLS>
[เดียว]<@@NBM>
[เช่น]<DRF>[เท้า]<*NCM>[ของ]<!PRP>[ม้า]<*NCM>
[มี]<#VRB>[เพียง]<!PRP>[นิ้ว]<@CLS>[เดียว]<@@NBM>[ที่]<PRL>
[เรียก]<#VRB>[ว่า]<!VPO>[กีบ]<*NCM>[เท้า]<**NCM>[ม้า]<**NCM>
```

*Figure 5: Example of A Corpus with A Phrase Tagged*

The symbols, that is used to indicate phrase type, consist of "*", "\&", "+", "\#", "\%", "\$", "!", "@", "=", "?" and "-" for head noun pattern, verbal noun pattern, adjective pattern, transitive verb pattern, intransitive verb pattern, adverb pattern, preposition pattern, amount pattern, time pattern, determiner pattern and classifier pattern respectively. Single symbol tagged for the starting element of phrase type and double symbol for the following element of phrase type. Some POSs maybe not tagged if it is not a member of non-recursive phrase for example "SUB" subordinated, "PRL" relative pronoun.

- **Dictionary for Shallow Parser**: Dictionary for the shallow parser is used to indicate all possible phrase types to POS. All data in a dictionary are gathered from the tagged corpus. Table 7 shows examples of a dictionary for the shallow parser.

*Table 7: Examples of A Dictionary for The Shallow Parser*

| POS | Phrase Types |
|---|---|
| VAT | + ++ && $ $$ ?? == |
| FAV | $ $$ && |
| NPN | * ** = == ?? ++ |
| SUB | SUB |
| VRB | # ## && == ** $$ |
| VPA | $ $$ |

- **Feature for CRF Training**: There are 3 types of features that are used for training. A combination of phrase type(PT), POS, and surface word is constructed to make a shallow parser feature. Table 8 shows the shallow parser feature pattern.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

*Table 8: Shallow Parser Feature Pattern*

| Feature Type | Combination of Pattern |
|---|---|
| 1 | $PT_1 : PT_0 : POS_1 : POS_0 : none : none$ |
| 2 | $PT_1 : PT_0 : POS_1 : POS_0 : none : WORD_0$ |
| 3 | $PT_1 : PT_0 : POS_1 : POS_0 : WORD_1 : none$ |

The notation "PT" is the phrase type and the subscription is the position that is determined. The notation "POS" is the POS and the notation "WORD" is the surface word.

### 5.6 EDU Segmentation by Syntactic Pattern

The result of the shallow parser is a phrase chunked in a given text. The syntactic information from phrase chunked is a source data for EDU segmentation. From observation, some points of syntactic structure can indicate the point of EDU segmentation. Table 9 shows examples of the syntactic pattern.

*Table 9: Examples of The Syntactic Pattern*

| Syntactic Pattern |
|---|
| VRBpat:PRPpat:NP::VRBpat |
| VRBpat:NP::VRIpat |
| VRBpat:NP:ADVpat:PRPpat:NP::VRIpat |
| VRIpat::PRPpat:VRBpat |
| VRIpat:PRPpat:NP::PRL:VRBpat |
| NP2::NP:VRBpat |
| NP2::CON:VRBpat |

The pattern consists of the sequence of phrase types that connect with the colon symbol. The EDU segment point is indicated by the double colon symbol. NP and NP2 in the pattern are the encapsulated phrase type that consists of the head noun, preposition, verbal noun, amount, time, adjective, and determiner pattern. In some contexts, the EDU can be constructed by an only head noun and adjective. That means some EDUs are looked like a noun phrase. In this work, we define that NP is a noun phrase on the EDU and NP2 is a noun phrase that is an EDU. We encapsulate NP and NP2 by using rules that NP2 can be composed of the head noun, adjective, and adverb that no preposition is before the head noun and it is not followed by a verb. NP can be composed of the head noun, preposition, verbal noun, amount, time, adjective, and determiner.

