THAI RHETORICAL STRUCTURE ANALYSIS

Somnuk Sinthupoun  
Department of Computer Science  
Maejo University  
Chiangmai, Thailand 50290  
somnuk@mju.ac.th

Ohm Sornil  
Department of Computer Science  
National Institute of Development Administration  
Bangkok, Thailand 10240  
osornil@as.nida.ac.th

Abstract—Rhetorical structure analysis (RSA) explores discourse relations among elementary discourse units (EDUs) in a text. It is very useful in many text processing tasks employing relationships among EDUs such as text understanding, summarization, and question-answering. Thai language with its distinctive linguistic characteristics requires a unique technique. This article proposes an approach for Thai rhetorical structure analysis. First, EDUs are segmented by two hidden Markov models derived from syntactic rules. A rhetorical structure tree is constructed from a clustering technique with its similarity measure derived from Thai semantic rules. Then, a decision tree whose features derived from the semantic rules is used to determine discourse relations.

Keywords- Thai Language, Rhetorical Structure Analysis, Elementary Discourse Unit, Rhetorical Structure Tree, Discourse Relation.

I. INTRODUCTION

Rhetorical structure analysis (RSA) studies relations among elementary discourse units (EDUs). It provides a framework for analyses of text and is very useful to many text processing tasks employing relationships among EDUs such as text understanding, summarization, and question-answering. Definition of EDU may vary. Some researchers consider an EDU to be a clause or a clause-like [6] excerpt while others consider them to be a sentence [14] in discourse parsing. A number of techniques are proposed to determine EDU boundaries for English language such as those using discourse cues [5, 12, 13], punctuation marks [6, 13], and syntactic information [6, 14, 15].

EDUs and their discourse relations (DRs) are commonly represented as a rhetorical structure tree (RS tree). It can be defined as follows: 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 [4, 6].

Many discourse relations can be used in writings. Some have a single nucleus such as elaboration and condition while others have multiple nucleuses such as contrast [25].

Marcu, et al. [7] determine discourse relations using Naive Bayes classifiers to learn all adjacent sentence pairs that contain the cue phrase (i.e. "but", "however") at the beginning of the second sentence, in the middle of a sentence, and at the end of the sentence. Pitler, et al. [9] determine local discourse relations using an N-gram model to compute transitional probabilities in both directions for each pair of EDUs. To account for remaining ambiguities, a unigram model based on previous known relations is used to predict the next one. Pitler, et al. [10] determine implicit discourse relations using naive Bayes, maximum entropy, and AdaBoost classifiers whose features include polarity tag, inquirer tag, verb classes, First-Last, First3, Modality, context and lexical features based on Penn Discourse Treebank [18].

For Thai language, Sukvaree, et al. [21] construct RS trees by using global and local spanning trees which determine relations by using DR marker tags. Wattanamethanont, et al. [15] purpose a technique to determine relations by using naive Bayes classifier whose features consist of DR marker, key phrase, and word co-occurrences.

This article proposes a new approach to Thai RSA which consists of three major steps: EDU segmentation, RS tree construction, and DR determination. Two hidden Markov models constructed from syntactic properties of Thai language are used in segmenting EDUs, a clustering technique with its similarity measure derived from semantic properties of Thai language is used to construct an RS tree, and a decision tree is used to determine the relation between two related EDUs in the RS tree.

II. ISSUES IN THAI RHETORICAL STRUCTURE ANALYSIS

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 Thai RSA are discussed in this section.

A. 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.
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.

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: ศาลได้มีคำสั่งให้แบ่งสมบูรณ์ (A court has ordered partition of marriage properties.)
EDU2: ผู้ชายหรือแม่หม้ายกล่าว (but contrast, a wife or a husband may contest.)

On the other hand,

EDU1: ศาลได้มีคำสั่งให้แบ่งสมบูรณ์ (A court has ordered partition of marriage properties.)
EDU2: แต่จะต้องยื่นคำร้องในเรื่อง (but elaborates only what the wife and the husband agree.)

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

D. Adjacent Markers

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

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

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

![Fig. 1. Adjacent markers issue](image)

E. Marker Ambiguities

One marker may infer multiple relations such as “but” (when) can infer condition or cause-result relation, and “if” (but) can infer “contrast” or “elaboration” relation whose example can be seen below:

EDU1: ศาลได้มีคำสั่งให้แบ่งสมบูรณ์ (A court has ordered partition of marriage properties.)
EDU2: แต่จะต้องยื่นคำร้องในเรื่อง (but contrast, a wife or a husband may contest.)

