EXPLOITING RHETORICAL RELATIONS TO MULTIPLE DOCUMENTS TEXT SUMMARIZATION

N. Adilah Hanin Zahri[1] , Fumiyo Fukumoto[2], Matsyoshi Suguru[2] and Ong Bi Lynn[1]

[1]School of Computer and Communication,
University of Malaysia Perlis, Perlis, Malaysia
[2]Interdisplinary Graduate School of Medicine and Engineering,
University of Yamanashi, Yamanashi, Japan

ABSTRACT

Many of previous research have proven that the usage of rhetorical relations is capable to enhance many applications such as text summarization, question answering and natural language generation. This work proposes an approach that expands the benefit of rhetorical relations to address redundancy problem for cluster-based text summarization of multiple documents. We exploited rhetorical relations exist between sentences to group similar sentences into multiple clusters to identify themes of common information. The candidate summary were extracted from these clusters. Then, cluster-based text summarization is performed using Conditional Markov Random Walk Model to measure the saliency scores of the candidate summary. We evaluated our method by measuring the cohesion and separation of the clusters constructed by exploiting rhetorical relations and ROUGE score of generated summaries. The experimental result shows that our method performed well which shows promising potential of applying rhetorical relation in text clustering which benefits text summarization of multiple documents.

KEYWORDS

Rhetorical Relations, Text Clustering, Extractive Text Summarization, Support Vector Machine, Probability Model, Markov Random Walk Model

1.INTRODUCTION

The study on rhetorical relations between sentences has been introduced to analyze, understand, and generate natural human-languages. Rhetorical relations hold sentences or phrases in a coherent discourse and indicate the informative relations regarding an event i.e. something that occurs at a specific place and time associated with some specific actions. Rhetorical relations are defined according to the objective expression the writer intends to achieve by presenting two text spans. There are several structures have been developed to describe the semantic relations between words, phrases and sentences such as Rhetorical Structure Theory (RST) [1], RST Treebank [2], Lexicalized Tree-Adjoining Grammar based discourse [3], Cross-document Structure Theory (CST) [4][5] and Discourse GraphBank [6]. Each structure defines different kind of relations to distinguish how events in text are related by identifying the transition point of a relation from one text span to another. In general, rhetorical relations is defined by the effect of the relations, and also by different constrains that must be satisfied in order to achieve this effect, and these are specified using a mixture of propositional and intentional language. For instance, in RST structure, the Motivation relation specifies that one of the spans presents an action to be

DOI : 10.5121/ijnsa.2015.7201

performed by the reader; the *Evidence* relation indicates an event (claim), which describes the information to increase the reader's belief of why the event occurred [2]. Rhetorical relations also describe the reference to the propositional content of spans and which span is more central to the writer's purposes.

The interpretation of how the phrases, clauses, and texts are semantically related to each other described by rhetorical relations is crucial to retrieve important information from text spans. Previous works have proven that these kind of coherent structures have benefit text summarization [7][8][9][10][11][12]. Text summarization is a process of automatically creating a summary that retains only the relevant information of the original document. Generating summary includes identifying the most important pieces of information from the document, omitting irrelevant information and minimizing details. Automatic document summarization has become an important research area in natural language processing (NLP), due to the accelerating rate of data growth on the Internet. Text summarization limits the need for user to access the original documents and improves the efficiency of the information search. The task becomes tougher to accomplish as the system also has to deal with multi-document phenomena, such as paraphrasing and overlaps, caused by repeated similar information in the document sets.

In general, rhetorical relations are used to produce optimum ordering of sentences in a document and remove redundancy from generated summaries.

Our work focused on different aspect of utilizing rhetorical relations to enhanced text summarization. In our study, we discovered that rhetorical relations not only describes how two sentences are semantically connected, but also shows the similarity pattern between two sentences. For instance, CST suggests that two text span connected as *Paraphrase* is offering same information, and on the other hand, two text span connected as *Overlap* is having partial similar information, as shown in Example 1 and Example 2 which adopted from CST structure:

Example 1: Paraphrase

S_1 *Smokes billows from the Pirelli building.*
S_2 *Smoke rises from the Milan skyscraper.*

Example 2: Overlap

S_3 *The plane put a hole in the 25th floor of the Pirelli building, and smoke was seen pouring from the opening.*
S_4 *The plane crashed into 25th floor of the Pirelli building in downtown Milan.*

Figure 1 and 2 exhibit the illustration of both *Paraphrase* and *Overlap* using set theory diagram.

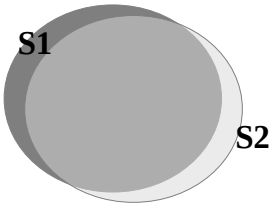

Figure 1. Similarity pattern of *Paraphrase*, where S1≈S2

Figure 2. Similarity pattern of Overlap, where $s \subset S3$ and $s \subset S4$

Figure 1 and 2 show that the similarity patterns between two sentences can be extracted from rhetorical relations which can be exploited during construction of similar text clusters to identify theme of common information in multiple documents for text summarization. Our objective is to improve the retrieval of candidate summary from clusters of similar texts and utilize the rhetorical relations to eliminate redundancy during summary generation.

We first examined and investigated the definition of rhetorical relations from existed structure and then redefined the rhetorical relations between sentences which will be useful for text summarization. We then perform an automated identification of rhetorical relations among sentences from the documents using machine learning technique, SVMs. We examined the surface features, *i.e.* the lexical and syntactic features of the text spans to identify characteristics of each rhetorical relation and provide them to SVMs for learning and classification module. We extended our work to the application of rhetorical relations in cluster-based text summarization. The next section provides an overview of the existing techniques. Section 3 describes the methodology of our system and finally, we report experimental result with some discussion.