### 5.7 EDU Reconstruction by Rule-Based

Noun list can be separated by space that needs to be reconstructed. Moreover, space is used in some writing style to separate some word such as subordinate word, noun phrase, or some parts of EDU that breaks the EDU structure by EDU segmentation by clue marker process. Some EDUs need to be reconstructed to increase the precision of the EDU structure. Rule-based is applied to analyze the partial EDU and then construct the new EDU structure. The rule consists of 3 parts: starting condition, combined condition, and commit condition. Table 10 shows examples of the EDU reconstruction rule.

*Table 10: Examples of the EDU Reconstruction Rule*

| No. | Starting | Combined | Commit |
|---|---|---|---|
| 1 | Line is end with "PRPpat" | Line is NP | Line is not NP and not bound |
| 2 | Line is end with "VRBpat" | Line is NP | Line is not NP and not bound |
| 3 | Line is NP | Line is NP | Line start with "VRBpat" and bound |

The EDU reconstruction process starts by finding the line that matches the starting condition. After that, if the next line match with the combined condition then merges this line with the starting line. Until the combined condition is not matched then check with the commit condition. If the commit condition is matched then check that the commit line could be bound with the new EDU structure. If it is bounded then merge the last line with the new EDU structure. Lastly, if the commit condition is not matched then all the processes are canceled and then try to the next rules.

## 6. EXPERIMENTAL RESULTS

Our corpus for training contains a total of 18,248 words with POS tagging. Dictionary contains 2,171 words for word segmentation and 52 words for the shallow parser. There are 235 POS patterns for the word correction process. There are 58,750 features of the CRF model for word segmentation and 11,975 features for the shallow parser. There are 114 syntactic patterns and 12 EDU reconstruction rules in the EDU segmentation process. The precision, recall, and F1 score are used to evaluate the algorithm. The measures can be defined as follows.

$$Precision = \frac{\#\,of\ correct\ tokens\ by\ algorithm}{\#\,of tokens\ determined\ by\ algorithm}$$

$$Recall = \frac{\#\,of\ correct\ tokens\ by\ algorithm}{\#\,of\ tokens\ in\ corpus}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

The results of word segmentation are shown in Table 11 and Figure 6 shows the bar chart of the results of word segmentation. CRF shows a good performance for word segmentation. However, CRF tends to segment words that could be one word to be more chunks. Word segmentation by machine learning [16, 18, 21, 22] produces some errors due to segmented words into more chunks. The extra process is needed to gain more precision in the word segmentation process. To increase the precision of results, the word segmentation correction and the POS re-tagging process are applied. The fragment word is reconstructed into a correct word by word segmentation correction process that can boost the precision and recall significantly. The re-tagging process can increase a little of the precision and recall.

The clue marker is useful to partially segment EDU. Some parts of EDU need syntactic information to indicate the point of EDU Segmentation and then reconstruct the EDU structure. Table 12 shows the results of EDU segmentation from all algorithms and Figure 7 shows the bar chart of the results of EDU segmentation.

*Table 11: Results of Word Segmentation*

| Algorithm | Precision | Recall | F1 |
|---|---|---|---|
| Word Segmentation | 0.92688 | 0.96153 | 0.94388 |
| Word Segmentation & Correction | 0.98946 | 0.98854 | 0.98899 |
| Word Segmentation & Correction & Re-Tagging | 0.99544 | 0.99451 | 0.99452 |

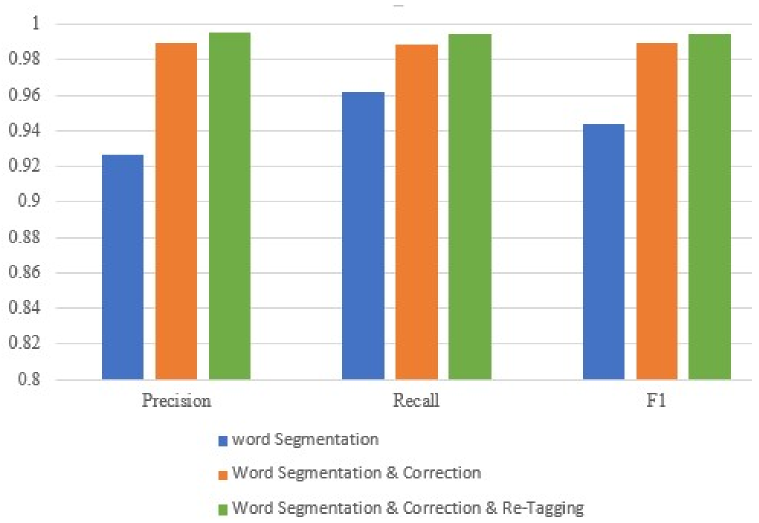


*Figure 6: Bar Chart of The Results of Word Segmentation*

The results show that the use of a clue marker alone produces the EDU segment with low precision to 0.41085 and recall to 0.36441. Many EDUs appear without a clue marker to indicate the boundary. However, a clue marker can be a starting tool to indicate the EDU boundary [28]. Syntactic information is used in various ways to indicate the boundary of EDU [29, 30]. Syntactic pattern from shallow parser can be a good option to identify the point of EDU segmentation. The use of syntactic pattern significantly increase the precision and recall, however, noun list EDU and the use of space in some writing style still indicate wrong EDU boundary. Finally, the EDU reconstruction process is used to combine noun list EDU and some fragment EDUs to produce a more precise EDU with 0.88865 precision and 0.91577 recall.