On the other hand,

EDU1: ศาลได้มีคำสั่งให้แบ่งสมบูรณ์ (A court has ordered partition of marriage properties.)
EDU2: แต่จะต้องยื่นคำร้องในเรื่อง (but elaborates only what the wife and the husband agree.)

---

*Fig. 2. Structure of the EDU “A teacher usually doesn’t drink alcohol.”*
III. STRUCTURES OF THAI EDUS

A Thai EDU consists of infrastructure and adjunct constituents. The twelve possible arrangements of Thai EDUs [22] are shown in Table 1. The structure of an EDU “A teacher usually doesn’t drink alcohol” is shown in Fig. 2.

Table 1: The possible arrangements of Thai EDUs.

| EDUs   | Examples                                      | Rules         |
|--------|----------------------------------------------|---------------|
| Vi     | ติ่ง (I’m hungry.)                           | NP_Vi          |
| S-Vi   | อุ่นๆ (It’s rainy.)                         |               |
| Vi-S   | คุณๆ สิ่งๆ (Are you pain?)                 |               |
| Vi-0   | มันๆ (It’s rain.)                           |               |
| S-Vt-O | ตื่นๆ (The car hit the boy.)                |               |
| O-S-Vt | ฉันๆ หน่อยๆ (I’ve already seen this photograph.) |               |
| Vt-O-I | ตื่นๆ (I’m hungry.)                         | NP_Vt          |
| S-Vt-O-I| ตื่นๆ (I’ve already seen this photograph.) |               |
| O-S-Vt-I| ฉันๆ หน่อยๆ (I’ve already seen this photograph.) |       |
| I-S-Vt-O| ฉันๆ หน่อยๆ (I’ve already seen this photograph.) | |
| N     | ตื่นๆ (Auntie)                             | NP_N           |
| N-N   | ตื่นๆ (Whose pen is this?)                |               |

IV. EDU SEGMENTATION

This section describes the EDU segmentation technique proposed in this research. To reduce the segmentation ambiguities caused from omissions of words or discourse markers, and the appearances of modifiers, noun phrases and verb phrases which are constituents of EDUs are first determined, according to the syntactic properties of Thai language. These phrases are then used to identify boundaries of EDUs.

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 (Oi) and nomen (N). The general structure of a noun phrase consists of five constituents which are: head (H), intransitive (Oi) and nomen (N). The general structure of a noun phrase consists of five constituents which are: head (H), intransitive (Oi) and nomen (N). The general structure of a noun phrase consists of five constituents which are: head (H), intransitive (Oi) and nomen (N).

There are twenty five possible arrangements of noun phrase and ten arrangements of verb phrases [22], which are shown in Table 2.

A. Phrase Identification

To perform phrase identification, word segmentation and part of speech (POS) tagging are performed using SWATH [20] which extracts words and classifies them into 44 types such as common noun (NCMN), active verb (VACT), personal pronoun (PPRS), definite determiner (DDAC), unit classifier (CNIU) and negate (NEG). A hidden Markov model (HMM) [18] 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 of emitting a particular visible state of POS tag tag(t), and a transition probability between hidden states aij:

\[ a_{ij} = p(PC(t+1)|PC(t)). \]  \hspace{1cm} (1)

\[ b_{jk} = p(tag(t)|PC(t)). \]  \hspace{1cm} (2)

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

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:

\[ p(tag^T | PC^T) = \prod_{t=1}^{T} p(tag(t)|PC(t)). \] \hspace{1cm} (4)

Then, the probability that the model produces a sequence \( tag^T \) of visible POS tag states is:

