A Statistical and Rule-based Method for Chunking Verbal Units in Thai Texts

Pimnapa Atsawintarangkun [1], Thanaruk Theeramunkong [1] and Choochart Haruechaiyasak [2]

[1] School of Information, Computer and Communication Technology Sirindhorn International Institute of Technology, Thammasat University, Pathum Thani, Thailand.

[2] Speech and Audio Technology Laboratory National Electronics and Computer Technology Center Pathum Thani, Thailand.

Correspondence:
Pimnapa Atsawintarangkun
School of Information, Computer and Communication Technology Sirindhorn International Institute of Technology, Thammasat University, Pathum Thani, Thailand.
Email: atsa.pimnapa@gmail.com

Abstract

This work focuses on how to extract a verbal unit, which is a group of words to express an action or state of being. A verbal unit is a basic and fundamental element of a clause or a sentence. In this work, we define three layers of verbal units including verbal sequences, verbal phrases (i.e., verbal chunks, causative forms and event occurrences), and elementary discourse units (EDUs). For the first layer, a verbal sequence is defined as a single verb or a sequence of contiguous verbs without any interrupting nouns or particles. As the second layer, a verb phrase (i.e., a causative form and event occurrence form) is defined as a phrase that may include auxiliary verbs, verbs and nouns as subjects or objects. In the third layer, a Thai elementary discourse unit is defined as a sentence-like or clause-like unit which includes only one actual verb per unit. We propose a hybrid approach by combining a statistical-based method and a rule-based method to chunk Thai verbal units. The statistical-based method used is based on a conditional random field while the rule-based method utilizes grammatical rules with chart parsing. These two methods can help each other to improve correctness. Compared are three approaches: statistical-based, a rule-based, and a hybrid approach. The experimental results show that the hybrid approach is the best approach to chunk verbal units.

Keywords: Chunking, Thai Verbal Sequence, CRF, Grammatical Rules, Hybrid Approach

1. Introduction

Tokenizing a text into a sequence of words is an important process towards text interpretation. This process is required in many applications such as text summarization, semantic search, and machine translation. Instead of splitting into words, recently there have been works on chunking into units which are larger than words. Text chunking is a process to divide a running text into non-overlapping groups of words, which have meaningful contents, such as named entities and verbal units.

In this work, we explore three layers of verbal units, called (1) verbal sequences, (2) verb phrases (i.e., verbal chunks, causative forms and event occurrences), and (3) elementary discourse units (EDUs). As the basic layer, a verbal sequence is defined as a single verb or a sequence of contiguous verbs without any interrupting nouns or particles. For example,

- Single verb

  Ex.: คิด (think), พูด (talk), กิน (eat)
- Compound verb

  Ex.: ติดตาม (pursue)
- Verb with a modifier (auxiliary)

  Ex.: จะทำงาน (will work), ตื่นแล้ว (woke up), ไม่เดิน (not walk)
- Serial verb

  Ex.: ไปทำงาน (go to work)

As the second layer, a verb phrase (i.e., a causative form and event occurrence form) is defined as a phrase that may include auxiliary verbs, verbs and objects.

- Causative form
  - "อาจทำให้ [+ นาม] + กริยา" (may cause s.o. [n] to do s.th.)
  - "ทำให้ [+ นาม] + กริยา" (make s.o. [n] to do s.th.)
  - "สั่งให้ [+ นาม] + กริยา" (order s.o. [n] to do s.th.)
- Event occurrence form (V = simple present tense verb)
  - "มีการ + กริยา" (there is Ving action)
  - "ทำการ + กริยา" (do Ving action)
  - "ทำให้มีการ + กริยา" (cause to do Ving action)

As the third layer, a Thai elementary discourse unit is defined as a sentence-like or clause-like unit which includes only one actual verb per unit. It may include a subject, object, prepositional phrases, and adverb phrases. Seven syntactic units for detecting Thai EDUs (T-EDUs) are described in [1]. The rest of this paper is organized as follows. In Section 2, some related works are reviewed. Section 3 describes a statistical-based approach for chunking verbal units. Our rule-based approach is presented in Section 4 and a hybrid approach is described in Section 5. Finally, Section 6 gives the conclusion.

2. Related Work

In this section, we discuss four topics: (1) characteristics of Thai language, (2) tagging and chunking, (3) Conditional Random Fields and (4) parsing and grammar.

2.1 Characteristics of Thai Language

Basically, a Thai sentence consists of a subject, a verb unit with some modifications, followed by an object [2]. Thai language has several difficulties in terms of computer processing. Firstly, three common levels of Thai grammatical units (i.e., words, phrases, and sentences) may usually be ambiguous since Thai text lacks word, phrase, and sentence boundaries. Secondly, Thai language normally omits pronouns in non-beginning sentences. Thirdly, there are several compound words that are composed of a sequence of nouns, their modifications or verbs. Their structures are the same with a sentence and they are called sentence-structured compound nouns. Therefore, it is difficult to differentiate between a sentence-structured compound noun and a sentence. Fourthly, many Thai function words are polysemous to content words and/or sometimes they are parts of Thai proper nouns, such as a person's name and an organization's name. For this type, the process of chunking might have a problem when one attempts to extract a verb, but it may appear as a part of a noun phrase.