## 7. LIMITATIONS

The source of this work is a freestyle text in an open domain document. However, the accomplishment of this work is based on some limitations as follows.

- **No misspelling word**: Misspelling words can occur in the text normally. The whole system expects accurate spelling text as a data source. The misspelling word correction is not in our interesting scope of work. Spelling word correction is done by manual before feeding to the whole system.
- **Ignore the additional information from parenthesis**: There is the use of parenthesis in the text to provide additional

data such as abbreviations, synonym, and foreign words. Text in parenthesis can be in free form depending on the aspect of the writer. In this work, text in parenthesis is ignoring and pruning by the corpus cleaning process.

- **Do not follow the structure information from a table**: Structure information, such as a table, is widely used in the text. A table is not a part of EDU in the text and is not in our interesting scope of work. Tables in source corpus are pruning by the corpus cleaning process.

*Table 12: Results of EDU Segmentation*

| Algorithm of EDU Segmentation | Precision | Recall | F1 |
|---|---|---|---|
| Using Clue Markers | 0.41085 | 0.36441 | 0.38623 |
| Using Clue Marker<br>& Syntactic Pattern | 0.68788 | 0.81263 | 0.74506 |
| Using Clue Marker<br>& Syntactic Pattern<br>& EDU Reconstruction | 0.88865 | 0.91577 | 0.90200 |

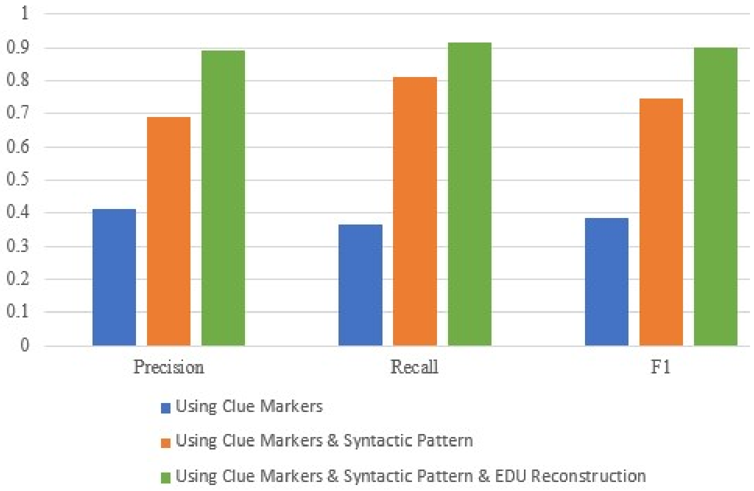


*Figure 7: Bar Chart of The Results of EDU Segmentation*

## 8. CONCLUSION

In this paper, we present a pipeline of the process for Thai EDU segmentation by using the syntactic information from shallow parser. The first step, word segmentation and POS tagging process are done by using CRF to identify the word boundary and its POS with 0.92688 precision and 0.96153 recall. The CRF shows the good result in word segmentation, however, some composition words tend to segment into fragment words. To improve the precision of word segmentation and POS tagging process, we implemented the word segmentation correction by using a POS pattern with a dictionary to merge some fragment words and improve the precision to 0.98946 and recall to 0.98854. POS re-tagging process is used to redetermine the POS label that further improved the precision and recall of POS tagging to 0.99544 and 0.99451 respectively.