| Noun Phrases | Noun Phrases (cont.) | Verb Phrases |
|--------------|----------------------|-------------|
| H-Ma         | H                    | Nuc         |
| H-Mi-Ma      | H-Mi                 | Nuc-Aux2    |
| H-Q-Ma       | H-Q                  | Nuc-M       |
| H-Ma-Q       | H-D                  | Nuc-Aux2-M  |
| H-D-Ma       | H-Mi-Q               | Nuc-M-Aux2  |
| H-Mi-Q-Ma    | H-Q-Mi               | Aux1-Nuc    |
| H-Q-Mi-Ma    | H-Mi-D               | Aux1-Nuc-Aux2 |
| H-Mi-D-Ma    | H-Q-D                | Aux1-Nuc-M  |
| H-Q-D-Ma     | H-D-Q                | Aux1-Nuc-Aux2-M |
| H-D-Q-Ma     | H-Mi-Q-D             | Aux1-Nuc-Aux2-M |
| H-Mi-Q-D-Ma  | H-Mi-D-Q             | Aux1-Nuc-Aux2-M |
| H-Mi-D-Q-Ma  | H-Mi-D-Q             | Aux1-Nuc-Aux2-M |
| H-Mi-D-Q-Ma  | H-Mi-D-Q             | Aux1-Nuc-Aux2-M |
\[ p(\text{tag}_T) = \arg \max_{PC(t)} \prod_{t=1}^{T} p(\text{tag}(t) | PC(t)) p(PC(t) | PC(t-1)) \]  

(5)

The Baum-Welch [18] learning algorithm is applied to determine model parameters, i.e., \( a_{ij} \) and \( b_{kj} \), from an ensemble of training samples.

Given a sequence of visible state \( tag_T \), the Viterbi algorithm [18] is used to find the most probable sequence of hidden states by recursively calculating

\[ p(\text{tag}(t)|PC(t)) \]

Each term \( \delta \) is used to find the most probable sequence of hidden state contributions to the arg is for index \( k \) which matches the visible state \( \text{tag}(t) \).

| S | O | I | Marker | Nuc | AuxI | End |
|---|---|---|--------|-----|------|-----|
| Start | 1 | 0 | 0 | 0 | 0 | 0 |
| H | 0 | 0 | 0 | 1 | 0 | 0 |
| D | 0 | 0 | 0 | 0 | 0 | 0 |
| Marker | 0 | 0 | 0 | 0 | 0 | 0 |
| AuxI | 0 | 0 | 0 | 0 | 0 | 0 |
| Nuc | 0 | 0 | 0 | 0 | 0 | 0 |
| End | 0 | 0 | 0 | 0 | 0 | 0 |

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

(6)

where \( b_{kj} \) represents the transition probability \( b_{kj} \) selected by the visible state emitted at time \( t \). Thus, the only nonzero contribution to the arg is for index \( k \) which matches the visible state \( \text{tag}(t) \).

\[ p(\text{tag}_T) = \arg \max_{EDUC(t)} \prod_{t=1}^{T} p(\text{tag}(t) | EDUC(t)) p(EDUC(t) | EDUC(t-1)) \]  

(7)

where \( EDUC(t) \) is EDU constituent at time step \( t \), and \( \text{tag}(t) \) 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 \( p(\text{tag}(t)|EDUC(t)) \) is the probability of phrase tag \( \text{tag}(t) \) given \( EDUC(t) \).

Figure 3 shows a phrase identification model of string “A friend’s going to borrow this book. Because she (friend-NCMN) hasn’t been able to (buy-VACT) it (book-NCMN) yet (NCMM).” (A friend’s going to borrow this book. Because she (friend-NCMN) hasn’t been able to (buy-VACT) it (book-VACT) yet (NCMM)).

The hidden state of a phrase model consists of H(NCMN-numerative-(1/2); PPRS-me-(1/2)), Discourse-marker(CONJ-because-(1/2)), Aux1(XVMM-is going to (1/2); NUMM-has been (1/2); DISC-this (1/2)), and Conjuction(CONJ-because-(1/2); NUMM-has been (1/2); DISC-this (1/2)).

B. 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. 4 shows an example of the EDU segmentation model for an EDU “friend-NCMM) book-NCMM).” (A friend’s going to borrow this book.)

The EDU segmentation model can be expressed as:

\[ p(\text{tag}_T) = \arg \max_{EDUC(t)} \prod_{t=1}^{T} p(\text{tag}(t) | EDUC(t)) p(EDUC(t) | EDUC(t-1)) \]  

where \( EDUC(t) \) is EDU constituent at time step \( t \), and \( \text{tag}(t) \) 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 \( p(\text{tag}(t)|EDUC(t)) \) is the probability of phrase tag \( \text{tag}(t) \) given \( EDUC(t) \).

Fig. 5 shows an example of the EDU segmentation model for an EDU “A friend’s going to borrow this book. Because she (friend-NCMN) hasn’t been able to (buy-VACT) it (book-NCMN).” (A friend’s going to borrow this book.)

