A Rule-Based Method for Thai Elementary Discourse Unit Segmentation (TED-Seg)

Nongnuch Ketui, Thanaruk Theeramunkong and Chutamanee Onsuwan

Abstract—Discovering discourse units in Thai, a language without word and sentence boundaries, is not a straightforward task due to its high part-of-speech (POS) ambiguity and serial verb constituents. This paper introduces definitions of Thai elementary discourse units (T-EDUs), grammar rules for T-EDU segmentation and a longest-matching-based chart parser. The T-EDU definitions are used for constructing a set of context free grammar (CFG) rules. As a result, 446 CFG rules are constructed from 1,340 T-EDUs, extracted from the NE- and POS-tagged corpus, Thai-NEST. These T-EDUs are evaluated with two linguists and the kappa score is 0.68. Separately, a two-level evaluation is applied; one is done in an arranged situation where a text is pre-chunked while the other is performed in a normal situation where the original running text is used for test. By specifying one grammar rule per one T-EDU instance, it is possible to make the perfect recall (100%) in a close environment when the testing corpus and the training corpus are the same, but the recall of approximately 36.16% and 31.69% are obtained for the chunked and the running texts, respectively. For an open test with 3-fold cross validation, the recall is around 57.70% and 54.14% for the L2R-LM and M-LM. For the running text in the open test, the f-score improves to 57.70% and 54.14% for the L2R-LM and M-LM.

Index Terms—Thai Elementary Discourse Unit, Discourse unit Segmentation, Chart parser.

1 INTRODUCTION

Most of the works on text processing usually define elementary discourse units (EDUs) as minimal building blocks for forming a discourse tree. Such discourse segmentation has been recognized as an important process in several text processing tasks [1]. In many applications, paragraphs or sentences are too large to use as processing units since they may contain too many pieces of information and need to be broken down into small units. By segmenting a text into a set of tractable discourse units and discovering their relationships, one could construct an abstraction-based summary, which is usually more comprehensive than an extraction-based summary. In this paper, we present definitions of Thai elementary discourse units (T-EDUs), grammar rules for T-EDU segmentation and a longest-matching-based chart parser. With our definition of EDUs in the Thai language, our process includes the NE- and POS-tagged, and T-EDU segmentation. During the first stage, word units and named entities (NEs) are detected in a given text. In the second stage, each NE and word unit will be assigned with suitable POS tag. Finally, the third stage involves construction of all possible parse trees based on grammar rules using chart parsing, and then a process of filtering out implausible ones by a left-to-right longest matching (L2R-LM) [2] and a maximal longest matching (M-LM). A set of 446 context-free grammar rules for T-EDUs are constructed based on the THAI-NEST news corpus [3]. Our method is tested on 1,340 EDU units. Section 2 describes related works on discourse segmentation. Characteristics of the Thai language are given in Section 3. Thai Elementary
Discourse Units (T-EDUs) are defined in Section 4. Section 5 presents Thai grammar rules and a method to construct the parse trees with chart parsing. In Section 6 the experimental evaluation are given with discussion. Finally, a conclusion and future work are given in Section 7.

2 RELATED WORK

In the field of discourse processing, Rhetorical Structure Theory (RST) in [4] has attracted much attention with successful applications for sentence compression and text summarization. Marcu [5] identified discourse markers and elementary textual units in unrestricted (English) text using a surface-based approach. Recently, elementary discourse units (EDUs), minimal building blocks of a discourse tree, have also been proposed for English in the RST framework such as [1]. The RST framework has been adopted for Japanese in [6].