In the second step, the shallow parser is used to determine the POS sequence to phrase types such as head noun, amount, adjective, and determiner pattern. The EDU segmentation process is done by using a clue marker, syntactic pattern, and reconstruction rule-based. The clue marker indicates the EDU segment with 0.41085 precision and 0.36441 recall. The syntactic pattern is used to identify the EDU segment with the improvement of precision and recall to 0.68788 and 0.81263 respectively. The fragment of noun list EDUs and some partial EDUs are reconstructed by rule-based and finally improve the precision and recall to 0.88865 and 0.91577 respectively.

From the corpus development, we found that some writing styles use the space to make it more readable in some words or some parts of the sentence. The non-formal use of space can cause some difficulties to identify the boundary of EDU and also increase ambiguity. A preposition and noun list are the most apparent of non-formal use of space in our corpus. Moreover, POS label designing is also an important component to make a more reliable grammar structure in the corpus. POS labels with embedded semantic can be a useful resource to analyze the structure of EDU and sentence.

## REFERENCES

[1] Ponrudee Netisopakul and Gerhard Wohlgenannt. The State of Knowledge Extraction from Text for Thai Language. In Advanced Applied Informatics (IIAI-AAI), pages 379–384. IEEE, 2017.

[2] Ponrudee Netisopakul and Gerhard Wohlgenannt. A Survey of Thai Knowledge Extraction for the Semantic Web Research and Tools. IEICE TRANSACTIONS on Information and Systems, 101(4):986–1002, 2018.

[3] Thana Sukvaree, Asanee Kawtrakul, and Jean Caelen. Thai text coherence structuring with coordinating and subordinating relations for text summarization. In International and Interdisciplinary Conference on Modeling and Using Context, 2007.

[4] Nongnuch Ketui, Thanaruk Theeramunkong, and Chutamanee Onsuwan. An EDU-based approach for Thai multi-document summarization and its application. ACM Transactions on Asian and Low-Resource Language Information Processing, 14(1):1–26, 2015.

[5] Daniel Marcu. A surface-based approach to identifying discourse markers and elementary textual units in unrestricted texts. Proceedings of COLING-ACL Workshop on Discourse Relations and Discourse Markers, pages 1–7, 1998.

[6] Daniel Marcu. A decision-based approach to rhetorical parsing. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, 1999.

[7] Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In Current and new directions in discourse and dialogue, pages 85–112. Springer, 2003.

[8] Fei Sha and Fernando Pereira. Shallow parsing with conditional random fields. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language TechnologyVolume 1, pages 134–141. Association for Computational Linguistics, 2003.

[9] Wirote Aroonmanakun. Thoughts on Word and Sentence Segmentation in Thai. In Proceedings of the Seventh Symposium on Natural Language Processing, pages 85–90, 2007.

[10] Yuen Poowarawan. Dictionary-based thai syllable separation. In Proceedings of the Ninth Electronics Engineering Conference, pages 409–418, 1986.

[11] Virach Sornlertlamvanich. Word segmentation for Thai in machine translation system. Machine Translation, National Electronics and Computer Technology Center, Bangkok, pages 50–56, 1993.

[12] Z Mustaffa, Y Yusof, and S. S. Kamaruddin. Application of LSSVM by ABC in energy commodity price forecasting. In 2014 IEEE 8th

International Power Engineering and Optimization Conference (PEOCO2014), pages 94–98, 2014.

[13] Yuhanis Yusof, Farzana Kabir Ahmad, Siti Sakira Kamaruddin, Mohd Hasbullah Omar, and Athraa Jasim Mohamed. Short term traffic forecasting based on hybrid of firefly algorithm and least squares support vector machine. In International Conference on Soft Computing in Data Science, pages 164–173. Springer, 2015.

[14] Azizi Ab Aziz, Faudziah Ahmad, Nooraini Yusof, Farzana Kabir Ahmad, and Shahrul Azmi Mohd Yusof. Designing a robot-assisted therapy for individuals with anxiety traits and states. In the 2015 International Symposium on Agents, Multi-Agent Systems and Robotics (ISAMSR), pages 98–103, 2016.

