Thai Rhetorical Structure Tree Construction

SOMNUK SINTHUPOUN\textsuperscript{1} AND OHM SORNIL\textsuperscript{2}

\textsuperscript{1}Maejo University, Thailand
\textsuperscript{2}National Institute of Development Administration, Thailand 10240

ABSTRACT

A rhetorical structure tree (RS tree) is a representation of elementary discourse units (EDUs) and discourse relations among them. An RS tree is very useful to many text processing tasks utilizing relations among EDUs such as text understanding, summarization, and question-answering. Thai language with its distinctive linguistic characteristics requires a unique RS tree construction technique. This article proposes an approach to Thai RS tree construction; it consists of two major steps: EDU segmentation and RS tree construction. Two hidden Markov models constructed from grammatical rules are employed to segment EDUs, and a clustering technique with its similarity measure derived from Thai semantic rules is used to construct a Thai RS tree. The proposed technique is evaluated using three Thai corpora. The results show the Thai RS tree construction effectiveness of 94.90%.

Keywords: Thai Language, Elementary Discourse Unit, Rhetorical Structure Tree.

1 INTRODUCTION

A rhetorical tree (RS tree) is a tree-like representation of elementary discourse units (EDUs) and discourse relations (DRs) among them. It can be defined as: RS tree = (status, DR, promotion, left, right) where status is a set of EDUs; DR is a set of discourse relations; promotion is a subset
of EDUs; and left and right can either be NULL or recursively defined objects of type RS tree [14, 16].

Definition of EDU may vary. Some researchers consider an EDU to be a clause or a clause-like [16] excerpt while others consider them to be a sentence [18] in discourse parsing. A number of techniques are proposed to determine EDU boundaries for English language such as those using discourse cues [1, 6, 15], punctuation marks [6, 16], and syntactic information [16, 18, 19].

Many discourse relations can be used in writings. Some have a single nucleus such as elaboration and condition while others have multiple nuclei such as contrast [13]. A number of techniques for determining relations between EDUs are proposed, such as those using verb semantics [20] to build verb-based events, using cue phrases/discourse markers (e.g., “because”, “however”) [15], and using machine learning techniques [16].


Ito, et.al. [10] construct RS trees by using linguistic clues and rules to identify relation types, i.e., clausal-sequence, conjunction, means and circumstance, and using features of subject and verb in the clauses to predicate adjacent child units of the relations.

For Thai language, Sukvaree, et.al. [21] propose a technique to construct an RS tree by using global and local spanning trees which makes decisions by discourse markers.

This article proposes a new approach to Thai RS Tree construction which consists of two major steps: EDU segmentation and RS tree construction. Two Hidden Markov models constructed from syntactic properties of Thai language are used to segment EDUs, and a clustering technique with its similarity measure derived from semantic properties of Thai language is then used to construct a Thai RS tree.

2 Issues in Thai RS Tree Construction

Thai language has unique characteristics both syntactically and semantically. This makes techniques proposed for other languages not
directly applicable to Thai language. A number of important issues with respect to constructions of Thai RS trees are discussed in this section.

2.1 No Explicit EDU Boundaries

Unlike English, Thai language has no punctuation marks (e.g., comma, full stop, semi-colon, and blank) to determine the boundaries of EDUs. Therefore, EDU segmentation in Thai language becomes a nontrivial issue.

![Thai tree diagram]

EDU1    EDU2                   EDU3
Thai :  [w1 w2 ... wm w(m+1) ... w(n-1) w(n)]
English :  [w1 w2 ... wn], [w(m+1) w(m+2) ... w(n)], [w(n+1) w(n+2) ... w(o)]

Where wi is a word in text.

2.2 EDU Constituent Omissions

Given two EDUs, an absence of subject, object or conjunction in the anaphoric EDU may happen, such as a situation where an anaphoric EDU omits the subject that refers back to the object of the cataphoric EDU. Accordingly, EDU boundaries are ambiguous.

![Thai text example]

Three possibilities:

1) [S(เพื่อน) V(จะขอยืม) O(หนังสือ)]EDU1 [because S(Ф) V(หาซื้อไม่ได้)]EDU2
2) [S(เพื่อน) V(จะขอยืม) O(หนังสือ)]EDU1 [because(Ф) S(Ф) V(หาซื้อไม่ได้)]EDU2
3) [S(เพื่อน) V(จะขอยืม) O(Ф)]EDU1 [because(Ф) S(หนังสือ) V(หาซื้อไม่ได้)]EDU2

In addition, the absence of subject, object or preposition which is a modifier nucleus of VP especially in the anaphoric EDU makes the use of word co-occurrence alone not sufficient to determine the relation between EDU1 and EDU2. For example,

EDU1: ศาลได้มีคำสั่งให้แยกสินสมรส (A court has ordered partition of marriage properties.)
EDU2:  Ф1 จะสั่งยกเลิกการแยก  Ф2 ได้ (Ф1 can cancel the partition of Ф2.)
In the example, EDU2 omits subject “ศาล” (court) and object ‘สินสมรส’ (marriage properties). Therefore, word co-occurrence alone is not sufficient to determine this relation.