In most languages, detection of sentence boundaries usually presents some difficulty when dealing with spoken language data. In Thai, with no sentence boundary marker, this challenge remains true even for written texts in [7], [8]. Charoenponsawat and Sornlertlamvanich [9] presented a method to disambiguate Thai sentence boundary detection using machine learning. Their experiment was carried out using the Orchid corpus [10]. The results showed that punctuation marks as features were not sufficient for boundary detection in written Thai sentences. In [8], a method using discourse cues and syntactic information was proposed to segment Thai text into elementary discourse units. However, their work was based on a small agriculture corpus of 128 EDUs. While Sinthupoun and Sornil [11] detected their EDUs using a larger corpus of written family law, their results were somewhat limited as law code and usually contains specific words and sentence structures which differ from other types of written documents. In real-world situation, a large pile of news documents is found on World Wide Web. However, up to now there is no research work that shows whether context-free grammar rules (CFG) can help in detecting elementary discourse units in the Thai language.

3 CHARACTERISTICS OF THAI

Thai language shares a similar structure of Subject-Verb-Object (SVO) with many of the world’s languages. However, Thai written texts have neither word nor sentence boundaries. A Thai text can be viewed as a running chunk of text, which expresses an event or a situation. The basic components of a Thai sentence are a subject, a verb, and an object with some modifications. Sentences may include auxiliary units. A sentence may act like a compound sentence with some embedded sentences and conjunctions. A nominal unit comes in the form of noun, noun phrase, and sometimes a word/phrase/sentence that can be represented in the form of noun units. A verbal unit (or predicate) is a verb or verb phrase, which may be accompanied by auxiliary units. In sum, the main challenges in segmenting a Thai text into discourse units may be classified as follows.

- Absence of discourse unit punctuation: Thai written texts contain no explicit word, clause or sentence boundaries.
- Synthetic nominal compounds: Some Thai named entities and noun phrases-verb phrases sequences share the same structure. They consist of a noun followed by a verb, which in the case of synthetic nominal compounds, acts as a modifier.
- Zero anaphoras: It is possible to omit subjects and/or objects in some structures, i.e., zero anaphora [12].
- Embedded clauses: Thai text structures may include many embedded clauses which are relative clauses, often indicating the modifications of nouns.
- Part-of-speech ambiguity: Many of the Thai words (in one form) can perform more than one function. Their functions depend on its position and its expression in unit [12].
- Word-sense ambiguity: Some words (in one form) have the same part-of-speech but may refer to two different meaning.
- Serial verb construction. A verb in Thai may consist of a sequence of concatenated verbs which expresses some meaning and relationship with regards to the main verb [12], [13].
A number of English corpora has been constructed for different purposes. Likewise, sentence segmentation in English has been studied extensively [1]. Large corpora based on Rhetorical Structure Theory (RST) have been developed and the proposed discourse-tagging framework with elementary discourse units (EDUs) has been used. While the works in [5] and [14] use the RST framework for the English language, with several dissimilarities and unique characteristics between English and Thai, we have modified the definitions of elementary discourse units [1]. Table 1 shows the types and their characteristics.

Thai Elementary discourse units (T-EDU) are composed of nominal units, such as subjects or objects and verb units in the form of Subject-Verb-Object (SVO). However, the subject and
the object may be omitted, such as VO, SV or V-only structures. Subjects or objects are noun phrases (NPs) which consist of head nouns, pronouns with numerals and classifiers or nouns with determiners or adjectives (pre/post-modification). A verb unit may have concatenated verbs; called a serial verb which signifies a relative action in order. A closed-class of functional items includes conjunctions, prepositions, speech-act particles, spatial or temporal phrase, and question phrases. Therefore, our analysis of T-EDUs will focus on predicates; T-EDU needs to have one verbal unit or verb phrase as the core verb. Some T-EDUs may include an auxiliary unit (additional information) and may be connected to another T-EDU with a discourse marker.

4 Thai Elementary Discourse Unit (T-EDUs)

In this work, we come up with a method to segment Thai texts into elementary discourse units using two sets of syntactic units, one for Thai EDU detection and the other for Thai Non-EDU detection. Here, six syntactic units for detecting T-EDUs are shown in Table 1.

1) T-EDU-1 (Simple clauses): This type refers to a simple clause, composed of a subject (S), a verb (V) and an optional object (O), respectively. A T-EDU of this type has in a pattern of either SV or SV O, corresponding to a sentence with an intransitive verb or one with a transitive verb, respectively.