[15] Ghaith Alkubaisi, Abdulsattar A. Jabbar, Siti Sakira Kamaruddin, and Husniza Husni. Conceptual framework for stock market classification model using sentiment analysis on twitter based on Hybrid Naïve Bayes Classifiers. International Journal of Engineering Technology, 7(2.14):57–61, 2018.

[16] Asanee Kawtrakul and Chalatip Thumkanon. A statistical approach to thai morphological analyzer. In Proceedings of of the 5th Workshop on Very Large Corpora, 1997.

[17] Eugene Charniak. Statistical language learning. MIT press, 1996.

[18] Wirote Aroonmanakun. Collocation and Thai Word Segmentation. In Proceedings of the Fifth Symposium on Natural Language Processing & The Fifth Oriental COCOSDA Workshop, 2002.

[19] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the eighteenth international conference on machine learning, ICML, pages 282–289, 2001.

[20] Fuchun Peng, Fangfang Feng, and Andrew McCallum. In Proceedings of the 20th international conference on Computational Linguistics, page 562.

[21] Canasai Kruengkrai, Virach Sornlertlamvanich, and Hitoshi Isahara. A Conditional Random Field Framework for Thai Morphological Analysis. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-2006), pages 2419–2424, 2006.

[22] Sarawoot Kongyoung, Anocha Rugchatjaroen, and Krit Kosawat. TLex+: A hybrid method using conditional random fields and dictionaries for Thai word segmentation. In International Conference on Knowledge, Information, and Creativity Support Systems, volume 685, pages 112—-125. Springer, 2015.

[23] Prachya Boonkwan and Thepchai Supnithi. Bidirectional Deep Learning of Context Representation for Joint Word Segmentation and POS Tagging. In International Conference on Computer Science, Applied Mathematics and Applications, pages 184–196. Springer, 2017.

[24] Pradit Mittrapiyanurak and Virach Sornlertlamvanich. The Automatic Thai Sentence Extraction. In Proceedings of the Fourth Symposium on Natural Language Processing, pages 23–28, 2000.

[25] Paisarn Charoenpornsawat and Virach Sornlertlamvanich. Automatic sentence break disambiguation for Thai. In International Conference on Computer Processing of Oriental Languages (ICCPOL), volume 33, pages 231–235, 2001.

[26] Glenn Slayden, Mei-Yuh Hwang, and Lee Schwartz. Thai Sentence-Breaking for LargeScale SMT. In Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing, number August, pages 8–16, 2010.

[27] Nina Zhou, AiTi Aw, Nattadaporn Lertcheva, and Xuancong Wang. A Word Labelling Approach to Thai Sentence Boundary Detection and POS Tagging. In Proceedings of the 26th International Conference on Computational Linguistics (COLING-16), pages 319–327, 2016.

[28] Jirawan Charoensuk, Thana Sukvakree, and Asanee Kawtrakul. Elementary Discourse Unit Segmentation for Thai Using Discourse Cues and Syntactic Information. In NCSEC2005, volume 6, pages 897–902, 2005.

[29] Somnuk Sinthupoun and Ohm Sornil. Thai rhetorical structure analysis. Journal of Computer Science and Information Security, 7(1), 2010.

[30] Nongnuch Ketui, Thanaruk Theeramunkong, and Chutamanee Onsuwan. A rule-based method for thai elementary discourse unit segmentation (ted-seg). In 2012 Seventh International Conference on Knowledge, Information and Creativity Support Systems, pages 195–202. IEEE, 2012.

[31] Nongnuch Ketui, Thanaruk Theeramunkong, and Chutamanee Onsuwan. Thai elementary

discourse unit analysis and syntactic-based segmentation. Information (Japan), 16(10):7423–7436, 2013.

[32] Rungsiman Nararatwong, Natthawut Kertkeidkachorn, Nagul Cooharojananone, and Hitoshi Okada. Improving Thai word and sentence segmentation using linguistic knowledge. IEICE Transactions on Information and Systems, 101(12):3218–3225, 2018.