2.3 Implicit Markers

The absences of discourse markers in Thai language are often occurred. In the example below, “แต่” (but) is a discourse marker which is omitted, but the relation between EDU1 and EDU2 is still able to determine.

EDU1: ศาลได้มีคำสั่งให้แยกสินสมรส (A court has ordered partition of marriage property.)
EDU2: ภริยาหรือสามีคัดค้าน (a wife or a husband may contest.)

Therefore, considering markers or cue phrases alone is not sufficient to determine the relation between EDUs.

2.4 Adjacent Markers

Given three EDUs with two markers, as shown in the example below, two RS Trees are possible.

EDU1: ศาลได้มีคำสั่งให้แยกสินสมรส (A court has ordered partition of marriage properties.)
EDU2: เด็กภริยาหรือสามีคัดค้าน (but if a wife or a husband contests.)
EDU3: ศาลจะสั่งยกเลิกการแยกได้ (the court can cancel the partition.)

The first possibility, EDU1 and EDU2 relate first by a discourse marker “แต่” (but), next (EDU1, EDU2) and EDU3 relate by a marker “ถ้า” (if). For the other possibility, EDU2 and EDU3 relate first by a marker “ถ้า” (if), next that between (EDU2, EDU3) and EDU1 relate by a marker “แต่” (but).

a) The RS tree with “but” applied first  
b) The RS tree with “if” applied first

Fig. 1. Adjacent markers issue
3 STRUCTURES OF THAI EDUs

A Thai EDU consists of infrastructure and adjunct constituents. The twelve possible arrangements of Thai EDUs [17] are shown in Table 1. The structure of an EDU “A teacher usually doesn’t drink alcohol” is shown in Fig. 2.

![Fig. 2. Structure of the EDU “A teacher usually doesn’t drink alcohol.”](image)

### Table 1: The possible arrangements of Thai EDUs.

<table>
<thead>
<tr>
<th>EDUs</th>
<th>Examples</th>
<th>Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vi</td>
<td>หิว (I’m hungry.)</td>
<td>NP_{x} Vi-NP_{x}</td>
</tr>
<tr>
<td>S-Vi</td>
<td>ฝนตก (It’s rain.)</td>
<td></td>
</tr>
</tbody>
</table>
A noun phrase (NP) is a noun or a pronoun and its expansions which may function as one of the four Thai EDU constituents, namely subject (S), object (O), indirect object (Oi) and nomen (N). The general structure of a noun phrase consists of five constituents which are: head (H), intransitive modifier (Mi), adjunctive modifier (Ma), quantifier (Q), and determinative (D).

A verb phrase (VP) is a verb and its expansions which may function as one of the three Thai EDU constituents, namely intransitive verb (Vi), transitive verb (Vt) and double transitive verb (Vtt). The general structure of a verb phrase consists of four constituents which are: nucleus (Nuc), pre-nuclear auxiliary (Aux1), post-nuclear auxiliary (Aux2), and modifier (M).

There are twenty five possible arrangements of noun phrase and ten arrangements of verb phrases [17], which are shown in Table 2.

4.1 Phrase Identification

To perform phrase identification, word segmentation and part of speech (POS) tagging are performed using SWATH [7] which extracts words and classifies them into 44 types such as common noun (NCMN), active verb (VACT), personal pronoun (PPRS), definite determiner
A hidden Markov model (HMM) [11] employs these POS tag categories to determine phrases. The model assumes that at time step $t$ the system is in a hidden state $PC(t)$ which has a probability $b_{jk}$ of emitting a particular visible state of POS tag $tag(t)$, and a transition probability between hidden states $a_{ij}$:

$$a_{ij} = p(PC_{j}(t+1)|PC_{i}(t)).$$  \hspace{1cm} (1)$$

$$b_{jk} = p(tag_{k}(t)|PC_{j}(t)).$$  \hspace{1cm} (2)$$

where $PC(t)$ is the phrase constituent at time step $t$, and $tag(t)$ is POS tag at time step $t$.

Table 2: The possible arrangements of Thai NPs and VPs.