The hidden state of a phrase model consists of H(NCMN-numerative-(1/2); PPRS-me-(1/2)), Discourse-marker(CONJ-because-(1/2)), Aux1(XVMM-is going to (1/2); NUMM-has been (1/2); DISC-this (1/2)), and Conjuction(CONJ-because-(1/2); NUMM-has been (1/2); DISC-this (1/2)).

The hidden state of a phrase model consists of H(NCMN) book-NCMM).” (A friend’s going to borrow this book.)

B. 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 “friend-NCMM) book-NCMM).” (A friend’s going to borrow this book.)

The EDU segmentation model can be expressed as:

\[ p(\text{tag}_T) = \arg \max_{EDUC(t)} \prod_{t=1}^{T} p(\text{tag}(t) | EDUC(t)) p(EDUC(t) | EDUC(t-1)) \]  

where \( EDUC(t) \) is EDU constituent at time step \( t \), and \( \text{tag}(t) \) 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 \( p(\text{tag}(t)|EDUC(t)) \) is the probability of phrase tag \( \text{tag}(t) \) given \( EDUC(t) \).
C. 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 "ต้องการจะซื้อตู้เย็นเล็กน้อย" (A friend’s going to borrow this book.), the result from the Viterbi result becomes: "NP of a nucleus or a verb phrase (VP). In the following example, object) or a preposition phrase (PP) functioning as a modifier (house), according to the rule я (O, O).

V. THAI RHETORICAL TREE CONSTRUCTION

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 [1, 3, 22, 23, 24]. 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.

A. Semantic Rules for EDU Relations

1) 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) “ของ” (dessert) is absent from the anaphoric EDU, according to the rule я (O, O).

Cataphoric EDU (Vt-O-I) : ฉันต้องต้องซื้อตู้เย็น(I’m going to sell him a house.)
Anaphoric EDU (Vt-O) : จะซื้อตู้เย็น(Which house are you going to sell?)

2) 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) “ของ” (house), according to the rule я (O, O).

Cataphoric EDU (Vt-O-I) : ฉันต้องต้องซื้อตู้เย็น(I want to borrow films.)
Anaphoric EDU (Vt-O) : ฉันต้องซื้อตู้เย็น(I have not been able to buy it.)
The following example is used to illustrate calculations related to semantic rules:

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

EDU2: และ (Before) ยา (Absence of Subject) พระทัย (Nucleus) สมบัติของเจ้า (Object) (and protect properties of the nation.)

EDU3: ดุสิตกรานเป็นกิจกรรมของ (Subject) จีนิน (Nucleus) สมบัติของเจ้า (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).

a) Features based on Absence rules:
Feature vectors of the cataphoric and anaphoric EDUs are filled for a matched Absence rule, as follows:

\[
If \Phi(C_{Cat}, C_{Ana}) \text{ is true then} \\
Cataphoric \space C_{Cat} = \text{Anaphoric}(Absence \space C_{Ana}) = 1 - \frac{|Pos_{Cat} - Pos_{Ana}|}{\text{Total # of sentences}} \tag{8}
\]

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:

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

b) Features based on Repetition rules:
Feature vectors of the cataphoric and anaphoric EDUs are filled for a matched Repetition rule, as follows:

\[
If \Psi(C_{Cat}, C_{Ana}) \text{ is true then} \\
Cataphoric \space C_{Cat} = \text{Anaphoric} \space C_{Ana} = \frac{|Pos_{Cat} - Pos_{Ana}|}{\text{Total # of sentences}} \times \frac{\text{Total # of repeating words}}{\text{Total # of words in sentences}} \tag{10}
\]

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

\[ Cataphoric:Obj ect = \text{Anaphoric:Subject} = (1 - \frac{1}{3}) \times \frac{3}{3} = 1 \times \frac{1}{3} \tag{11} \]

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:

\( a) \) 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:

\[
\text{if } |Pos_{Cat} - Pos_{Ana}| < MD \text{ then} \\
\text{RS}_{Absence} = \frac{\text{Magnitude of EDU}_{Cataphoric} \times \text{Magnitude of EDU}_{Anaphoric}}{\text{Magnitude of EDU}_{Cataphoric} + \text{Magnitude of EDU}_{Anaphoric}} \tag{14}
\]

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

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

\[
\text{if } |Pos_{Cat} - Pos_{Ana}| < MD \text{ then} \\
\text{RS}_{Addition} = \frac{\text{Magnitude of EDU}_{Cataphoric} + \text{Magnitude of EDU}_{Anaphoric}}{\text{Magnitude of EDU}_{Cataphoric} + \text{Magnitude of EDU}_{Anaphoric}} \tag{15}
\]

3) Similarity Calculation
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].

\[ (IJC\text{S}\text{IS}) \text{ International Journal of Computer Science and Information Security,} \]
C. 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. 8.