2. PREVIOUS WORK

The coherent structure of rhetorical relations has been widely used to enhance the summary generation of multiple documents [13][14][15]. For instance, a paradigm of multi-document analysis, CST has been proposed as a basis approach to deal with multi-document phenomenon, such as redundancy and overlapping information during summary generation [8][9][10][11][12]. Many of CST based works proposed multi-document summarization guided by user preferences, such as summary length, type of information and chronological ordering of facts. One of the CST-based text summarization approaches is the incorporation of CST relations with MEAD summarizer [8]. This method proposes the enhancement of text summarization by replacing low-salience sentences with sentences that have maximum numbers of CST relationship in the final summary. They also observed the effect of different CST relationships against summary extraction. The most recent work is a deep knowledge approach system, CST-based SUMMarizer or known as CSTSumm [11]. Using CST-analyzed document, the system ranks input sentences according to the number of CST relations exist between sentences. Then, the content selection is performed according to the user preferences, and a multi-document summary is produced CSTSumm shows a great capability of producing informative summaries since the system deals better with multi-document phenomena, such as redundancy and contradiction. Most of the CST-based works observed the effects of individual CST relationships to the summary generation, and focuses on the user preference based summarization. Most of the corpus used in the previous works was manually annotated for CST relationships. In other words, this technique requires deep linguistic knowledge and manually annotated corpus by human.

On the other hand, cluster-based approaches have been proposed to generate summary with wide diversity of each topic discussed in a multiple document. A cluster-based summarization groups

the similar textual units into multiple clusters to identify themes of common information and candidates summary are extracted from these clusters [16][17][18]. Centroid based summarization method groups the sentences closest to the centroid in to a single cluster [9][19]. Since the centroid based summarization approach ranks sentences based on their similarity to the same centroid, the similar sentences often ranked closely to each other causing redundancy in final summary. In accordance to this problem, MMR [20] is proposed to remove redundancies and re-rank the sentences ordering. In contrast, the multi-cluster summarization approach divides the input set of text documents in to a number of clusters (sub-topics or themes) and representative of each cluster is selected to overcome redundancy issue [30]. Another work proposed a sentences-clustering algorithm, *SimFinder* [21][22] clusters sentences into several cluster referred as themes. The sentence clustering is performed according to linguistic features trained using a statistical decision [23]. Some work observed time order and text order during summary generation [24]. Other work focused on how clustering algorithm and representative object selection from clusters affects the multi-document summarization performance [25]. The main issue raised in multi-cluster summarization is that the topic themes are usually not equally important. Thus, the sentences in an important theme cluster are considered more salient than the sentences in a trivial theme cluster. In accordance to this issue, previous work suggested two models, which are Cluster-based Conditional Markov Random Walk Model (Cluster-based CMRW) and Cluster-based HITS Model [26]. The Markov Random Walk Model (MRWM) has been successfully used for multi-document summarization by making use of the "voting" between sentences in the documents [27][28][29]. Differ with former model, Cluster-based CMRW incorporates the cluster-level information into the link graph, meanwhile Cluster-based HITS Model considers the clusters and sentences as hubs and authorities [26].

3. FRAMEWORK

3.1. Redefinition of Rhetorical Relations

Our main objective is to exploit rhetorical relations in order to build clusters of similar text that will enhance text summarization. Therefore, in this work, we make used the existing coherent structure of rhetorical relations. Since that previous works proposed various structure and definition of rhetorical relations, the structure that defines rhetorical relations between two text spans is mostly appropriate to achieve our objective. Therefore, we adopted the definition of rhetorical relation by CST [5] and examined them in order to select the relevant rhetorical relations for text summarization. According to the definition by CST, some of the relationship presents similar surface characteristics. Relations such as *Paraphrase*, *Modality* and *Attribution* share similar characteristic of information content with *Identity* except for the different version of event description. Consider the following examples:

Example 3

S_5 *Airbus has built more than 1,000 single-aisle 320-family planes.*

S_6 *It has built more than 1,000 single-aisle 320-family planes.*

Example 4

S_7 *Ali Ahmedi, a spokesman for Gulf Air, said there was no indication the pilot was planning an emergency landing.*

S_8 *But Ali Ahmedi said there was no indication the pilot was anticipating an emergency landing.*

Example 3 and 4 demonstrate an example of sentences pair that can be categorized as *Identity*, *Paraphrase*, *Modality* and *Attribution* relations. The similarity of lexical and information in each sentences pair is high, therefore these relations can be concluded as presenting the similar relation. We also discovered similarity between *Elaboration* and *Follow-up* relations defined by CST. Consider the following example:

Example 5

S_9 *The crash put a hole in the 25th floor of the Pirelli building, and smoke was seen pouring from the opening.*

S_{10} *A small plane crashed into the 25th floor of a skyscraper in downtown Milan today.*

Example 5 shows that both sentences can be categorized as *Elaboration* and *Follow-up*, where S_9 describes additional information since event in S_{10} occurred. Another example of rhetorical relations that share similar pattern is *Subsumption* and *Elaboration*, as shown in Example 6 and Example 7, respectively.

Example 6

S_{11} *Police were trying to keep people away, and many ambulances were at the scene.*

S_{12} *Police and ambulance were at the scene.*

Example 7

S_{13} *The building houses government offices and is next to the city's central train station.*

S_{14} *The building houses the regional government offices, authorities said.*

S_{11} contains additional information of S_{12} in Example 6, hence describes that sentences pair connected as *Subsumption* can also be defined as *Elaboration*. However, the sentences pair belongs to *Elaboration* in Example 7 cannot be defined as *Subsumption*. The definition of *Subsumption* denotes the second sentence as the subset of the first sentence, however, in *Elaboration*, the second sentence is not necessary a subset of the first sentence. Therefore, we keep *Subsumption* and *Elaboration* as two different relations so that we can precisely perform the automated identification of both relations.

We redefined the definition of the rhetorical relations adopted from CST, and combined the relations that resemble each other which have been suggested in our previous work [30]. *Fulfillment* relation refers to sentence pair which asserts the occurrence of predicted event, where overlapped information present in both sentences. Therefore, we considered *Fulfillment* and *Overlap* as one type of relation. As for *Change of Perspective*, *Contradiction* and *Reader Profile*, these relations generally refer to sentence pairs presenting different information regarding the same subject. Thus, we simply merged these relations as one group. We also combined *Description* and *Historical Background*, as both type of relations provide description (historical or present) of an event. We combined similar relations as one type and redefine these combined relations. Rhetorical relations and their taxonomy used in this work is concluded in Table 1.

Table 1. Type and definition of rhetorical relations adopted from CST.