<table>
<thead>
<tr>
<th>Noun Phrases</th>
<th>Noun Phrases (cont.)</th>
<th>Verb Phrases</th>
</tr>
</thead>
<tbody>
<tr>
<td>H-Ma</td>
<td>H</td>
<td>Nuc</td>
</tr>
<tr>
<td>H-Mi-Ma</td>
<td>H-Mi</td>
<td>Nuc-Aux2</td>
</tr>
<tr>
<td>H-Q-Ma</td>
<td>H-Q</td>
<td>Nuc-M</td>
</tr>
<tr>
<td>H-Ma-Q</td>
<td>H-D</td>
<td>Nuc-Aux2-M</td>
</tr>
<tr>
<td>H-D-Ma</td>
<td>H-Mi-Q</td>
<td>Nuc-M-Aux2</td>
</tr>
<tr>
<td>H-Mi-Q-Ma</td>
<td>H-Q-Mi</td>
<td>Aux1-Nuc</td>
</tr>
<tr>
<td>H-Q-Mi-Ma</td>
<td>H-Mi-D</td>
<td>Aux1-Nuc-Aux2</td>
</tr>
<tr>
<td>H-Mi-D-Ma</td>
<td>H-Q-D</td>
<td>Aux1-Nuc-M</td>
</tr>
<tr>
<td>H-Q-D-Ma</td>
<td>H-D-Q</td>
<td>Aux1-Nuc-Aux2-M</td>
</tr>
<tr>
<td>H-D-Q-Ma</td>
<td>H-Mi-Q-D</td>
<td>Aux1-Nuc-M-Aux2</td>
</tr>
<tr>
<td>H-Mi-Q-D-Ma</td>
<td>H-Mi-D-Q</td>
<td></td>
</tr>
<tr>
<td>H-Mi-D-Q-Ma</td>
<td>H-Q-Mi-D</td>
<td></td>
</tr>
<tr>
<td>H-Q-Mi-D-Ma</td>
<td>H-Q-Mi-D</td>
<td></td>
</tr>
</tbody>
</table>

The probability of a sequence of $T$ hidden states $PC^T = \{PC(1), PC(2), ..., PC(T)\}$ can be written as:

$$p(PC^T) = \prod_{t=1}^{T} p(PC(t) | PC(t-1))$$  \hspace{1cm} (3)$$

The probability that the model produces the corresponding sequence of POS tag $tag^T$, given a sequence of PCs $PC^T$ can be written as:
Then, the probability that the model produces a sequence \( T \) of visible POS tag states is:

\[
p(tag^T | PC^T) = \prod_{t=1}^{T} p(tag(t) | PC(t))
\]

(4)

The Baum-Welch [11] learning algorithm is applied to determine model parameters, i.e., \( a_{ij} \) and \( b_{jk} \), from an ensemble of training samples.

Given a sequence of visible state \( tag^T \), the Viterbi algorithm [11] is used to find the most probable sequence of hidden states by recursively calculating \( p(tag^t | PC^t) \) of visible POS states. Each term \( p(tag(t) | PC(t)) p(PC(t) | PC(t-1)) \) involve only \( tag(t), PC(t), \) and \( PC(t-1) \) by the following definition:

\[
\delta_t(j) = \begin{cases} 
0, & t = 0 \text{ and } j \neq \text{ initial state} \\
1, & t = 0 \text{ and } j = \text{ initial state} \\
\arg \max_i \delta_{t-1}(i) a_{ij} b_{jk}, & \text{otherwise}
\end{cases}
\]

(6)

Figure 3 shows a phrase identification model of string “เพื่อนจะขอยืมหนังสือเล่มนี้ เพราะยังไม่ซื้อ
ดังนั้นจึงต้องยืมหนังสือฉัน” (A friend’s going to borrow this book. Because she (Ф₁) hasn’t been able to buy it (Ф₂). Therefore she (Ф₃) must borrow it from me.) POS tags of the string is “เพื่อน (A friend-NCMN) จะขอ (is going to-XVMM) ยืม (borrow-VACT) หนังสือ (book-NCMN) เล่ม (numerative-CNIT) นี้ (this-DDAC) เพราะ (Because-CONJ) เธอ (she(Ф₁)-PPRS) ไม่ (hasn’t been-NEG) สามารถ (able-to-XVMM) ซื้อ (buy-VACT) มัน (it(Ф₂)) ดังนั้น (Therefore-CONJ) เธอ (she(Ф₁)-PPRS) จึงต้อง (must-XVMM) ยืม (borrow-VACT) หนังสือ (book-NCMN) ฉัน (me-PPRS)”.

The hidden state of a phrase model consists of H(NCMN-book (2/4), -friend (1/4); PPRS-me (1/4)), D(CNIT-numerative (1/2); DDAC-this (1/2)), Discourse-marker(CONJ-because (1/2), -therefore (1/2)), Aux1(XVMM-is going to (1/4), -must (1/4), -able to (1/4); NEG-hasn’t been (1/4)) and Nuc(VACT-borrow (2/3), -buy (1/3)).
4.2 EDU Boundary Determination

After we determine NPs and VPs, another HMM on EDU constituents (shown in Fig. 5.) is then created to determine the boundaries of EDUs. This model can handle the subject and object omission problems, discussed earlier.