Table 4. Hierarchical clustering algorithms studied in this research.

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

VI. DISCOURSE RELATION DETERMINATION

In this section, we describe our technique to determine relations based on features according to semantic rules in Table 3. A decision tree (C5.0 algorithm) employs these features to determine a relation.

A. Feature Extraction

A feature score for discourse relation determination is calculated from contents of the EDUs, based on the three types of rules. The feature set consists of two subsets. The first subset is for the cataphoric EDU which consists of: Subject, Object, Preposition, Nucleus, Marker Before and Marker After. The other subset is for the anaphoric EDU which consists of: 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. The value of each element is dependent upon the type of rules matched (multiple matching is allowed), as follows:

1) Features based on Absence rules:
Feature values are filled as follows:

If \( \Theta(C_{\text{cat}}, C_{\text{ana}}) \) is true then
- Cataphoric: \( C_{\text{cat}} = 1 \)
- Anaphoric: Absence of \( C_{\text{ana}} = 1 \)

In the example, considering EDU1 and EDU2 yields:

\[
\text{Cataphoric: Subject} = 1 \\
\text{Anaphoric: Absence of Subject} = 1
\]  

2) Features based on repetition rules:
Feature values are filled as follows:

If \( \cap (C_{\text{cat}}, C_{\text{ana}}) \) is true then
- Cataphoric: \( C_{\text{cat}} = 1 \)
- Anaphoric: \( C_{\text{ana}} = 1 \)

In the example, considering EDU1 and EDU3 yields:

\[
\text{Cataphoric: Object} = 1 \\
\text{Anaphoric: Subject} = 1
\]  

3) Features based on addition rules:
Feature values are filled as follows:

If Cataphoric: \( \sqcup \) (Marker, After) or Anaphoric: \( \sqcup \) (Marker, Before) is true then
- Cataphoric: Marker After = Anaphoric: Marker Before = Marker

In the example, considering EDU1 and EDU2 yields:

\[
\text{Cataphoric: Marker After} = \text{Anaphoric: Marker Before} = \text{same (and)}
\]

However, if one side of the pair is a relation, only addition and repetition rules are considered.

Fig. 7. Discourse relations in a rhetorical structure tree.

VII. EXPERIMENTAL EVALUATION

A. Evaluation of Thai EDU Segmentation

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. 21) and precision (Eq. 22) measures, which are widely used to measure performance.

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

\[
\text{Precision} = \frac{\# \text{correct (phrases or EDUs) identified by HMM}}{\# \text{(phrase or EDUs) identified by HMM}}
\]
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.

B. Evaluation of EDU Constituent Grouping

In order to evaluate the effectiveness of the EDU constituent grouping, three corupses 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 (Vt, Vt and Vt) are known. From the results, in general all rules, except NP0-NP3-Vt-NP1 and NP5-NP5-Vt-NP0, perform well.

Table 5: Performance of grouping EDU constituents

| Rules          | Absence Data | Repetition Data | Family Law |
|----------------|--------------|-----------------|------------|
| NPv-NPv-NPv-NPv | NPv (100%)   | NPv & NPv (100%)| NPv (100%) |
| NPv-NPv-NPv-NPv | NPv & NPv (100%) | NPv & NPv (100%) | NPv & NPv (100%) |
| N-Pv-NPv-NPv   | NPv, NPv & NPv (100%) | NPv & NPv & NPv (100%) | NPv & NPv & NPv (100%) |
| N-Pv-NPv-NPv   | NPv (100%), NPv & NPv (91.37%) | NPv (100%), NPO & NPv (79.59%) | NPv (100%), NPO & NPv (90.21%) |
| N-N            | NPv (100%)   | NPv (100%)      | NPv (100%) |

To further resolve ambiguities with respect to these two rules, a probability table of terms in positions of NP and NP following Vt (P(Vt|NP, NP)) 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%).