Relations by CST	Proposed Relations	Definition of Proposed Relation
Identity, Paraphrase, Modality, Attribution	Identity	Two text spans have the same information content
Subsumption, Indirect Speech, Citation	Subsumption	S_1 contains all information in S_2, plus other additional information not in S_2
Elaboration, Follow-up	Elaboration	S_1 elaborates or provide more information given generally in S_2.
Overlap, Fullfillment	Overlap	S_1 provides facts X and Y while S_2 provides facts X and Z; X, Y, and Z should all be non-trivial
Change of Perspective, Contradiction, Reader Profile	Change of Topics	S_1 and S_2 provide different facts about the same entity.
Description, Historical Background	Description	S_1 gives historical context or describes an entity mentioned in S_2.
-	No Relations	No relation exits between S_1 and S_2.

By definition, although *Change of Topics* and *Description* does not accommodate the purpose of text clustering, we still included these relations for evaluation. We also added *No Relation* to the type of relations used in this work. We combined the 18 types of relations by CST into 7 types, which we assumed that it is enough to evaluate the potential of rhetorical relation in cluster-based text summarization.

3.2. Identification of Rhetorical Relations

The type of relations exist among sentences from multiple documents are identified by using a machine learning approach, Support Vector Machine (SVMs) [31]. This technique is adopted from our previous work [30], where we used CST-annotated sentences pair obtained from CST Bank[1] [5] as training data for the SVMs. Each data is classified into one of two classes, where we defined the value of the features to be 0 or 1. Features with more than 2 value will be normalized into [0,1] range. This value will be represented by 10 dimensional space of a 2 value vector, where the value will be divided into 10 value range of [0.0,0.1], [0.1,0.2], …, [0.9,1.0]. For example, if the feature of text span S_j is 0.45, the surface features vector will be set into 0001000000. We extracted 2 types of surface characteristic from both sentences, which are lexical similarity between sentences and the sentence properties. Although the similarity of information between sentences can be determined only with lexical similarity, we also included sentences properties as features to emphasis which sentences provide richer and specific information, *e.g.* location and time of the event. We provided these surface characteristics to SVMs for learning and classification of the text span S_1 according to the given text span S_2

3.2.1 Lexical Similarity between Sentences

More than one similarity measurements is used to measure the amount of overlapping information among sentences. Each measurement computes similarity between sentences from different aspects.

1. Cosine Similarity

[1]http://tangra.si.umich.edu/clair/CSTBank/phase1.htm

Cosine similarity measurement is defined as follows:

$$\cos(S_1, S_2) = \frac{\sum_i (s_{1,i} \times s_{2,i})}{\sqrt{\sum_i (s_{1,i})^2} \times \sqrt{\sum_i (s_{2,i})^2}}$$

where S_1 and S_2 represents the frequency vector of the sentence pair, S_1 and S_2, respectively. The cosine similarity metric measures the correlation between the two sentences according to frequency vector of words in both sentences. We observed the similarity of word contents, verb tokens, adjective tokens and bigram words from each sentences pair. The cosine similarity of bigram s is measured to determine the similarity of word sequence in sentences. The words ordering indirectly determine the semantic meaning in sentences.

2. Overlap ratio of words from S_1 in S_2 , and vice versa

The overlap ratio is measured to identify whether all the words in S_2 are also appear in S_1, and vice versa. This measurement will determine how much the sentences match with each other. For instance, given the sentences pair with relations of *Subsumption*, the ratio of words from S_2 appear in S_1 will be higher than the ratio of words from S_1 appear in S_s. We add this measurement because cosine similarity does not extract this characteristic from sentences. The overlap ratio is measured as follows:

$$WOL(S_1) = \frac{\#commonwords(S_1, S_2)}{words(S_1)}$$

where "*#commonword*" and "*#words*" represent the number of matching words and the number of words in a sentence, respectively. The feature with higher overlap ratio is set to 1, and 0 for lower value. We measured the overlap ratio against both S_1 and S_2.

3. Longest Common Substring

Longest Common Substring metric retrieves the maximum length of matching word sequence against S_1, given two text span, S_1 and S_2, .

$$LCS(S_1) = \frac{len(MaxComSubstring(S_1, S_2))}{length(S_1)}$$

The metric value shows if both sentences are using the same phrase or term, which will benefit the identification of *Overlap* or *Subsumption*.

4. Ratio overlap of grammatical relationship for S_1

We used a broad-coverage parser of English language, MINIPAR [32] to parse S_1 and S_2, and extract the grammatical relationship between words in the text span. Here we extracted the number of *surface subject* and the *subject of verb* (*subject*) and *object of verbs*(*object*). We then compared the grammatical relationship in S_1 which occur in S_2, compute as follows:

$$SubjOve(S_1) = \frac{\#commonSubj1(S_1, S_2)}{Subj1(S_1)}$$

$$ObjOve(S_1) = \frac{\#commonObj1(S_1, S_2)}{Obj1(S_1)}$$

The ratio value describes whether S_2 provides information regarding the same entity of S_1 , *i.e. Change of Topics*. We also compared the *subject* in S_1 with *noun* of S_2 to examine if S_1 is discussing topics about S_2.

$$SubjNounOve(S_1) = \frac{\#commonSubj(S_1)Noun(S_2)}{Obj(S_1)}$$

The ratio value will show if S_1 is describing information regarding subject mention in S_2, *i.e. Description.*

3.2.2 Sentences Properties

The type of information described in two text spans is also crucial to classify the type of discourse relation. Thus, we extracted the following information as additional features for each relation.

1. Number of entities

Sentences describing an event often offer information such as the place where the event occurs (location), the party involves (person, organization or subject), or when the event takes place (time and date). The occurrences of such entities can indicate how informative the sentence can be, thus can enhance the classification of relation between sentences. Therefore, we derived these entities from sentences, and compared the number of entities between them. We used Information Stanford NER (CRF Classifier: 2012 Version) of Named Entity Recognizer [46] to label sequence of words indicating 7 types of entities (*PERSON*, *ORGANIZATION*, *LOCATION*, *TIME*, *DATE*, *MONEY* and *PERCENT*).

The Stanford NER generally retrieves proper nouns from corresponding sentences and categorize into one of the mentioned class, as shown in the following example:

On Jan./DATE *5*/DATE, *a 15-year-old boy crashed a stolen plane into a building in Tampa* /LOCATION, *Florida*/LOCATION.