Fig. 5 shows an example of the EDU segmentation model for an EDU “เพื่อน จะขอ ยืม หนังสือ เล่ม นี้” (A friend’s going to borrow this book.)

The EDU segmentation model can be expressed as:
where $EDUC(t)$ is EDU constituent at time step $t$, and $tag(t)$ is the phrase tag at time step $t$.

The expression, $p(EDUC(t)|EDUC(t-1))$ is the probability of EDU constituent ($EDUC$) at time $t$ given the previous $EDUC(t-1)$, and $p(tag(t)|EDUC(t))$ is the probability of phrase tag $tag(t)$ given $EDUC(t)$.

Fig. 5. An example of a Thai EDU segmentation model.

\[
p(tag^t) = \max_{EDUC(n)} \prod_{t} p(EDUC(t)|EDUC(t-1))
\]  

(7)

Fig. 6. The results of Viterbi tagging on the Thai EDU segmentation model in Fig. 5.
4.3 EDU Constituent Grouping

Once EDU boundaries are determined, syntactic rules in Table 1 are then applied to group EDU constituents into a larger unit that will be used to match the semantic rules in further steps. For example, a string "เพื่อน-จะขอ-ยืม-หนังสือ-เล่ม-นี้" (A friend’s going to borrow this book.), the result from the Viterbi tagging on the EDU segmentation model is S, Vt, Vt, O, O, O. The matched rule of "NP_O-NP_S-Vt-NP_O" is applied, and the result becomes: "NP_S – (V, V) _O – (NP, NP, NP) _O."

5 THE REFERENCES SECTION

In this section, we describe our proposed technique based on semantic rules derived from Thai linguistic characteristics to construct an RS tree from a corpus. The rules are classified into three types which are Absence, Repetition, and Addition rules [2, 3, 4, 12, 17]. Given a pair of EDUs, an author may write by using any combination of the rules. A similarity measure is calculated from these rules, and a hierarchical clustering algorithm employing this measure is used to construct an RS tree.

5.1 Semantic Rules for EDU Relations

Absence Rules

In Thai language, it has been observed that frequently in writings some constituents of an EDU may be absent while its meaning remains the same. In the example below, the NP (object) “ขนม” (dessert) is absent from the anaphoric EDU, according to rule Ф (O, O).

Cataphoric EDU (Vt-O) : อยากจะทำขนมไหม (Would you like to make a dessert?)
Anaphoric EDU (Vt) : อยากจะทำ (Yes, I do.)

Repetition Rules

It has been observed that frequently an anaphoric EDU relates to its cataphoric EDU by a repetition of NP (subject, object) or a preposition phrase (PP) functioning as a modifier of a nucleus or a verb phrase (VP). In the following example, two EDUs relate by a repetition of an object (NP) “บ้าน” (house), according to the rule ɉ (O, O).
Cataphoric EDU (Vtt-O-I): ผมกำลังจะขายบ้านให้เขา (I’m going to sell him a house.)
Anaphoric EDU (Vt-O): จะขายบ้านหลังไหน (Which house are you going to sell?)

Addition Rules
It has been observed that frequently an anaphoric EDU relates to its cataphoric EDU by an addition of a discourse marker, and possibly accompanied by Absence and/or Repetition rules. In the example below a discourse marker “เพราะ” (because) is added in front of the anaphoric EDU, according to the rule Δ (Marker, Before).

Cataphoric EDU (Vtt-O-I): ฉันอยากจะยืมหนัง (I want to borrow films.)
Anaphoric EDU (Vt-O): เพราะหาซื้อไม่ได้ (because I have not been able to buy it.)

Table 3 lists Repetition, Absence, and Addition rules, for example, я (S, S) means that the subject of the cataphoric EDU is repeated in the anaphoric EDU; Ф(S, S) means that the subject is present in the cataphoric EDU but absent from the anaphoric EDU; and Δ (Marker, Before) means that a discourse marker is added in front of this particular EDU.

5.2 EDU Similarity

Similarity between two EDUs can be calculated from the semantic rules in Table 3, as follows:

5.2.1 Feature Calculations
Given a pair of EDUs, for each rule, an EDU calculates a feature vector which consists of the following elements: Subject, Absence of Subject, Object, Absence of Object, Preposition, Absence of Preposition, Nucleus, Modifier Nucleus, Head, Absence of Head, Modifier Head, Absence of Modifier Head, Marker Before, and Marker After elements. The value of an element is dependent upon the type of rule, as follows:
Table 3: Repetition, Absence, and Addition rules.