C. 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. 23) and precision (Eq. 24) 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.

Recall = \frac{\text{#correct internal nodes identified by HMM}}{\text{#correct internal nodes identified by linguists}} \quad (23)

Precision = \frac{\text{#correct internal nodes identified by RS Tree}}{\text{#internal nodes identified by RS Tree}} \quad (24)

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) : (S) (And will deliver letters)

Cataphoric EDU ((S)-Vt-O) : (S) (A Postman will sort letters)

Table 6 shows calculations of recall and precision of RS trees created by the Minimum Variance and Unweighted Arithmetic Average algorithms, in Fig. 8.

Table 6: RS tree construction performance of two clustering algorithms

| The correct RS tree | Minimum Variance | Unweighted Arithmetic Average |
|---------------------|------------------|-----------------------------|
| 3'                  | 3'               | 3'                          |
| 4'                  | 4'               | 4'                          |
| 5'                  | 5'               | 5'                          |
| 6'                  | 6'               | 6'                          |
| 7'                  | 7'               | 7'                          |
| 8'                  | 8'               | 8'                          |
| 9'                  | 9'               | 9'                          |
| 10'                 | 10'              | 10'                         |

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.
In order to evaluate the effectiveness of the Thai DR determination, linguists manually tag a relation to each internal node of RS trees constructed from Family Law, with a total of 624 EDU/relation pairs. A C5.0 decision tree algorithm [11] is trained with 424 random pairs, and the rest are used to measure performance. Ten discourse relations are studied in this research. Since markers are found helpful in determining relations, the test set is divided into EDU/relation pairs with markers and those without markers. The performance is reported accordingly.

Table 8 shows the results of determining ten discourse relations. The performance of EDU pairs from the Family law with and without marker is 85.09% and 82.81%, respectively.

According to a sensitivity analysis, Marker ranks at the top for determining discourse relations. This also shows in the results where the accuracy of determining relations for EDU/relation pairs with markers is higher than for those without markers.

VIII. CONCLUSIONS

Rhetorical structure analysis explores relations among elementary discourse units (EDUs) in a text. It is very useful for many textual analysis applications such as automatic text summarization and question-answering.
This article proposes a novel technique to analyze rhetorical structure of Thai texts which combines machine learning techniques with linguistic properties of Thai language. Relations among EDUs are expressed hierarchically as a rhetorical structure tree.

First, phrases are determined and then are used to segment 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 a hidden Markov model constructed from the possible phrase-level arrangements of Thai EDUs. Linguistic rules are applied after the EDU segmentation to group related constituents into large units. 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.

Once an RS tree is constructed, a decision tree algorithm whose features derived from the semantic rules is used to determine discourse relations between EDUs in the tree. The technique is experimentally evaluated, and the overall effectiveness is at 82.81%.