As Stanford NER only recognizes proper nouns, the common noun such as "*boy*"" in the context is not labeled as PERSON. Thus, in order to harvest maximum information from a text span, we make use of the lexical units obtained from lexical database, FrameNet [33]. We extracted lexical unit from FrameNet which matches the 7 class defined by Stanford NER class. The manual lexical unit extraction is carried out by 2 human judges. Table 2 shows the example of frames used in the experiment. We used data from FrameNet to retrieve the unidentified type of information from common noun in sentences. We hereafter refer to the information retrieved here and by Stanford NER as sentences entity. We computed the number of sentences entities appearing in both S_1 and S_2. Based on the study of training data from CSTBank[1] [5], there are no significant examples of annotated sentences indicates which entity points to any particular

discourse relation. Therefore, in the experiment, we only observed the number of sentences entities in both text spans. The features with higher number of entities are set to 1, and 0 for lower value.

Table 2. Information adopted from FrameNet

NER Class	FrameNet	
	No. Frames	Example of Frames
PERSON	12	People (*e.g.* person, lady, boy, man, woman) People by vocation (*e.g.* police officer, journalist) Behind the scene (*e.g.* film-maker, director, producer) Kinship (*e.g.* father, mother, sister, brother) Leadership (*e.g.* captain, chairman, president, chief) Origin (*e.g.* European, Dutch, American, Chinese) People by residence (*e.g.* roommate, neighbour, housemate)
ORGANIZATION	6	Business (*e.g.* company, corporation, firm) Organization (*e.g.* government, agency, committee) Military (*e.g.* army, naval, military, navy)
LOCATION	12	Building (*e.g.* pyramid, airport, terminal, house) Locale by event (*e.g.* theatre, battlefield, venue) Locale by ownership (*e.g.* land, estate, property) Locale by use (*e.g.* museum, gallery, college, headquarters) Part Orientational (*e.g.* west, east, north) Political Locale (*e.g.* village, municipality, city)
TIME	2	Calenderic unit (*e.g.* morning, evening, noon) Location in time (*e.g.* time)
DATE	2	Calenderic unit (*e.g.* winter, spring, summer) Natural fatures (*e.g.* spring, fall)
MONEY	1	Money (*e.g.* money, cash, funds)
PERCENT	0	-

2. Number of conjunctions

We observed the occurrence of 40 types of conjunctions. We measured the number of conjunctions appear in both S_1 and S_2, and compare which sentence contains more conjunctions. We assumed that the higher the number of conjunctions, the more information is provided in the corresponding text span. The comparison of the number of conjunctions will help to determine relation *i.e. Elaboration*.

Table 3. List of conjunctions

because	since	now that	as	in order that
so	so that	why	although	though
even though	whereas	while	but	if
unless	whether or not	even if	in case	after
and	before	but	for	nor
once	only if	until	when	whenever
where	wherever	yet	or	either or
neither nor	whether or	not only	but also	both and

3. Lengths of sentences

   We define the length of S_j by the number of word occurs in the corresponding text span, and compare the length of both sentences. The length of both text spans will show whether both text span are *Identity*, where the length will be the same, or one of the text spans presents more information than another, where S_1 will be longer, *i.e. Subsumption*.
   We defined the length of S_j as follows:

$$Length(S_j) = \sum_i w_i$$

   where *w* is the word appearing in the corresponding text span.

4. Type of Speech

   We determined the type of speech, whether the text span, S_1 cites another sentence by detecting the occurrence of quotation marks to identify *Citation* or *Indirect Speech* which are the sub-category of *Identity*.

3.3. Rhetorical Relation-based Text Clustering

The aim of this work is to expand the benefits of rhetorical relations between sentences to cluster-based text summarization. Rhetorical relation between sentences not only indicates how two sentences are connected to each other, but also shows the similarity patterns in both sentences. Therefore, by exploiting these characteristics, our idea is to construct similar text clustering based on rhetorical relations among sentences. We consider that the following relations are most appropriate for this task:

(i) *Identity*
(ii) *Subsumption*
(iii) *Elaboration*
(iv) *Overlap*

These relations indicates either equivalence or partial overlapping information between text spans, as shown in Table 1. Connections between two sentences can be represented by multiple rhetorical relations. For instance, in some cases, sentences defined as *Subsumption* can also be define as *Identity*. Applying the same process against the same sentence pairs will be redundant. Therefore to reduce redundancy, we assigned the strongest relation to represent each connection between 2 sentences according to the following order:

(i) whether both sentences are identical or not
(ii) whether one sentence includes another

(iii) whether both sentences share partial information
(iv) whether both sentences share the same subject of topic
(v) whether one sentence discusses any entity mentioned in another

The priority of the rhetorical relations assignment can be concluded as follows:

Identity > Subsumption > Elaboration > Overlap

We then performed clustering algorithm to construct groups of similar sentences. The algorithm is summarized as follows:

i) Rhetorical relations identified by SVMs is assign to between two sentences. For sentences pair which is assigned with multiple relations, the strongest relations is assigned as stated in the above (refer to Figure 3(a)).

ii) Suppose each sentence is a centroid of its own cluster. Sentences connected to the centroid as *Identity* (*ID)*, *Subsumption* (*SUB*), *Elaboration* (*ELA*) and *Overlap* (*OVE*) relations is identified and sentences with these connections are evaluated as having similar content, and aggregated as one cluster (refer Figure 3(b)).

iii) Similar clusters is removed by retrieving centroids connected as *Identity*, *Subsumption* or *Elaboration.*

iv) Clusters from (iii) is merged to minimize the occurrence of the same sentences in multiple clusters (refer Figure 3(c)).

v) Step (iii) and (iv) are iterated until the number of clusters is convergence

The algorithm of similar text clustering is illustrated in Figure 3. In this work, we performed and observed 2 types of text clustering, which are:

i) *RRCluster 1*, which consist of *Identity* (*ID)*, *Subsumption* (*SUB*), *Elaboration* (*ELA*) and *Overlap (OVE)*
ii) *RRCluster2*, which consist of *Identity (ID)*, *Subsumption (SUB)* and *Elaboration (ELA)*