2) T-EDU-2 (Subject zero-anaphora clauses): This type refers to a clause with its subject omitted. In general, it is possible for a clause to share its subject with its preceding clause, especially in coordination. Here, ϕ is a zero anaphora (some parts are omitted).

3) T-EDU-3 (Clauses with attribution verb): This type refers to a clause with an attribution verb (VA). Normally it expresses a speech act or a cognitive act. Here, X is the speech act particle and P is a prepositional phrase.

4) T-EDU-4 (Comparative clauses): This type refers to a clause with a special verb (VC) and a comparative clue (R), such as “more (VC) than (R)” and “higher (VC) than (R)”.

5) T-EDU-5 (Question clauses): This type refers to an interrogative sentence, a clause with a question word.

6) T-EDU-6 (Embedded conjunction clauses): This type refers to an embedded conjunction clause, a clause with a conjunction (C), i.e., the conjunction is found embedded inside a clause, rather than in the front. Like conjunctions, its function is to connect two clauses.

Thai Non-EDUs (T-Non-EDU) which are similar to T-EDU but they should not be treated as T-EDUs, can be classified into two types as follows.

1) T-Non-EDU-1: (Clausal subject/objects) Clausal subjects and objects containing verbs should not be counted as T-EDUs e.g., clausal objects of prepositional phrases.

2) T-Non-EDU-2: (Synthetic nominal compounds) Noun phrases cannot be treated as T-EDUs even they have the form of SV, SVO or VO.

5 Thai EDU Grammar and Parser

A Thai sentence, as it is difficult to define, can be viewed as a running chunk of text with no sentence boundaries. We have developed a set of T-EDU grammar rules to transform an input text into T-EDUs. The proposed T-EDU grammar uses the syntactic categories defined by a Thai E-Class [15], where patterns, dictionary and statistic approaches were applied to tag Thai entities and part-of-speech. Punctuation and space are removed from an input text. First, to build an EDU dataset for this work, unit segmentation is manually performed to obtain T-EDUs by making use of the definitions in Table 1. Second, based on the obtained T-EDUs and the POS tags given to the words in these T-EDUs, we straightforwardly construct CFG rules, each of which corresponds to the POS sequence of each obtained T-EDU. However, we slightly generalize the rules by defining some common rules which share frequently occurring sequences. Table 3 shows the
syntactic and meaning of the terminals and non-terminals for rule construction. There are two groups of CFG rules, i.e., shared rules and specific rules.

As for the first group, shared CFG rules are defined to grasp common parts of T-EDUs, mostly detecting T-Non-EDUs. Some shared rules are given in the top part of Fig. 1. Here, corresponding to the first type of T-Non-EDU, npfix contains a sequence of a particle prefix (a prefix that makes a verb become a noun) and a verb or vact (a speech act or a cognitive verb act) unit. As the second type of T-Non-EDU, the nppre is a sequence of a specific noun which can act as a prefix (npfix), followed by ‘verb’ or ‘verb adjective’ or ‘verb adverb’. Note that nppre (T-Non-EDU-2) is similar to T-EDU-1.

For the second group, specific CFG rules are used to define T-EDUs, classified by T-EDU types. Some examples are given in the bottom part of Fig. 1. For example, the first rule is the first type of Thai EDUs (T-EDU-1). It is general with noun/npfix/nppre and verb. T-EDU-2 rule always starts from a verb unit. The attribution verbs which are represented by an attribution verb (vact/vactx), are a specific element of T-EDU-3 rule. T-EDU-4 rule refers to an adverb (vcomp) which follows a main verb, and defines the comparative meaning. The question words (que) are the essential elements which appear in T-EDU-5 rule. T-EDU-6 rules include a conjunction (conje) or a pronoun (prone) which embeds between the other terminals. Conjunction phrases (conjp), embedded phrases (embp), temporal phrases (tempp) and spatial phrase (spatp) are included to the specified CFG rules. It should be noted that those phrases are marker cues that classify possible relations among T-EDUs which are crucial in the task of text summarization (our goal).