<table>
<thead>
<tr>
<th>Repetition (Я)</th>
<th>Absence (Ф)</th>
<th>Addition (Д)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Я (S, S)</td>
<td>Ф (S, S)</td>
<td>Д (Marker, After)</td>
</tr>
<tr>
<td>Я (O, S)</td>
<td>Ф (O, S)</td>
<td>Д (Marker, Before)</td>
</tr>
<tr>
<td>Я (S, O)</td>
<td>Ф (S, O)</td>
<td>Д (Key Phrase, After)</td>
</tr>
<tr>
<td>Я (O, O)</td>
<td>Ф (O, O)</td>
<td>Д (Key Phrase, Before)</td>
</tr>
<tr>
<td>Я (S, Prep)</td>
<td>Ф (Only H, H)</td>
<td></td>
</tr>
<tr>
<td>Я (O, Prep)</td>
<td>Ф ((H, M), H)</td>
<td></td>
</tr>
<tr>
<td>Я (Prep, S)</td>
<td>Ф ((H, M), M)</td>
<td></td>
</tr>
<tr>
<td>Я (Prep, O)</td>
<td>Ф (S, Prep)</td>
<td></td>
</tr>
<tr>
<td>Я ((S, Prep), (S, Prep))</td>
<td>Ф (O, Prep)</td>
<td></td>
</tr>
<tr>
<td>Я ((O, Prep), (S, Prep))</td>
<td>Ф (Prep, S)</td>
<td></td>
</tr>
<tr>
<td>Я ((Prep, Prep), (S, Prep))</td>
<td>Ф (Prep, O)</td>
<td></td>
</tr>
<tr>
<td>Я ((S, Prep), (O, Prep))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Я ((O, Prep), (O, Prep))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Я (Only H, Only H)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Я (H, M)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Я (Only M, Only Nuc)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Я (Only M, Only M)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Я ((Nuc, M), (Nuc, M))</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The following example is used to illustrate calculations related to semantic rules:

EDU1: ชาวบ้าน (Subject) ประกอบ (Nucleus) อุตสาหกรรมในครอบครัว (Object) (The villagers perform the family-industry.)

EDU2: และ (Before) Ф (Absence of Subject) หวงแหน (Nucleus) สมบัติของชาติ (Object) (and protect properties of the nation.)

EDU3: อุตสาหกรรมในครอบครัว (Subject) จึงเป็น (Nucleus) สมบัติของชาติ (Object) (Therefore, the family-industry is a property of the nation.)

To describe the calculations related to semantic rules, the following notations will be used. $C_{Cat}$ is a constituent of the cataphoric EDU, $C_{Ana}$ is a constituent of the anaphoric EDU, $Pos_{Cat}$ is the position of cataphoric EDU, and $Pos_{Ana}$ is the position of anaphoric EDU. $X:Y$ where $X$ can be either Cataphoric or Anaphoric, and $Y$ is an element in the vector of $X$, e.g., Cataphoric:Subject is the Subject element in the vector of the cataphoric EDU. $X:rule$ is an Addition rule applied to $X$ (i.e., a cataphoric or an anaphoric EDU).

Features based on an Absence rules:

Feature vectors of the cataphoric and anaphoric EDUs are filled for a
matched Absence rule, as follows:

\[ \text{if } \Phi(C_{Cat}, C_{Ana}) \text{ is true then} \]

Cataphoric: \( \text{C}_{Cat} = \text{Anaphoric}(\text{Absence of } C_{Ana}) \)

\[ \text{then } \frac{|\text{Pos}_{Cat} - \text{Pos}_{Ana}|}{\text{Total # of sentences}}. \] \hfill (8)

In this example, the properties of EDU1 and EDU2 match with the rule \( \Phi(S, S) \) with the absence of subject “ชาวบ้าน” (villager) in the anaphoric EDU, thus:

Cataphoric: Subject = Anaphoric: Absence of Subject = \( \frac{|1 - 2|}{3} \) \hfill (9)

**Features based on Repetition rules:**

Feature vectors of the cataphoric and anaphoric EDUs is filled for a matched Repetition rule, as follows:

\[ \text{if } \Re(C_{Cat}, C_{Ana}) \text{ is true then} \]

Cataphoric: \( \text{C}_{Cat} = \text{Anaphoric}: C_{Ana} \)

\[ = \frac{|\text{Pos}_{Cat} - \text{Pos}_{Ana}|}{\text{Total # of repeating words}} \cdot \frac{\text{Total # of sentences}}{\text{Total # of words in sentences}} \] \hfill (10)