2.2 Tagging and Chunking

Text chunking is the process to determine the boundaries of groups and phrases in a sentence. Various methods including rule-based and machine learning are used for text chunking while there is no method that is suitable for all cases. Many research works combine two or more methods to increase performance of chunking.

In the past, there have been a number of works on text chunking using rules in various languages such as Croatian (โครเอเชียน) and Arabic. These existing works achieved good performance mainly due to their fixed-word-order characteristics. Many researchers have used various machine learning techniques for chunking. Most of them utilized Support Vector Machines (SVMs) [3], Maximum Entropy Markov Models (MEMM) [4] or Conditional Random Fields (CRFs). Most of the existing works showed that the CRF-based approach is a prospective solution for chunking, such as those for English [5], Chinese [6], Korean [7], and Vietnamese [8]. However, there are only a few research works on developing tools for text chunking in Thai language. Considering the high performance achieved by related works using a hybrid approach [9, 10], we focus on chunking based on the hybrid method, which combines a statistical-based (CRFs) and a rule-based approach.

2.3 Conditional Random Fields

CRFs [11, 12] are undirected graphical models for segmenting and labeling structured data. The CRFs represent discriminative probabilistic models for computing the conditional distribution $p(y+x)$ which specifies dependencies over the observation sequence x. Let $x = (x_1,...,x_T)$ be a sequence of observations, i.e., input variables, such as a sequence of characters to be segmented and let $y = (y_1, ..., y_T)$ be a set of label sequences. A linear-chain CRF can be represented with probability $p(y|x)$ distribution written in the following form.

$$P(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k (y_{t-1}, y_t, x, t) \right) \tag{1}$$

where $Z(x)$ is a normalization parameter that is a sum over all possible state sequences and can be written as follows.

$$Z(x) = \sum_{y} \exp\left(\sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k (y_{t-1}, y_t, x, t) \right) \tag{2}$$

k is the index of feature function $f_k(y_{t-1}, y_t, x, t)$.

In this study, we use an open-source CRF tool called CRF++ [13] to perform our experiments. CRF++ is an open source implementation of CRF for segmenting and labeling sequential data. In CRF++, there are several advantages, including less memory usage both in training and testing, encoding/decoding in practical time, etc. CRF++ is written in C++ language and is applied in several tasks such as Named Entity Recognition, Information Extraction, and Text Chunking. CRF++ applies the limited memory quasi-Newton gradient-climber method (LBFGS) for fast training. To find the optimal path, the Viterbi algorithm is applied to obtain the most probable label sequence over the observation sequence x as follows.

$$y^* = \arg\max_{y} P(y|x) \tag{3}$$

2.4 Parsing and Grammar

Parsing [14] is a procedure of structural analysis to determine whether or not an input sentence is grammatical according to the grammar. The most popular grammar is Context-Free Grammar. Context-Free Grammar (CFG) is the most significant grammar formalism to describe language syntax. A CFG is often used as a base form since CFG is usually used for linguistic description, and most grammar formalisms are derived from or can be somehow related to CFG.

In this study, we focus on a chart parser. A chart represents the interaction between edges and vertices. A vertex indicates the position between any pair of words in a sentence and an edge represents an underlying rule. Chart parsing is a popular method which analyzes complex sentences or language structures that involve several rules. A chart parser constructs a graph and creates a set of seed inactive edges. A chart is relatively efficient because each constituent is generated exactly once, and it can search for only certain types of constituents, e.g., a noun phrase or verb phrase. The chart maintains the record of all the constituents derived from the input sentence so far in the parse.

3. Statistical-based Approach

3.1 Methodology

Three preliminary works using CRFs on three layers of verbal unit detection are proposed: (1) Thai verbal sequences chunking, (2) Thai verb phrases chunking [15, 16], and (3) Thai EDUs chunking [1].

With the training corpus, we employ three different feature sets: word (in Thai EDU, we called Token.), Part-Of-Speech (POS) tag, and their combination (by using both word and POS tag). To facilitate the learning and testing processes, the training corpus is formatted into three columns based on three feature sets as shown in this figure.