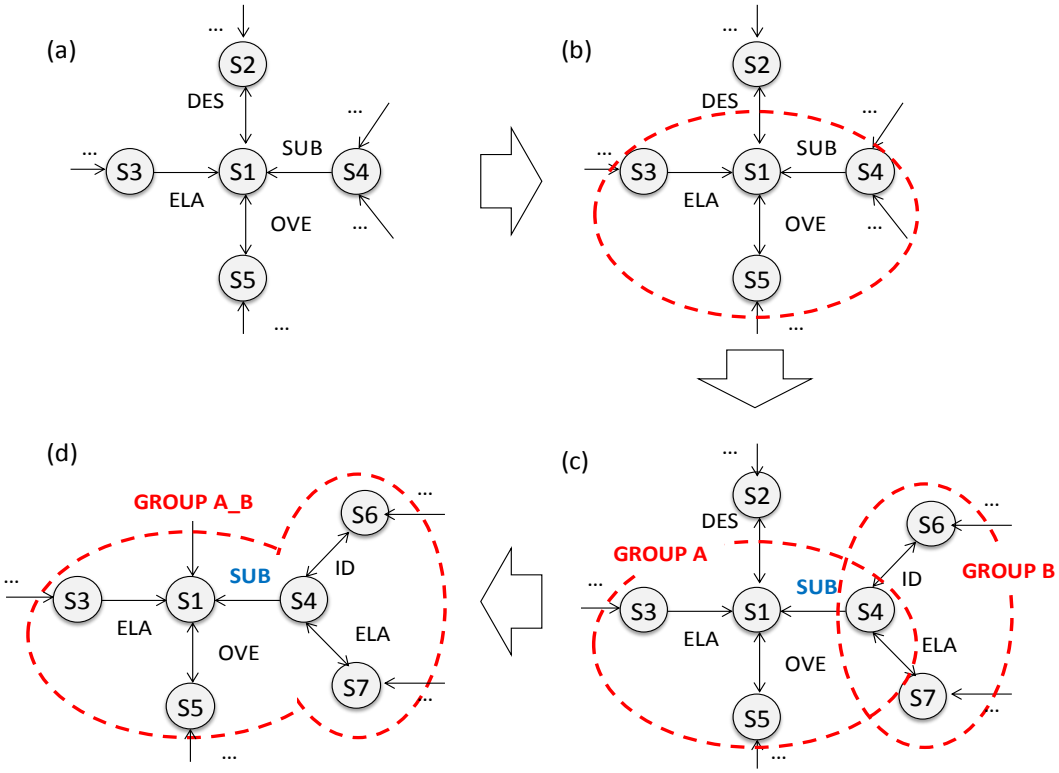

Figure 3. Rhetorical relation-based clustering algorithm

3.4. Cluster-based Summary Generation

We performed a cluster-based text summarization using clusters of similar text constructed by exploiting rhetorical relations between sentences. We used Cluster-based Conditional Markov Random Walk Model [26] to measure the saliency scores of candidate summary. Here we defined the centroid as relevant candidate summary since each centroid represents the whole cluster. The Conditional Markov Random Walk Model is based on the two-layer link graph including both the sentences and the clusters. Therefore, the presentation of the two layer graph are is denoted as $G^* = < V_s, V_C, E_{SS}, E_{SC}$. Suppose $V_s = V = v_i$ is the set of sentences and $V_c = C = c_j$ is the set of hidden nodes representing the detected theme clusters, where $E_{SS} = E = e_{ij} \mid v_i \in V_s$ corresponds to all links between sentences. $E_{SC} = e_{ij} \mid v_i \in V_s, c_j \in V_c, c_j = clus(v_i)$ corresponds to the correlation between a sentence and its cluster. The score is computed measured as follows:

$$SenScore = \mu \cdot \sum_{all\ j \neq i} SenScore(v_j) \cdot \tilde{M}^*_{ij,i} + \frac{(1-\mu)}{|V|}$$

μ is the damping factor set to 0.85, as defined in the PageRank algorithm. $\tilde{M}*_{j,i}$ refers to row-normalized matrix $\tilde{M}^*_{j,i} = (\tilde{M}^*_{j,i})_{|V| \times |V|}$ to describe $\tilde{G}^*$ with each entry corresponding to the transition probability, shown as follows:

$$\tilde{M}^{*}_{ij,i} = p(i \rightarrow j \mid clus(v_i), clus(v_i))$$

Here, *clus(v_i)* denotes the theme cluster containing sentence v_i. The two factors are combined into the transition probability from v_i to v_j defined as follows:

$$p(i \rightarrow j \mid clus(v_i), clus(v_i)) = \frac{f(i \rightarrow j \mid clus(v_i), clus(v_i))}{\sum_{k=1}^{|V|} f(i \rightarrow k \mid clus(v_i), clus(v_k))}, if \sum f \neq 0$$

$f(i \rightarrow j \mid clus(v_i), clus(v_i))$ denotes the new affinity weight between two sentences v_i and v_j, where both sentences belong to the corresponding two clusters. The conditional affinity weight is computed by linearly combining the affinity weight conditioned on the source cluster, *i.e.* $f(i \rightarrow j \mid clus(v_i))$ and the affinity weight conditioned on the target cluster *i.e.* $f(i \rightarrow j \mid clus(v_j))$, defined in the following equation.

$$\begin{aligned} f(i \rightarrow j \mid clus(v_i), clus(v_i)) &= \lambda \cdot (f(i \rightarrow j \mid clus(v_i)) + (1-\lambda) \cdot f(i \rightarrow j \mid clus(v_i)) \\ &= \lambda \cdot f(i \rightarrow j) \cdot \pi(clus(v_i)) \cdot \omega(v_i, clus(v_i)) \\ &\quad + (1-\lambda) \cdot f(i \rightarrow j) \cdot \pi(clus(v_j)) \cdot \omega(v_j, clus(v_j)) \\ &= f(i \rightarrow j) \cdot (\lambda \cdot \pi(clus(v_i)) \cdot \omega(v_i, clus(v_i)) \\ &\quad + (1-\lambda) \cdot \pi(clus(v_j)) \cdot \omega(v_j, clus(v_j)) \end{aligned}$$

Where $\lambda \in [0,1]$ is the combination of weight controlling the relative contributions from the source cluster and the target cluster[2]. $\pi(clus(v_i)) \in [0,1]$ refers to the importance of cluster $clus(v_i)$ in the whole document set *D* and $\omega(v_i, clus(v_i)) \in [0,1]$ denotes the strength of the correlation between sentence v_i and its cluster $clus(v_i)$. In this work, $\pi(clus(v_i))$ is set to the cosine similarity value between the cluster and the whole document set, computed as follows:

$$\pi(clus(v_i)) = sim_{cosine}(clus(v_i), D)$$

Meanwhile, $\omega(v_i, clus(v_i))$ is set to the cosine similarity value between the sentence and the cluster where the sentence belongs, computed as follows:

$$\omega(v_i, clus(v_i)) = sim_{cosine}(v_i, clus(v_i))$$

The saliency scores for the sentences are iteratively computed until certain threshold, θ is reached[3].