In the example, the properties of EDU1 and EDU3 match with the rule \( \Re(O, S) \) with a repetition of an object “อุตสาหกรรมในครอบครัว” (family-industries) in the cataphoric EDU as a subject in the anaphoric EDU, thus:

Cataphoric: Object = Anaphoric: Subject = \((1 - \frac{1 - 3}{3}) \cdot (\frac{1}{3} + \frac{1}{3})\) \hfill (11)

**Features based on Addition rules:**

Feature vectors of the cataphoric and anaphoric EDUs is filled for a matched Addition rule, as follows:

\[ \text{if Cataphoric:} \text{D (Marker, After) is true then} \]

Cataphoric: Marker After = Anaphoric: Marker Before = 1

else if Anaphoric: \text{D (Marker, Before) is true then} \text{Anaphoric: Marker Before = Cataphoric: Marker After = 1} \hfill (12)

In this example, the properties of EDU1 and EDU2 match with the rule \( \text{D (Marker, Before) at EDU2, thus:} \)

Anaphoric: Marker Before = Cataphoric: Marker After = 1 \hfill (13)

### 5.2.2 Rule Scoring

After for each rule, the two vectors of the EDU pair are calculated, the vectors are then combined into a rule score which depends on the type of rule and the distance between the two EDUs, as follows:
Absence and Repetition Rules:
These rules consist of two parts (cataphoric and anaphoric). If both parts of an Absence or a Repetition rule are true, then the rule is true. But if a part of an Absence or a Repetition rule is false, then the rule is false, thus:

\[
\text{if } |\text{Pos}_{\text{Cat}} - \text{Pos}_{\text{Ana}}| < \text{MD} \text{ then}
\]

\[
\text{RS}_{\text{Absence}} = [\text{Magnitude of } \text{EDU}_{\text{Cataphoric}} * \text{Magnitude of } \text{EDU}_{\text{Anaphoric}}]
\]

(14)

where \( \text{Pos}_{\text{Cat}} \) and \( \text{Pos}_{\text{Ana}} \) are the positions of cataphoric and anaphoric EDUs, and \( \text{MD} \) is the maximum distance between the EDUs (from experiments \( \text{MD} = 4 \) in this research)

Addition Rules:
In this type of rules, if one part of the rule is true, then the rule is true, thus:

\[
\text{if } |\text{Pos}_{\text{Cat}} - \text{Pos}_{\text{Ana}}| < \text{MD} \text{ then}
\]

\[
\text{RS}_{\text{Addition}} = [\text{Magnitude of } \text{EDU}_{\text{Cataphoric}} + \text{Magnitude of } \text{EDU}_{\text{Anaphoric}}]
\]

(15)

5.2.3 Rule Scoring
Once rule scores are available, similarity between two EDUs (cataphoric and anaphoric) can be calculated as a sum of all the rule scores (each normalized into a range from 0 to 1) according to the CombSum method [8].

6 Rhetorical Tree Construction
A hierarchical clustering algorithm is applied to create an RS tree where each sample (an EDU in this case) begins in a cluster of its own; and while there is more than one cluster left, two closest clusters are combined into a new cluster, and the distance between the newly formed cluster and each other cluster is calculated. Hierarchical clustering algorithms studied in this research are shown in Table 4, and two example RS trees created from two different algorithms are shown in Fig. 7.
Table 4. Hierarchical clustering algorithms studied in this research.

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Distance Between Two Clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Linkage</td>
<td>The smallest distance between a sample in cluster A and a sample in cluster B.</td>
</tr>
<tr>
<td>Unweighted Arithmetic Average</td>
<td>The average distance between a sample in cluster A and a sample in cluster B.</td>
</tr>
<tr>
<td>Neighbor Joining</td>
<td>A sample in cluster A and a sample in cluster B are the nearest. Therefore, define them as neighbors.</td>
</tr>
<tr>
<td>Weighted Arithmetic Average</td>
<td>The weighted average distance between a sample in cluster A and a sample in cluster B.</td>
</tr>
<tr>
<td>Minimum Variance</td>
<td>The increase in the mean squared deviation that would occur if clusters A and B were fused.</td>
</tr>
</tbody>
</table>

7 EXPERIMENTAL EVALUATION

7.1 Rule Scoring

In order to evaluate the effectiveness of the EDU segmentation process, a consensus of five linguists, manually segmenting EDUs of Thai family law, is used. The dataset consists of 10,568 EDUs in total.

The EDU segmentation model is trained with 8,000 random EDUs, and the rest are used to measure performance.

The training continues until the estimated transition probability changes no more than a predetermined value of 0.02, or the accuracy achieves 98%.