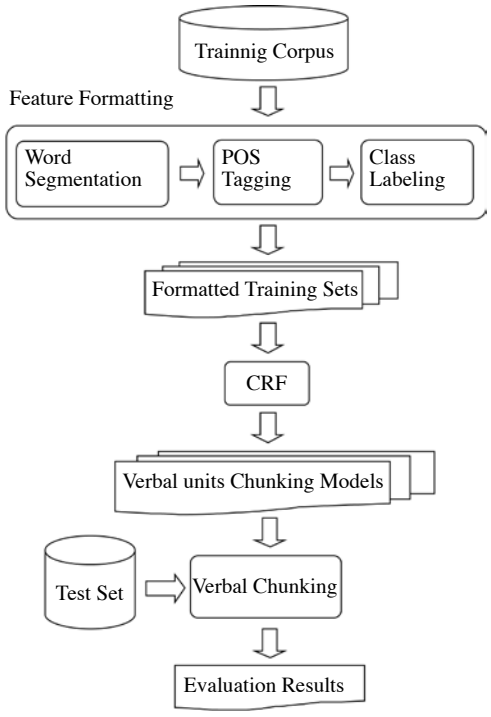

Fig. 1. The overall process of statistical-based approach.

Running Text

Word	POS tag	Class label
แท็กส์	NPRP	O
จะ	XVBM	B
แพ้	VACT	I
หรือ	EITT	O
ชนะ	VSTA	B
<space>	SPACE	O
จะ	XVBM	B
มี	VSTA	I
ผล	NCMN	O
ต่อ	RPRE	O
การ	FIXN	O
ตัดสินใจ	VACT	O

Fig. 2. An example of the formatted training set for chunking verbal units.

The first column contains Thai words. The second column contains the POS tags. For Thai verbal sequence and verb phrase, both columns are obtained from SWATH [17] while both columns of Thai EDU are obtained from Thai E-Class. The third column represents the annotated class labels which are separated into three classes: the beginning position of a verbal unit (B), the intra-verbal unit position (I), and the non-verbal unit position (O). The third column is obtained based on the definition of three layers of Thai verbal unit described in [1]. We train verbal unit chunking models using CRFs and use three feature sets based on n-gram where n is a window size. Finally, we utilize the test set to investigate the performance of three feature sets. To e valuate the performance of our models, we employ three performance measures: precision, recall, and F1.

3.2 Results

Table 1 shows the performance of verbal sequence chunking using CRFs based on 3-, 5-, 7-gram. The word feature set performs well and obtains more information due to actual words. However, the performance of the combination feature set is higher than the word and POS tag feature set, which rises to 81.48 %. Since the combination feature set consists of a word and POS tag, the POS tag feature set can help chunking patterns of words which cannot previously be handled with the word feature set. Consequently, the combination feature set, based on 3-gram, can achieve the best performance because the pattern of the verbal sequence contains a few words, which is about 1-5 words per verbal sequence.

Table 1 Performance in verbal sequences chunking using CRF.

		Word	POS	Combination
3-gram	Precision	82.97	73.23	84.00
	Recall	68.03	68.83	79.14
	F1	74.74	70.94	81.48
5-gram	Precision	81.38	72.36	83.29
	Recall	67.60	66.94	78.03
	F1	73.84	69.51	80.55
7-gram	Precision	81.43	72.77	83.10
	Recall	67.55	67.30	77.90
	F1	73.83	69.89	80.39

Table 2 shows the experimental results of 3-, 5- and 7-gram models. We compare the performance of Thai verb phrase chunking by using two different POS tag sets in the second column (POS tag feature set). The first POS tag set is the finer POS tagset from SWATH (Smart Word Analysis for Thai). There are 47 types. The second POS tagset is the coarser POS tag set. Some finer tags in the first tag are combined to be in the same group. For example, NOUN constitutes several subtypes of nouns, which are proper noun, cardinal number, ordinal number, label noun, common noun, and title noun. There are 16 types of the coarser POS. The performance of the Thai verb phrase chunking using the finer POS tagset shows that the POS tag feature set based on 3-gram gave the lowest performance of 37.32% in the F1 value, while the combination feature set based on 5-gram yielded the best performance of 77.56%. For the performance of chunking using the coarser POS tagset, the 3-gram has the lowest performance while the performance of 7-gram is higher than 5-gram (in word and combination feature set). Thus, we use 7-gram to compare performance in each feature set. The POS tag feature set yields the lowest performance of 45.89%, based on the F1 value. The word feature set gives better performance than the POS tag feature set, a performance improvement to 75.32%. The word feature set performs well and obtains more information due to actual words. However, the performance of the combination feature set is higher than the word and POS tag feature set, which rises to 77.03%. The combination feature set based on 5-gram and 7-gram yields better performance than the 3-gram because the pattern of a verb phrase contains more words than a verbal sequence. The verb phrase contains about 1-8 words per verb phrase while the verbal sequence contains about 1-5 words. Moreover, this table shows that the F1 of the finer POS is slightly better than the F1 of the coarser POS. It would therefore be fair to say that a larger POS tagset is preferable for Thai language. When we use only POS tags, the finer tagset achieves much better results than the coarser tagset. However, both finer and coarser tagsets show comparative results when both word and POS are considered.