4.EXPERIMENT

4.1. Data

CST-annotated sentences are obtained from Cross-document Structure Theory Bank [5]. Our system is evaluated using 2 data sets from Document Understanding Conference, which are DUC'2001 and DUC'2002 [34].

4.2. Result and Discussion

4.2.1 Identification of Rhetorical Relations

SVMs classified the rhetorical relation of a sentence pair, S_1 and S_2, by considering the relationship type of S_1 according to S_2, and vice versa. In this word, we focused on the strength of the connection, rather than the number of the rhetorical relations belong to each connection. Since that a sentence pair might contain multiple relations, we assigned the strongest relations to present each connection. We conducted analysis to verify the most significant features against every relation. We calculated the sum of the vector component products to evaluate the effectiveness of each feature. The absolute value of weight directly reflects the importance of a feature in discriminating the two classes. The easy interpretation of the obtained weight values allows to identify the best features in case of a high-dimensional feature space. The evaluation results shown in Table 4 demonstrates the top 5 of most significant features for each relation. For instance, *Identity* indicates that both sentences are the same type of speech, which is indirect speech, while the cosine similarity and word overlap metrics indicates a value of 0.7 and above. From this evaluation, we concluded that the following features show most significant characteristics during classification of most relations:

(i) Similarity measurements
(ii) Grammatical relationship
(iii) Number of entities

[2]We set $\lambda = 0.5$ for fair evaluation with methods adopted from (Wan and Yang, 2008)
[3]In this study, the threshold, θ is set to 0.0001

Identity	Type of Speech (S_1)= Indirect and Type of Speech (S_1) = Indirect $0.7 \leq$ Cosine similarity ≤ 0.8 $0.9 \leq$ Subject Overlap(S_1) ≤ 1.0 Overlap Word (S_2) $\geq$ Overlap Word (S_1) Named Entities (S_1) $\geq$ Named Entities (S_2)
Subsumption	Length (S_1) $\geq$ Length(S_2) Type of Speech (S_1) = Indirect and Type of Speech (S_1) = Indirect Named Entities (S_1) $\geq$ Named Entities (S_2) $0.2 \leq$ Longest Common Substring ≤ 0.3 $0.9 \leq$ Subject Overlap (S_1) ≤ 1.0
Elaboration	Type of Speech (S_1) = Indirect and Type of Speech (S_1) = Indirect Named Entities (S_1) $\geq$ Named Entities (S_2) Length (S_1) $\geq$ Length (S_2)

	Overlap Word (S_2) $\geq$ Overlap Word(S_1) 0.4 $\leq$ Subject Overlap (S_1) $\leq$ 0.5
Overlap	0.9 $\leq$ Subject Overlap (S_1) $\leq$ 1.0 0.1 $\leq$ Longest Common Substring $\leq$ 0.2 0.1 $\leq$ Bigram similarity $\leq$ 0.2 0.2 $\leq$ Overlap Word (S_2) $\leq$ 0.3 0.2 $\leq$ Cosine similarity $\leq$ 0.3
Change of Topic	Type of Speech (S_1) = Indirect and Type of Speech (S_1) = Indirect 0.0 $\leq$ Longest Common Substring $\leq$ 0.1 0.9 $\leq$ Subject Overlap (S_1) $\leq$ 1.0 0.0 $\leq$ Cosine similarity $\leq$ 0.1 0.0 $\leq$ Overlap Word (S_2) $\leq$ 0.1
Description	Type of Speech (S_1)= Indirect and Type of Speech (S_1) = Indirect Subject Overlap (S_1) $\leq$ 0.0 Named Entities (S_1) $\geq$ Named Entities (S_2) Length (S_2) $\geq$ Length (S_1) 0.0 $\leq$ Bigram similarity $\leq$ 0.1
No Relations	Subject Overlap (S_1) $\leq$ 0.0 Subject Noun Overlap (S_1) $\leq$ 0.0 0.0 $\leq$ Cosine Similarity $\leq$ 0.1 Bigram Similarity $\leq$ 0.0 Overlap Word (S_1) $\leq$ 0.0

The rhetorical relations assigned by SVMs are manually evaluated by 2 human judges. Since no human annotation is available for DUC data sets, 5 times of random sampling consisting 100 sentence pairs is performed against each document set of DUC'2001 and DUC'2002). The human judges performed manual annotation against sentence pairs, and assessed if SVMs assigned the correct rhetorical relation to each pair. The correct rhetorical relation refers to either one of the relations assigned by human judges in case of multiple relations exist between the two sentences. As a baseline method, the most frequent relation in each set of sampling data is assigned to all sentence pairs. We evaluated the classification of rhetorical relations by measuring the Precision, Recall and F-measure score.

Identity shows the most significant performance of Precision, where the value achieved more than 90% in both data sets. Meanwhile, the Precision value for *Description* performed the worst compared to others in both data sets. As for Recall value, *Identity*, *Subsumption*, *Elaboration* and Description yield more than 80%, meanwhile *Change of Topic* and *No Relation* performed the worst with Recall of 60% in both data sets. We found that SVMs was unable to identify Change *of Topics*, when multiple subjects (especially contained personal pronoun) occurred in a sentence. According to F-Measure, SVMs performed well during the classification of *Identity*, *Subsumption* and *Elaboration* with the Precision values achieved are above 70% for most data set. Overall, compared to other relations, the *Identity* classification by SVMs performed the best in each evaluation metric as expected. Sentence pair with *Identity* relation shows significant resemblance in similarity value, grammatical relationship and number of entities. For instance, the similarity between sentence pair is likely close to 1.0, and there are major overlap in subject and the object of the sentences. *Subsumption* and *Elaboration* indicate promising potential of automated

classification using SVMs with F-measure achieved higher than 70%. We observed that characteristics such as similarity between sentences, grammatical relationship and number of entities are enough to determine the type of rhetorical relation of most data sets. Therefore, we considered the ratio of rhetorical relations except for *No Relations* show a great potential for automated classification with small number of annotated sentences.