After the grammar rules are defined, the parser scans the input string from left-to-right to recognize valid strings called terminals and

![Fig. 1. Some CFG rules classified into shared rules and specific rules. Here, [] means mandatory (select one of them), {} means optional (select zero or one of them) and * represents the repetition.](image)
then uses them to construct the parse trees with their grammar rules. Here, we use a parallel parsing method for analyzing ill-formed inputs under loosely-coupled hardware environment in [16], based on the top-down parsing.

### 6 Experimental Evaluation

This section first describes characteristics of the THAI-NEST corpus, and the NE- and POS-tagged corpus with T-EDU markers. Then, the results which are evaluated with the longest-matching-based chart parser are given.

#### 6.1 Experimental Settings

The THAI-NEST corpus [3] is a set of 10,000 news articles of seven categories; crimes, sports, foreign affairs, politics, entertainment, economics and education. Gathered from 17 on-line news sources such as Daily News, Thairath and Matichon, the texts were manually tagged with name entities (NEs), parts of speech (POS) and semantic roles in the 4W1H format. Using Thai E-Class [15] to segment the data set. However, the output from Thai E-Class has some POS ambiguity problems such as ‘unknown’ tokens and multiple POS tagging. Since our experiment requires a good grammar, we manually reclassify the unknown and multiple POS tokens into one exact POS. Then, as a gold-standard data set, a trained Thai experimenter segments the running Thai text into different T-EDU types.

To ensure that the T-EDU corpus is clear and reliable. Two annotators, who are linguists and do not have any knowledge of our research, segmented the running text according to our definition guidelines into 6 T-EDU types, conjunction phrases, embedded phrases, temporal phrases and spatial phrases. Using the kappa coefficient (\(K\)) [17] measures pairwise agreement among a set of coders making category judgments, correcting for expected chance agreement. Equation 1 shows the calculation of kappa score.

\[
K = \frac{P(A) - P(E)}{1 - P(E)} \tag{1}
\]

where \(P(A)\) is the proportion of times that the coders agree and \(P(E)\) is the proportion of times that we would expect them to agree by chance. In this work, kappa scores are 0.68, showing a good level of agreement.

Table 4 shows characteristics of the data set and its three folds (subsets). We randomly divided a whole data set into three sets (\(C_1\), \(C_2\), \(C_3\), respectively). In the experiments, chunking texts (segmented T-EDUs, \(CC_1\), \(CC_2\), \(CC_3\), \(CC_{123}\)) as well as the running texts (\(CR_1\), \(CR_2\), \(CR_3\), \(CR_{123}\)).
TABLE 6
Precision (P), Recall (R), and F-Score (F) from T-EDU detection. Four experiments are; (1) Close test with the chunking text, (2) Close test with the running text, (3) Open test with the chunking text (3-fold), and (4) Open test with the running text (3-fold).

<table>
<thead>
<tr>
<th>No.</th>
<th>Training Set (Test Set)</th>
<th>All possible matching</th>
<th>L2R-LM</th>
<th>M-LM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>P</td>
<td>R</td>
<td>F</td>
</tr>
<tr>
<td>1</td>
<td>CC1 (CC1)</td>
<td>36.16</td>
<td>100.00</td>
<td>53.11</td>
</tr>
<tr>
<td>2</td>
<td>CC1 (CR2)</td>
<td>31.69</td>
<td>100.00</td>
<td>48.12</td>
</tr>
<tr>
<td>3</td>
<td>CC12 (CC3), CC13 (CC2),</td>
<td>28.90</td>
<td>67.83</td>
<td>40.52</td>
</tr>
<tr>
<td></td>
<td>CC23 (CC1)</td>
<td>25.18</td>
<td>67.83</td>
<td>36.73</td>
</tr>
</tbody>
</table>


CR3, CR123) are used as the input data in the testing. Here, ‘CC1’ is a chunked version of the running text ‘CR1’, ‘CC2’ is a chunked version of ‘CR2’ and ‘CC3’ is that of ‘CR3’. The average number of the tokens (4,063) per T-EDU units (1,340) in the whole data set is 3.03.