Table 3 shows the result of each feature set in EDU chunking. The experimental results show that the POS tag has the lowest performance (F1 value) among three feature sets while the combination gives the highest performance in the F1 value. In the combination, the performance of the 2-gram model is 91.31% for the F1 value. From the experimental results, the token feature set achieves better performance than the entity/POS tag feature set because it has more information obtained from actual words. However, the performance of the combination feature set is higher than the token and entity/POS tag feature set because it is composed of a token and entity/POS tag. The entity/POS tag feature set can help segment patterns of units which cannot be handled by only the token feature set.

Table 2 Performance of Thai verb phrases chunking.

Feature set	n-gram	Results					
		Precision		Recall		F1	
		Finer POS	Coarser POS	Finer POS	Coarser POS	Finer POS	Coarser POS
word	3-gram	90.19		64.10		74.94	
	5-gram	91.32		63.70		75.05	
	7-gram	91.26		64.12		75.32	
POS tag	3-gram	65.91	63.60	26.03	11.67	37.32	19.71
	5-gram	85.81	73.39	49.93	33.38	63.13	45.89
	7-gram	84.35	76.41	48.94	31.90	61.94	45.01
Combination	3-gram	89.20	89.26	67.53	66.34	76.87	76.12
	5-gram	91.39	90.47	67.37	66.87	77.56	76.90
	7-gram	91.18	91.13	67.13	66.71	77.33	77.03

Table 3 Performance of Thai EDUs chunking.

Feature Sets	n-gram	Result		
		P	R	F1
Token	1-gram	52.83	17.39	26.17
	2-gram	91.56	87.96	89.72
	3-gram	89.70	84.91	87.24
	4-gram	89.43	84.37	86.83
	5-gram	89.47	84.33	86.82
Entity/POS tag	1-gram	58.81	43.98	50.33
	2-gram	59.48	45.33	51.45
	3-gram	59.76	48.60	53.60
	4-gram	60.97	51.90	56.07
	5-gram	59.31	50.99	54.84
Combination	1-gram	90.96	86.95	88.91
	2-gram	92.91	89.76	91.31
	3-gram	92.56	89.21	90.86
	4-gram	91.76	88.21	89.95
	5-gram	91.38	87.56	89.44

3.3 Discussion and error analysis

A number of incorrect results occurred when CRFs were used for chunking verbal units. They can be grouped into three cases. The first case is 'wrong word segmentation'. For example, the string "มี" should be part of the word "สามี" and the string "ทำ" should be part of "ทำเนียบ" as shown in Fig. 3.

The second case is incorrect POS tagging. For example, an ambiguous word "กำลัง" can be an auxiliary or noun. In Fig. 4, the word "กำลัง" should be a noun but it is tagged as an auxiliary verb. As for the second example, a word "โทษ" can be a verb or noun. In this figure, it should be a noun but it is tagged as a verb.

As for the third case, the incorrect results are derived from a pattern of some verbal units that are not provided in the training set. Two examples of this case are shown in Fig. 5.

Word	POS tag	Answer	Result
ตัว	CNIT	O	O
สา	NCMN	O	O
มี	VSTA	O	B
ภรรยา	NCMN	O	O

Word	POS tag	Answer	Result
ที่	PREL	O	O
ทำ	VACT	O	B
เนียบ	VACT	O	I
ประธานาธิบดี	NCMN	O	O

Fig. 3. Two examples of incorrect result of the first case.

Word	POS tag	Answer	Result
ใช้	VACT	B	B
กำลัง	XVBM	O	I

Word	POS tag	Answer	Result
และ	JCRG	O	O
โทษ	VACT	O	B
ใน	RPRE	O	O
มาตรา	CMTR	O	O

Fig. 4. Two examples of incorrect result of the second case.

Word	POS tag	Answer	Result
เอา	VACT	B	O
ไว้	XVAE	I	O
พร้อม	VSTA	I	O
เมื่อไหร่	PNTR	O	O

Word	POS tag	Answer	Result
การ	FIXN	O	O
เป็น	VSTA	O	O
แฟน	NCMN	O	O
ต้อง	XVMM	B	O
หึง	VACT	I	O

Fig. 5. Two examples of incorrect result of the third case.

4. Rule-based approach

4.1 Methodology

The overall process of our rule-based approach using grammatical rules is shown in Fig. 6.