We found that the lack of significant surface characteristic is the main reason of misclassification of relations such as *Overlap, Change of Topics* and *Description*. Therefore, we conducted further analysis using confusion matrix [35] to determine the accuracy of classification by SMVs. Confusion matrix compares the classification results by the system and actual class defined by human, which useful to identify the nature of the classification errors.

Table 6 and 7 describe the evaluation result of confusion matrix for DUC'2001 and DUC'2002, respectively. The analysis is done against each relation independently. Each table shows the classification nature of rhetorical relations according to the number of sentences pair. We also included the accuracy and reliability value of every relations. For instance, according to evaluation of DUC'2001 in Table 6, from 44 pairs of sentences with *Identity* relation, our system has been able to classify 43 pairs of them as *Identity* correctly, while 1 pair misclassified as *Subsumption*. As a result, the Accuracy and Reliability value achieved for *Identity* are 1.000 and 0.977, respectively.

Table 5. Evaluation result for identification of rhetorical relations

Relations	DUC'2001			DUC'2002		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Baseline	0.875	0.114	0.201	0.739	0.108	0.188
Identity	0.980	1.000	0.989	0.849	1.000	0.917
Subsumption	0.721	0.984	0.830	0.685	0.900	0.773
Elaboration	0.664	0.952	0.778	0.652	0.901	0.743
Overlap	0.875	0.532	0.653	0.739	0.556	0.633
Change of Topics	0.591	0.709	0.640	0.618	0.589	0.597
Description	0.841	0.947	0.886	0.817	0.856	0.826
No Relations	1.000	0.476	0.632	0.966	0.475	0.628

Table 6. Evaluation of Confusion Matrix for DUC'2001

		Classification by System							Accuracy
		ID	SUB	ELA	OVE	CHT	DES	NOR	
Actual Class	ID	43	0	0	0	0	0	0	1.000
	SUB	1	61	0	0	0	0	0	0.984
	ELA	0	2	48	0	0	1	0	0.941
	OVE	0	3	12	57	3	2	0	0.533
	CHT	0	5	6	6	51	3	0	0.718
	DES	0	0	0	0	2	59	0	0.967
	NOR	0	3	5	3	30	2	35	0.449
		0.977	0.726	0.676	0.864	0.593	0.881	1.000	

Table 7. Evaluation of Confusion Matrix for DUC'2002

		Classification by System							Accuracy
		ID	SUB	ELA	OVE	CHT	DES	NOR	
Actual Class	ID	55	0	0	0	0	0	0	1.000
	SUB	6	51	0	0	0	0	0	0.895
	ELA	0	4	35	0	0	0	0	0.897
	OVE	2	12	6	54	2	2	0	0.557
	CHT	1	4	9	10	40	2	1	0.597
	DES	0	0	0	0	8	70	0	0.886
	NOR	0	3	6	10	13	7	36	0.480
Reliability		0.859	0.689	0.614	0.730	0.635	0.864	0.973	

Despite the errors discovered during the identification of rhetorical relations, the classification by SVMs shows a promising potential especially for *Identity*, *Subsumption*, *Elaboration* and *No Relation*. In future, the increment of annotated sentences with significant characteristics of each relation will improve the identification of rhetorical relation. For instance, in this experiment, *Overlap* refers to sentences pair that shares partial information with each other. Therefore, we used Bigram similarity and Longest Common Substring metric to measure the word sequences in sentences. However, these metrics caused sentences with long named entity, *e.g.* ``*President George Bush*'' and ``*Los Angeles''*, as having consecutive words which contributed to false positive result of *Overlap* relation. The increment of annotated sentences consists of consecutive common nouns and verbs will help to precisely define *Overlap* relation. Moreover, improvement such as the usage of lexical database to extract lexical chain and anaphora resolution tool can be used to extract more characteristics from each relation.

4.2.2 Rhetorical Relation-based Clustering

We evaluated our method by measuring the cohesion and separation of the constructed clusters. The cluster cohesion refers to how closely the sentences are related within a cluster, measured using Sum of Squared Errors (SSE) [49]. The smaller value of SSE indicates that the sentences in clusters are closer to each other. Meanwhile, Sum of Squares Between (SSB) [49] is used to measure cluster separation in order to examine how distinct or well-separated a cluster from others. The high value of SSB indicates that the sentences are well separated with each other. Cosine similarity measurement is used to measure the similarity between sentences in both SSE and SSB evaluation. We also obtained the average of Silhouette Coefficient (SC) value to measure the harmonic mean of both cohesion and separation of the clusters [36][37]. The value range of the Silhouette Coefficient is between 0 and 1, where the value closer to 1 is the better.

Table 8 shows the evaluation results for cohesion and separation of the clusters. *RRCluster1* refers to the clusters constructed by *Identity*, *Subsumption* and *Elaboration*, while *RRCluster1* refers to the clusters constructed by *Identity*, *Subsumption*, *Elaboration* and *Overlap*. We also used K-Means clustering for comparison [38]. K-means iteratively reassigns sentences to the closest clusters until a convergence criterion is met. Table 8 indicates that *RRCluster2*, which generates clusters of sentences with strong connections *Identity*, *Subsumption* and *Elaboration*, demonstrates the best SSE value (4.181 for DUC'2001 and 3.624 for DUC'2002), which shows the most significant cohesion within clusters. In contrast, *RRCluster1* which includes *Overlap* during clustering indicates the most significant separation between clusters with the best SSB value (397.237 for DUC'2001 and 257.118 for DUC'2002). *RRCluster1* generated bigger clusters, therefore resulted wider separation from other clusters. The average Silhouette Coefficient shows that our method, *RRCluster1* (0.652 for DUC'2001 and 0.636 for DUC'2002) and *RRCluster2* (0.628 for DUC'2001 and 0.639 for DUC'2002) outranked K-Means (0.512 for DUC'2001 and 0.510 for DUC'2002) for both data sets.

In addition, we examined the clusters by performing a pair-wise evaluation. We sampled 5 sets of data consisting 100 sentences pairs and assessed if both sentences are actually belong to the same clusters. Table 9 shows the macro average Precision, Recall and F-measure for pair-wise evaluation. *RRCluster2*, which excludes Overlap relation during clustering, demonstrated a lower Recall value compared to *RRCluster1* and K-Means. However, the Precision score of *RRCluster2* indicates better performance compared to K-Means. Overall, *RRCluster1* obtained the best value for all measurement compared to *RRCluster2* and K-Means for both data sets. We achieved optimum pair-wise results by including *Overlap* during clustering, where the F-measure obtained for DUC'2001 and DUC'2002 are 0.770 and 0.766, respectively.