Table 5 shows the chunking data used to construct rules. For the whole training data (CC123), 446 CFG rules are constructed and the maximum grammar length is 21 terminals. At least one terminal is found in CFG rules because T-EDU definition focuses on a verb unit. While the average length number of CFG rules is 5.01.

There are two main experiments: close test and open test. The close test uses the whole data set to generate the grammars and test. On the other hand, in the open test, we construct the grammar rules with the training set and use the test set separately to measure the performance of grammar rules.

6.2 Evaluation
Three performance measures are precision, recall and F-score. Equation 2 shows the calculation of F-score value.

\[
F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2}
\]

Here, the precision is defined as the number of T-EDUs correctly segmented by the algorithm divided by the total number of T-EDUs returned from the algorithm and the recall is defined as the number of T-EDU correctly segmented by the algorithm divided by the total number of T-EDUs human manual. We compare the output in these all possible matching, L2R-LM and M-LM. For all possible matching, all possible T-EDUs which are generated by grammar rules are selected and evaluated. The left-to-right longest matching (L2R-LM) is the standard algorithm that filters out implausible T-EDU boundaries [2]. For L2R-LM, its process starts from the left to right sequence and then finds the maximum length of terminals. Here, we develop and propose an improved longest matching algorithm; called the maximal longest matching algorithm (M-LM). Its technique considers firstly the maximum number of T-EDU tokens, filters them out, and then iterates to find the rest of maximum T-EDU tokens until no more token are found.

6.3 Results
Table 6 shows the performances of T-EDU detection by the three techniques. All possible matching, which selects all possible T-EDUs from the system, gets the lowest value. Both of L2R-LM and M-LM reach the perfect performance at 100% for the close test with chunking input. For the close test with a running text, L2R-LM and M-LM can improve the precision from 31.69% to 93.97% and 94.03%, respectively. The recall drops slightly from 100% to 94.18% and 92.91%. The performance of the open-test with running input using L2R-LM and M-LM are 57.70% and 54.14% of F-Score values, respectively. On the other hand, the open-test with chunking input, the performance of M-LM in filtering out T-EDUs is
slightly higher than L2R-LM. This is presumably because it selects firstly the longest T-EDUs from the possible one. Two experiments of the close tests (No.1 and 2) show the highest number of CFG rules which means that the three techniques could use them very effectively to select the possible parse trees.

7 Conclusion and Future Work

In this paper, we present definitions of Thai elementary discourse units (T-EDUs), grammar rules for T-EDU segmentation and a longest-matching-based chart parser. T-EDU grammar rules are constructed from the terminals based on the bottom-up level. Some implausible parse trees are filtered out by grammar rules using chart parsing, and then T-EDU selection is based on the left-to-right longest matching algorithm (L2R-LM) and the maximal longest matching algorithm (M-LM). A set of empirical evaluations is performed on 1,340 T-EDUs from seven categories of news articles. The result shows that the L2R-LM and M-LM can improve the precision to 93.97% and 94.03% for the running text in the close test, respectively. However, the recall drops slightly to 94.18% and 92.91%, respectively. For the running text in the open test, the f-score improves to 57.70% and 54.14% for the L2R-LM and M-LM. We are in the process of expanding the CFG rules and exploring appropriate techniques for selecting T-EDUs. Our future work involves the task of applying statistic approaches to CFG rules for T-EDU segmentation and applying T-EDUs to the task of news summarization.

Acknowledgments

This work was partially supported by the National Research University Project of Thailand Office of Higher Education Commission, as well as the National Electronics and Computer Technology Center (NECTEC) under Project Number NT-B-22-KE-38-54-01.

References