In this approach, we sum up the rules by analysing the chunking results from the manually-tagged corpus. We use a chart parser [4] to analyse patterns of verbal sequences as a set of grammar rules in the training corpus. A chart represents the interaction between edges and vertices. A vertex indicates the position between any pair of words in a sentence and an edge represents an underlying rule. Chart parsing is a popular method which analyses complex sentences or language structures that involve several rules. The advantage of the rule-based approach is that it requires fewer resources. Moreover, it is possible to set a probability for each edge to rank the most probable rule. In the first experiment, we construct 605 grammatical rules from the tagged corpus. The rules are used as inputs to our chart parser. Then, inactive edges of verbal sequences are generated from the chart parsing process. We use the verbal sequence

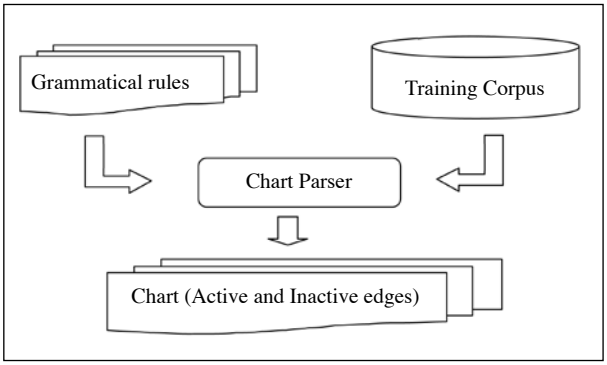

Fig. 6. Overall process of rule-based approach.

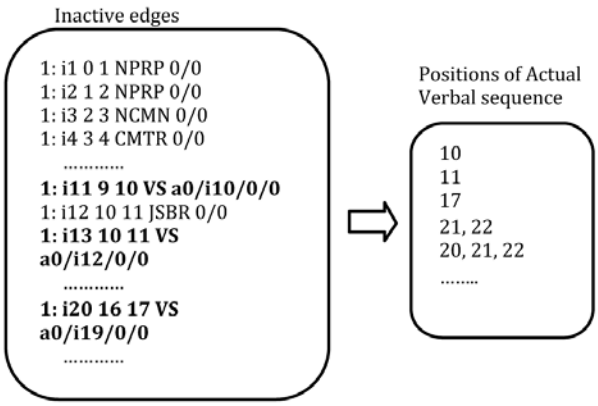

Fig. 7. Actual verbal-sequence/inactive edges.

inactive edges in the chart parser as shown in the Fig. 7 to find the next component. The positions of correct verbal sequences from the tagged corpus are listed as illustrated in Fig. 8. After that, we compare the position of correct verbal sequences with the position of verbal sequences from the chart parser as shown in Fig. 9.

In the second experiment, we use the set of grammatical rules and the procedure of the first experiment. However, we apply the longest matching technique to eliminate unnecessary and overlapped verbal units. Then, we compare positions of actual verbal sequences with the position of verbal sequences obtained from the chart parser with the longest matching technique. The dataset in this experiment is THAI-NEST corpus [18]. The dataset contains 1,879 processing units. Each processing unit includes approximately 50-55 words and must end with "space". We chunk them manually and then construct the rules by analyzing the chunking results. Then we use the chart parser to analyze patterns of the verbal units. A series of experiments are conducted to evaluate our approach.

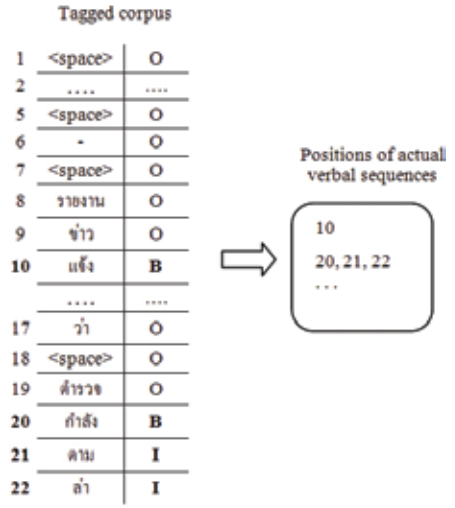

Fig. 8. List of actual verbal sequences extracted from tagged corpus.

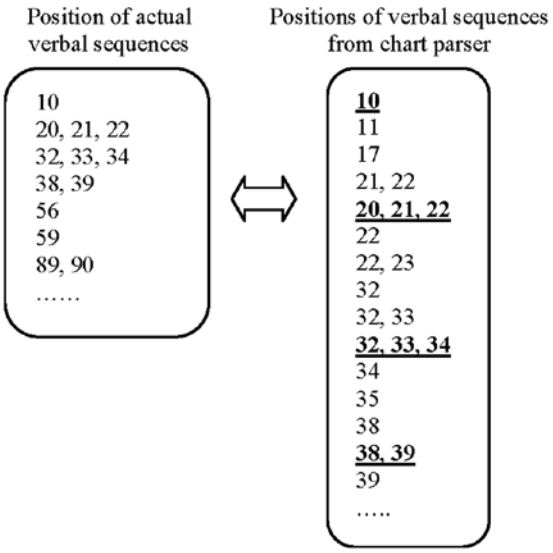

Fig. 9. Comparison of actual verbal sequences with those obtained from chart parser.

4.2 Results

Table 4 shows performance of chunking verbal sequences using a rule-based approach. Precision, recall and F1 shown in the table are measured in the verbal sequence unit level. In the first experiment, we process 101,171 words with 8,110 verbal sequences. We find out 34,118 verbal sequences among these units. All 8,110 verbal units can be detected, i.e., the recall is 100%. Compared to the first experiment, the precision of the second experiment, which applies the longest matching technique, is higher because unnecessary verbal units are eliminated, but the recall becomes lower.