REFERENCES

[1] D. Chamniraksant, “Clauses in the Thai Language”. Unpublished master’s thesis, Chulalongkorn University, Thailand, 1969.
[2] D. K. Harman, editor. “The second Text Retrieval conference (TREC-2)”, Gaithersburg, MD, USA, March 1994. U.S. Government Printing Office, Washington D.C, 1994.
[3] D. Mahathana, “A study of sentence groups in Thai essays”. Unpublished master’s thesis, Chulalongkorn University, Thailand, 1980.
[4] D. Marcu, “Build Up Rhetorical Structure Trees”. In American Association for Artificial Intelligence, 1996.
[5] D. Marcu, “A decision-based approach to rhetorical parsing”. In the 37th Annual Meeting of the Association for Computational Linguistics, ACL, Maryland, pp. 365-372 , 1999.
[6] D. Marcu, “The theory and Practice of Discourse Parsing and Summarization”. The MIT Press, Cambridge, MA, 2000.
[7] D. Marcu, and A. Echihabi, “An Unsupervised Approach to Recognizing Discourse Relations”. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 368-375, 2002.
[8] E. A. Fox, and J. A. Shaw, “Combination of multiple search”. In The second Text Retrieval conference (TREC-2), Gaithersburg, MD, USA, March 1994. U.S. Government Printing Office, Washington D.C, pages 243-249, 1994.
[9] E. Piter, M. Raghupathy, H. Mehta, A. Nenkova, A. Lee and A. Joshi, “Easily Identifiable Discourse Relations”. Department of Computer and Information Science Technical Reports, No. MS-CIS-08-24, University of pennsylvania, 2008.
[10] E. Piter, A. Louis and A. Nenkova, “Automatic sense prediction for implicit discourse relation in text”. In Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 683–691,Suntec, Singapore, 2-7 August 2009.
[11] J.R. Quinlan, Data mining tools See5 and C5.0, http://www.rulequest.com/see5-info.html, February 3, 2003.
[12] J. Chairoensuk, and A. Kawtrakul, “Thai Elementary Discourse Unit Segmentation by Discourse Segmentation Cues and Syntactic Information”. In the Sixth Symposium on Natural Language Processing 2005 (SNLP 2005), Chiang Rai, Thailand, December 13-15, 2005.
[13] L. Alonso, and I. Castellón, Towards a delimitation of discursive segment for Natural Language Processing applications. International Workshop on Semantics, Pragmatics and Rhetorics, San Sebastián 22-24 November, 2001.
[14] L. Polanyi, “A formal model of the structure of discourse”. Journal of Pragmatics, 12, 601-638, 1988.
[15] M. Wattanamethanont, T. Sukvaree, and A. Kultrakul, “Discourse relation recognition by using Naïve Bayesian classifier”. The 9th National Computer Science and Engineering Conference (NCSEC 2005), University of the Thai Chamber of Commerce (UTCC), Bangkok, Thailand, October 27-28, 2005.
[16] R. Sorcut, and D. Marcu, “Sentence Level Discourse Parsing using Syntactic and Lexical Information”. In Proceedings of the 2003 Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAAACL), May 27-June 1, Edmonton, Canada, 2003.
[17] R. Subba , B. Di Eugenio, S. N. K. “Learning FOL rules based on rich verb semantic representations to automatically label rhetorical relations (EACL)”. In Workshop on learning Structured Information in Natural Language Applications , 2006.
[18] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and W. Webber. “The penn discourse treebank 2.0”. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC), 2008.
[19] S. Levinson, R. Rabiner, and M. Sondhi, “An introduction to the application of the theory of probabilistic function of a Markov processes to automatic speech recognition”. Bell System Technical Journal, 62:1035-1074, 1983.
[20] T. Charoenpovorn, V. Sorntlertamvanich, H. Isahara, “Building A Large Thai Text Corpus—Part-Of-Speech Tagged Corpus: ORCHID—”. In proceedings of the Natural Language Processing Pacific Rim symposium, 1976.
[21] T. Sukvaree, J. Charoensuk, M. Wattanamethanont, and A. Kultrakul, “RST based Text Summarization with Ontology Driven in Agriculture Domain”. Department of Computer Engineering, Kasetsart University, Bangkok, Thailand, 2004.
[22] Y. Panupong, Inter-Sentence Relations in Modern Conversational Thai. The Stam Society, Bangkok, 1970.
[23] W. Aroonmanakun, “Referent Resolution for Zero Pronouns in Thai”. Southeast Asian Linguistic Studies in Honour of Vichin Panupong. (Abramson, Arthur S., ed.). pp. 11-24. Chulalongkorn University Press, Bangkok. ISBN 974-636-995-4, 1997.
[24] W. Aroonmanakun, “Zero Pronoun Resolution in Thai: A Centering Approach”. In Burnham, Denis, Interdisciplinary Approaches to Language Processing: The International Conference on Human and Machine Processing on Human and Machine Processing of Language and Speech. NECTEC. Bangkok, 127-147, 2000.
[25] W. C. Mann, and S. A. Thompson, “Rhetorical structure theory. Toward a functional theory of text organization”. Text, 8(3): 243-281, 1988.

AUTHORS PROFILE

Somnuk Sinthupoun is an instructor at the Department of Computer Science, Maejo University, Thailand. He holds an M.S. in Computer Science from National Institute of Development Administration (NIDA,), and a B.S. in Computer Science from Maejo University. His main research interests include artificial intelligence, information retrieval, data mining, and related areas.

Ohm Sornil is an Assistant Professor at the Department of Computer Science, National Institute of Development Administration, Thailand. He holds a Ph.D. in Computer Science from Virginia Polytechnic Institute and State University, an M.S. in Computer Science from Syracuse University, an M.B.A. in Finance from Mahidol University, and a B.Eng. in Electrical Engineering from Kasetsart University. His main research interests include computer and network security, artificial intelligence, information retrieval, data mining, and related areas.