The performances of both phrase identification and EDU segmentation are evaluated using recall (Eq. 16) and precision (Eq. 17) measures, which are widely used to measure performance.

\[
\text{Recall} = \frac{\# \text{correct } (\text{phrases or EDUs}) \text{ identified by HMM}}{\# (\text{phrase or EDUs}) \text{ identified by linguists}} \quad (16)
\]

\[
\text{Precision} = \frac{\# \text{correct } (\text{phrases or EDUs}) \text{ identified by HMM}}{\text{total } \# (\text{phrases or EDUs}) \text{ identified by HMM}} \quad (17)
\]

The results show that the proposed method achieves the recall values of 84.8% and 85.3%; and the precision values of 93.5% and 94.2% for phrase identification and EDU segmentation, respectively.
7.2 Evaluation of EDU Constituent Grouping

In order to evaluate the effectiveness of the EDU constituent grouping, three corpuses are used which consist of Absence data (84 EDUs), Repetition data (117 EDUs) and a subset of the Family law with 367 EDUs). The Absence data contains EDUs mostly those following the Absence rules while the Repetition data contains mostly those following the Repetition rules. Five linguists create training and testing data sets by manually grouping EDU constituents.

Table 5 shows the results of grouping EDU constituents (subject (S), object (O), indirect object (I) and nomen (N)) by using rules based on NPs, assuming the positions of verb phrases (Vi, Vt and Vtt) are known. From the results, in general all rules, except NP0-NP5-Vtt-NP1 and NP1-NP5-Vtt-NP0, perform well.

Table 5: Performance of grouping EDU constituents

<table>
<thead>
<tr>
<th>Rules</th>
<th>Absence Data</th>
<th>Repetition Data</th>
<th>Family Law</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP5-Vi-NP5</td>
<td>NP5 (100%)</td>
<td>NP5 (100%)</td>
<td>NP5 (100%)</td>
</tr>
<tr>
<td>NP0-NP5-Vtt-NP0</td>
<td>NP5 &amp; NP0 (100%)</td>
<td>NP5 &amp; NP0 (100%)</td>
<td>NP5 &amp; NP0 (100%)</td>
</tr>
<tr>
<td>NPS-Vtt-NPO-NPI</td>
<td>NP5 &amp; NP0 &amp; NP1 (100%)</td>
<td>NP5 &amp; NP0 &amp; NP1 (100%)</td>
<td>NP5 &amp; NP0 &amp; NP1 (100%)</td>
</tr>
<tr>
<td>NP5-Vt-NP5</td>
<td>NP5 (100%)</td>
<td>NP5 (100%)</td>
<td>NP5 (100%)</td>
</tr>
<tr>
<td>NP5-NP5-Vtt-NP0</td>
<td>NP5 &amp; NP5 (91.37%)</td>
<td>NP5 &amp; NP5 (79.59%)</td>
<td>NP5 &amp; NP5 (90.21%)</td>
</tr>
</tbody>
</table>

To further resolve ambiguities with respect to these two rules, a probability table of terms in positions of NP1 and NP0 following Vtt (P(Vtt| NP1, NP0)) is used. The results of determining functions of EDU constituents by using the rules based on NPs together with the probability table show higher performance for Absence data (92.24%), Repetition data (85.78%), and Family law (93.71%).

7.3 Evaluation of Thai RS Tree Construction

In order to evaluate the effectiveness of the proposed Thai RS tree construction process, linguists manually construct the rhetorical structure trees of three texts used above with a total of 568 EDUs. The
algorithms are evaluated by using recall (Eq. 18) and precision (Eq. 19) measures. Recall and precision are calculated with respect to how close an RS tree constructed from the proposed technique to that created by a consensus of the linguists.

\[
\text{Recall} = \frac{\# \text{correct internal nodes identified by RS Tree}}{\# \text{internal nodes identified by linguists}} \tag{18}
\]

\[
\text{Precision} = \frac{\# \text{correct internal nodes identified by RS Tree}}{\text{Total # of internal nodes identified by RS Tree}} \tag{19}
\]

For the Absence and Repetition data sets, though relations between EDUs follow mostly Absence rules and Repetition rules, respectively, in reality when examined in details, many types of rules are used together in writing. For example,

Anaphoric EDU (S-Vt-O) : บุรุษไปรษณีย์ (S) จะคัดเลือก (Vt) จดหมาย (O)
(A Postman will sort letters)

Cataphoric EDU ((S)-Vt-O) : และ (conj) (Φ S) รับส่ง (Vt) จดหมาย (O)
(And will deliver letters)

Table 6 shows calculations of recall and precision of RS trees created by the Minimum Variance and Unweighted Arithmetic Average algorithms, in Fig. 7.

Table 7 shows the results of evaluating Thai RS Tree construction on the three data sets. The performance on the Family law dataset which combines many kinds of rules in its content is 94.90% recall and 95.21% precision. The results also show that Unweighted Arithmetic Average clustering algorithm gives the best performance for Thai RS Tree construction.