Table 4 Performance of rule-based approach.

	Precision	Recall	F1
Rule-based approach	23.77	100	38.41
Rule-based approach with longest matching strategy	46.79	74.08	57.35

4.3 Discussion and error analysis

With the 605 grammatical rules extracted from the tagged corpus, we can extract actual verbal sequences, i.e., 100% recall. However, precision is quite low because unnecessary verbal sequences are output. To clarify some typical errors, we show some examples of unnecessary verbal sequences detected. From a running text shown in this figure, the chart parser can chunk three verbal sequences from the rules sumed up from the tagged corpus. The first verbal sequence is "เป็น". The second verbal sequence is "ต้องเป็น". The last verbal sequence is "จะต้องเป็น". In fact, the actual verbal sequence is only the last one. This example shows that the first two verbal sequences are the unnecessary ones.

The secondary source of errors is that some verbal sequences appear to be noun phrases. Three examples are illustrated in Fig. 11. The word in rectangle is the wrong verbal sequence. For example, the noun phrase in the example 'โต๊ะแข่งขันสนุกเกอร์" means 'a snooker competing table' but the chart parser indicates that the word "แข่งขัน" (meaning: compete) is a verbal sequence.

Compared to the first experiment, the precision of the second experiment is higher because unnecessary and overlapped verbal sequences are eliminated by the longest matching technique. The longest matching technique decreases the recall rate because some correct verbal sequences do not match with the longest pattern.

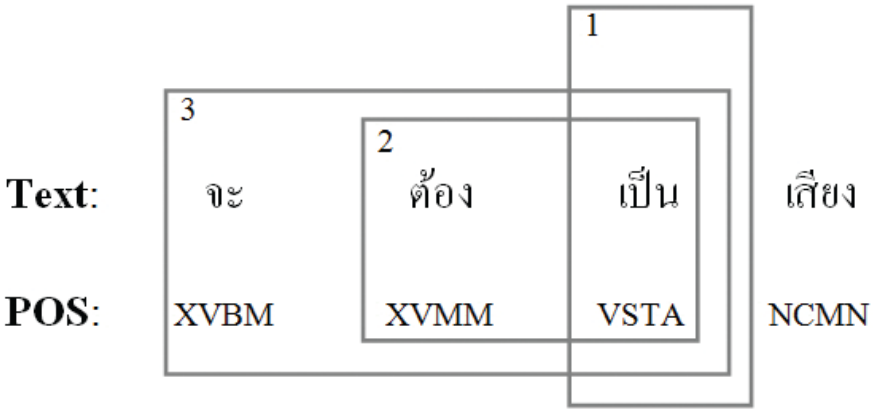

Fig. 10 An example of a running text and three candidate verbal sequences

Noun phrase:	โต๊ะ	แข่งขัน	สนุกเกอร์
POS:	NCMN	VACT	NCMN

Fig. 11. Examples of some verbal sequences that appear to be noun phrases.

5. Hybrid approach

5.1 Methodology

For the baseline method of this approach, we combine the result of the statistical-based approach and the result from the rule-based approach with the longest matching strategy. Before combining, we format the result from the rule-based experiment to IOB format. Before performing the hybrid approach, we conduct a preliminary experiment to find the optimal probability of the CRFs label that is used as the threshold. In the experiment, the probability of the CRFs label is defined as 0.8, 0.85, 0.9, and 0.95, The result of the experiment shows that the probability of 0.9 of the CRFs label is suitable to set as a threshold of the procedure. The following is the procedure in the hybrid process.

1) If the label suggested from CRFs is the same as the label suggested from the rule-based approach.
   • The result label is unambiguous. It is the same for both CRFs and rule-based approach.
2) If the CRFs label is not the same as the grammar label and the probability of the CRFs label is higher than 0.9 (this value is the threshold obtained from the experiment for Thai chunking).
   • The result label is the CRFs label.
3) If the CRFs label is not the same as the grammar label and the probability of CRFs label is lower than 0.9.
   • The result label is the grammar label.

An example of the hybrid procedure is illustrated in Fig. 12.

Then, we evaluate the performance of this experiment by comparing the correct answer of verbal sequence chunk with the combined result. After the combination process in the baseline method, the result shows that some verbal sequences may not start with 'B' (Beginning position). To solve this problem, we propose a method to transform the result. For the first transformation method, we merge 'B's (Beginning position) and 'I' (Intra-position) in the actual verbal sequence into one label 'I' as shown in Fig. 13. Since we will transform all B's to I's, we call this method 'B-to-I' method.