8 CONCLUSIONS

Thai rhetorical structure tree (RST) construction is an important task for many textual analysis applications such as automatic text summarization and question-answering. This article proposes a novel two-step technique to construct Thai RS tree combining machine learning techniques with linguistic properties of the language.
### Table 6: RS tree construction performance of two clustering algorithms

<table>
<thead>
<tr>
<th>The correct RS tree</th>
<th>Minimum Variance</th>
<th>Unweighted Arithmetic Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>3'</td>
<td>3'</td>
<td>3'</td>
</tr>
<tr>
<td>4'</td>
<td>4'</td>
<td>4'</td>
</tr>
<tr>
<td>1'</td>
<td>1'</td>
<td>1'</td>
</tr>
<tr>
<td>9'</td>
<td>9'</td>
<td>6'</td>
</tr>
<tr>
<td>2'</td>
<td>2'</td>
<td>2'</td>
</tr>
<tr>
<td>5'</td>
<td>5'</td>
<td>5'</td>
</tr>
<tr>
<td>6'</td>
<td>6'</td>
<td>6'</td>
</tr>
<tr>
<td>7'</td>
<td>7'</td>
<td>7'</td>
</tr>
<tr>
<td>8'</td>
<td>8'</td>
<td>8'</td>
</tr>
</tbody>
</table>

**Precision = 9/9**

**Recall = 9/9**

### Table 7: Performance of the RS tree construction

<table>
<thead>
<tr>
<th>Data</th>
<th>Num EDUs</th>
<th>Clustering Method</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absence</td>
<td>84</td>
<td>Neighbor Joining</td>
<td>87.23</td>
<td>89.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Single Linkage</td>
<td>82.97</td>
<td>84.78</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Un weighted Arithmetic Average</td>
<td>87.23</td>
<td>89.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minimum Variance</td>
<td>89.40</td>
<td>91.30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Weighted Arithmetic Average</td>
<td>87.23</td>
<td>89.13</td>
</tr>
<tr>
<td>Repetition</td>
<td>117</td>
<td>Neighbor Joining</td>
<td>89.70</td>
<td>91.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Single Linkage</td>
<td>83.82</td>
<td>85.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unweighted Arithmetic Average</td>
<td>89.70</td>
<td>91.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minimum Variance</td>
<td>77.94</td>
<td>79.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Weighted Arithmetic Average</td>
<td>89.70</td>
<td>91.04</td>
</tr>
<tr>
<td>Family-Law</td>
<td>367</td>
<td>Neighbor Joining</td>
<td>85.98</td>
<td>86.26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Single Linkage</td>
<td>64.01</td>
<td>64.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unweighted Arithmetic Average</td>
<td>94.90</td>
<td>95.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minimum Variance</td>
<td>63.37</td>
<td>63.57</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Weighted Arithmetic Average</td>
<td>90.44</td>
<td>90.73</td>
</tr>
</tbody>
</table>

**Precision = 6/10**

**Recall = 6/9**
Fig. 7. RS trees from two hierarchical clustering algorithms
First, phrases are determined and then are used to segment elementary discourse units (EDUs). The phrase segmentation model is a hidden Markov model constructed from the possible arrangements of Thai phrases based on part-of-speech of words, and the EDU segmentation model is another hidden Markov model constructed from the possible phrase-level arrangements of Thai EDUs. Linguistic rules are applied after EDU segmentation to group related constituents into a large unit. Experiments show the EDU segmentation effectiveness of 85.3% and 94.2% in recall and precision, respectively.

A hierarchical clustering algorithm whose similarity measure derived from semantic rules of Thai language is then used to construct an RS tree. The technique is experimentally evaluated, and the effectiveness achieved is 94.90% and 95.21% in recall and precision, respectively.

REFERENCES

8. Fox, E. A. and Shaw, J. A. Combination of multiple searches. In the second Text Retrieval conference (TREC-2), Gaithersburg, MD, USA,


17. Panupong, V.: Inter-Sentence Relations in Modern Conversational Thai. The Siam Society, Bangkok (1970)


THAI RHETORICAL STRUCTURE TREE CONSTRUCTION

SOMNUK SINTHUPOUN
DEPARTMENT OF COMPUTER SCIENCE,
MAEJO UNIVERSITY, CHIANGMAI, THAILAND 50290
E-MAIL: <somnuk@mj.ac.th>

OHM SORNIL2
DEPARTMENT OF COMPUTER SCIENCE,
NATIONAL INSTITUTE OF DEVELOPMENT ADMINISTRATION
BANGKOK, THAILAND 10240
E-MAIL: <osornil@as.nida.ac.th>