As the second method, we change 'I's (Intra-position) which are not preceded by 'B' (Beginning position) in the result of the hybrid method to 'O' (non-verbal unit position) as displayed in Figure 14. Since all I's without leading 'B' are changed to 'O', we call this transformation 'I-to-O' method. For the third transformation method, we change the first 'I' which is not preceded by 'B' to 'B' as illustrated in Figure 15. Since the first 'I' without leading 'B' is changed to 'B', we call this transformation 'I-to-B' method. Then we evaluate performance based on these three transformation methods.

Word	POS	Answer	CRFs label	CRF Prob.	Rule-based with longest matching	Baseline Method
รับผิดชอบ	VSTA	B	B	0.819515	B	B
ปฏิบัติ	VACT	I	I	0.755801	I	I
และ	JCRG	O	O	0.997928	O	O
ควบคุม	VACT	B	O	0.715398	B	B

Fig. 12. Example of the hybrid procedure.

Actual Answer	Baseline Method
O	O
O	I
O	I
B	B
I	I
I	I
B	B
O	O
B	I

Actual Answer	'B-to-I' method
O	O
O	I
O	I
I	I
I	I
I	I
I	I
O	O
I	I

Fig. 13. 'B-to-I' transformation method (change 'B' and 'I' → 'I').

Baseline Method		'I-to-O' method
O		O
I	→	O
I		O
B		B
I		I
I		I
B	→	B
O		O
I		O

Fig. 14 'I-to-O' transformation method (change 'I'→ 'O' if 'I' is not preceded by 'B')

Baseline Method		'I-to-B' method
O		O
I	→	B
I		I
B		B
I		I
I		I
B	→	B
O		O
I		B

Fig. 15. 'I-to-B' transformation method (change the first 'I' to 'B').

5.2 Results

Table 5 summarizes the result of our four hybrid methods. In the baseline method and the 'I-to-O' method, we obtain the best performance for precision, recall, and F1 which measures verbal sequence level. However, accuracy in the tag level of the 'I-to-O' method is higher than the baseline experiment.

5.3 Discussion and error analysis

Some errors are detected after the baseline method. That is, the result label is assigned to 'O' at the beginning position or the inside position of the verbal sequence. Some examples are displayed in Fig. 16.

As for the first method, precision and recall of this method is lower than the baseline method because this method cannot distinguish the adjoining verbal sequences as illustrated in Fig. 17. In fact, the dash rectangle area contains three verbal sequences. After we change 'B' and 'I' into only 'I', the dash rectangle area contains only one verbal sequence.

In the second method, the performance of this method for precision, recall, and F1-score is the same as the baseline method since the evaluation of the baseline method does not detect a sequence of words, which does not begin with 'B' to be the verbal unit. However, accuracy in the tag level is slightly higher.

As for the third method, recall increases from the baseline method while precision decreases from the baseline method. An example that may trigger increasing a recall is shown in Fig. 18. Fig. 19 shows an example that may cause the precision to decrease. A precision decrease in the case of a part of NP is tagged as 'I' in the baseline method.

Table 5 Performance of the hybrid approach.

	Tag Level	Unit level		
	Accuracy	Precision	Recall	F1
The baseline method	97.86	85.36	91.41	88.28
The 'B-to-I' method	98.12	84.08	90.76	87.30
The 'I-to-O' method	97.89	85.36	91.41	88.28
The 'I-to-B' method	97.90	84.00	92.04	87.84

Word	POS	Answer	CRF Result	CRF Prob.	Rule-based	Baseline Method
เครื่อง	CNIT	O	O	0.998934	O	O
บิน	VSTA	O	O	0.905315	B	O
ขนส่ง	VACT	O	O	0.858003	I	I
ทหาร	NCMN	O	O	0.999817	O	O

Fig. 16. An example of error in the baseline method.

Answer	Baseline Method
O	O
O	B
O	I
B	B
I	I
I	I
B	B
O	O
B	B

Answer	'B-to-I' method
O	O
O	I
O	I
I	I
I	I
I	I
I	I
O	O
I	I

Fig. 17. An example of error in the first method.

Word	POS	Answer	Baseline Method	'I-to-B' method
สถานี	NCMN	O	O	O
บริการ	VACT	O	O	O
ปรับ	VACT	B	I	B
ขึ้น	XVAE	I	I	I

Fig. 18. An example that increases recall in the 'I-to-B' method.

Word	POS	Answer	Baseline Method	'I-to-B' method
การ	FIXN	O	O	O
เรียก	VACT	O	O	O
พบ	VACT	O	I	B

Fig. 19. An example that decreases precision in 'I-to-B' method.

6. Conclusion

In this work, we define three layers of verbal units including verbal sequences, verbal phrases (i.e., verbal chunks, causative forms, and event occurrences), and elementary discourse units (EDUs). As a statistical-based approach, we propose an approach using CRFs on three layers of verbal unit detection including Thai verbal sequences chunking, verb phrases chunking, and Thai EDUs chunking. For the rule-based approach, we use a chart parser to analyze patterns of verbal sequence. Two experiments are conducted in this approach. Moreover, we propose a hybrid approach by combining a statistical-based approach and a rule-based approach to chunk Thai verbal sequences. The results show that the hybrid approach, a combination of a statistical-based and a rule-based method, is the best approach to chunk verbal units.

7. Acknowledgement

This research is financially supported by Thailand Advanced Institute of Science and Technology-Tokyo Institute of Technology (TAIST&Tokyo Tech), National Science and Technology Development Agency (NSTDA) Tokyo Institute of Technology (Tokyo Tech) and Sirindhorn International Institute of Technology (SIIT), Thammasat University (TU). This work is also supported by NECTEC under the project number NT-B-22-KE-38-54-01, and NCRT grant via Thammasat University.

8. References

[1] Atsawintarangkun, P., Ketui, N., Theeramunkong, T. and Haruechaiyasak, C. Analysis of Thai Elementary Discourse Units and Their Detection. In Proceedings of the Sixth International Conference on Knowledge, Information and Creativity Support Systems (KICSS 2011), Beijing, China, 22 - 24, pp. 122 - 129, October 2011.

[2] Thai to English Rule-based Machine Translation. I2R Project Report, Institute for Infocomm Research (I2R) Singapore and National Electronics and Computer Technology Center (NECTEC).

[3] Sameer S. Pradhan, Wayne Ward, Kadri Hacioglu, James H. Martin, and Daniel Jurafsky. Shallow semantic parsing using support vector machines. In Proceedings of the Human Language Technology Conference/North American chapter of the Association for Computational Linguistics annual meeting (HLT/NAACL-2004), May 2004.

[4] Andrew McCallum, D. Freitag and F. Pereira. Maximum entropy Markov models for information extraction and segmentation. In Proceedings of International Conference on Machine Learning, Stanford, pp. 591 - 598, California (2000).

[5] F. Sha and F. Pereira. Shallow parsing with conditional random fields. In Proceedings of Human Language Technology Conference'2003, Edmonton, Canada, pp. 134 - 141, May 27 - June 1, 2003.

[6] Wenliang Chen, Yujie Zhang, and Hitoshi Isahara, "Chinese Chunking based on Conditional Random Fields". NLP 2006, Yokohama, Japan, pp. 149 - 152, March 2006.

[7] Yong-Hun Lee, Mi-Young Kim, and Jong-Hyeok Lee. Chunking Using Conditional Random Fields in Korean Texts. In Lecture Notes in Artificial Intelligence IJCNLP 2005.

[8] H.T. Nguyen, T.P. Nguyen, M.L. Nguyen, and Q.T. Ha. Vietnam Noun Phrase Chunking based on Conditional Random Field. In proceeding of The First International Conference on Knowledge and System Engineering (KSE), Hanoi, Vietnam, 2009.

[9] Park, S.-B. and Zhang, B.-T. Text Chunking by Combining Hand-Crafted Rules and Memory-Based Learning. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pp. 497 - 504, 2003.

[10] Awan, W. and Hussain, S. A hybrid approach to Urdu Verb Phrase chunking. In Proceedings of the 8th Workshop on Asian Language Resources, COLING 2010, Beijing, China (2010).

[11] Sutton, Charles and Mccallum, Andrew. Introduction to Conditional Random Fields for Relational Learning. In Introduction to Statistical Relational Learning MIT Press (2006).

[12] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the 18th international Conference on Machine Learning, Morgan Kaufmann, pp. 282 - 289, San Fransisco, 2001.

[13] CRF++: Yet another crf toolkit, 2005. http://crfpp.sourceforge.net.

[14] Indurkhya, N. and Damerau, F.J. Handbook of Natural Language Processing, Second Edition. Taylor & Francis, 2010.

[15] P. Atsawintarangkun, T. Theeramunkong, C. Haruechaiyasak and T. Kobayashi. Effect of Coarser and Finer POS Categories on Thai VP Chunking. In Proceedings of the International Conference on Information and Communication Technology for Embedded Systems (ICICTES 2012), Bangkok, Thailand, pp. 22 - 24, March 2012.

[16] P. Atsawintarangkun, T. Theeramunkong, C. Haruechaiyasak and T. Kobayashi. Thai Verb Phrase Chunking Based on Conditional Random Fields. In Proceedings of The Joint International Symposium on Natural Language Processing and Agricultural Ontology Service 2011(SNLP-AOS 2011), Bangkok, Thailand, 9 - 10, pp. 40 - 44, February 2012.

[17] Smart Word Analysis for Thai: http://www.cs.cmu.edu/~paisarn/software.html.

[18] Nattapong Tongtep and Thanaruk Theeramunkong. Multi-stage annotation using pattern-based and statistical-based techniques for automatic thai annotated corpus construction. In Proceedings of the 9th Workshop on Asian Language Resources collocated with IJCNLP 2011, pp. 50 - 58, Chiang Mai, Thailand, Asian Federation of Natural Language Processing, November 2011.