Table 8. Evaluation result for cohesion and separation of clusters

Data Set	Evaluation	Clustering Method		
		K-Means	RRCluster1 (ID,SUB,ELA,OVE)	RRCluster2 (ID, SUB, ELA)
DUC'2001	Average SSE	7.271	4.599	4.181
	Average SSB	209.111	397.237	308.153
	Average SC	0.512	0.652	0.628
DUC'2002	Average SSE	6.991	3.927	3.624
	Average SSB	154.511	257.118	214.762
	Average SC	0.510	0.636	0.639

Table 9. Evaluation result for pair-wise

Data Set	Evaluation	Clustering Method		
		K-Means	RRCluster1 (ID,SUB,ELA,OVE)	RRCluster2 (ID, SUB, ELA)
DUC'2001	Precision	0.577	0.783	0.805
	Recall	0.898	0.758	0.590
	F-Measure	0.702	0.770	0.678
DUC'2002	Precision	0.603	0.779	0.750
	Recall	0.885	0.752	0.533
	F-Measure	0.716	0.766	0.623

We made more detailed comparison between clusters constructed by K-Means and our method. The example of the clustered sentences by each method from the experiment is shown in Table 10. K-Means is a lexical based clustering method, where sentences with similar lexical often be clustered as one group although the content semantically different. The 5th sentences from K-Means cluster in Table 10 demonstrates this error. Meanwhile, our system, *RRCluster1* and *RRCluster2* performed more strict method where not only lexical similarity, but also syntactic similarity, *i.e* the overlap of grammatical relationship is taken into account during clustering. According to Table 8, Table 9 and Table 10, the connection between sentences can allow text clustering according to the user preference. For instance, *RRCluster2* performed small group of similar sentences with strong cohesion in a cluster. In contrast, *RRCluster1* method performed clustering of sentences with *Identity*, *Subsumption*, *Elaboration* and *Overlap*, which are less strict than *RRCluster2*, however presents strong separation between clusters. In other words, the overlapping information between clusters are lower compared to *RRCluster2*. Thus, the experimental results demonstrate that the utilization of rhetorical relations can be another alternative of cluster construction other than only observing word distribution in corpus.

Table 10. Comparison of sentences from K-Means and proposed methods clusters

K-Means		
√	Centroid	*Tropical Storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night.*
√	1	*Earlier Wednesday Gilbert was classified as a Category 5 storm, the strongest and deadliest type of hurricane.*
√	2	*Such storms have maximum sustained winds greater than 155 mph and can cause catastrophic damage.*
√	3	*As Gilbert moved away from the Yucatan Peninsula Wednesday night , the hurricane formed a double eye, two concentric circles of thunderstorms often characteristic of a strong storm that has crossed land and is moving over the water again.*
√	4	*Only two Category 5 hurricanes have hit the United States the 1935 storm that killed 408 people in Florida and Hurricane Camille that devastated the Mississippi coast in 1969, killing 256 people.*
x	5	*"Any time you contract an air mass , they will start spinning . That's what makes the tornadoes , hurricanes and blizzards , those winter storms",Bleck said.*
RRCluster1		
√	Centroid	*Tropical Storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane*

		Saturday night.
√	1	*On Saturday, Hurricane Florence was downgraded to a tropical storm and its remnants pushed inland from the U.S. Gulf Coast.*
√	2	*The storm ripped the roofs off houses and flooded coastal areas of southwestern Puerto Rico after reaching hurricane strength off the island's southeast Saturday night.*
√	3	*Hurricane Gilbert, one of the strongest storms ever, slammed into the Yucatan Peninsula Wednesday and leveled thatched homes, tore off roofs, uprooted trees and cut off the Caribbean resorts of Cancun and Cozumel.*
√	4	*It reached tropical storm status by Saturday and a hurricane Sunday.*
√	5	*Tropical Storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night.*
RRCluster2		
√	Centroid	*Tropical Storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night.*
√	1	*On Saturday, Hurricane Florence was downgraded to a tropical storm and its remnants pushed inland from the U.S. Gulf Coast.*
√	2	*The storm ripped the roofs off houses and flooded coastal areas of southwestern Puerto Rico after reaching hurricane strength off the island's southeast Saturday night.*
√	3	*It reached tropical storm status by Saturday and a hurricane Sunday.*
√	4	*Tropical Storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night.*

4.2.3 Cluster-based Summary Generation

We generated short summaries of 100 words for DUC'2001 and DUC'2002 to evaluate the performance of our clustering method, and to observe if rhetorical relation-based clustering benefits the multi-document text summarization. The experimental results also include the evaluation of summaries based on clusters generated by Agglomerative Clustering, Divisive Clustering and K-Means as comparison, adopted from [26]. The ROUGE-1 and ROUGE-2 score of clustering method shown in Table 11.

For DUC'2001 data set, our *RRCluster1* performed significantly well for ROUGE-1 and ROUGE-2 score, where we outperformed others with highest score of 0.3602 and 0.0736, respectively. Divisive performed the worst compared to other methods. As for DUC'2002 data set, Agglomerative obtained the best score of ROUGE-1 with 0.3854, while *RRCluster2* yield the lowest score of 0.3591. In contrast, *RRCluster1* gained the best score of ROUGE-2 with 0.0873.

We observed that our proposed *RRCluster1* performed significantly well with ROUGE-2. During the classification of rhetorical relations, we also considered word sequence of Bigram to determine rhetorical relations, therefore resulted a high score of ROUGE-2. However, the ROUGE-1 score of our proposed methods performed poorly for DUC'2002 data sets, especially for *RRCluster2*. This technique, which considers *Identity, Subsumption* and *Elaboration* during text clustering certainly constructed clusters with high cohesion, but also limits the clustering to sentences with only strong connections. This lead to the construction of many small clusters with possibility of partial overlaps of information with other clusters. As a result, the structure of clusters in *RRCluster2* caused the low value of both ROUGE-1 and ROUGE-2 scores.

Although our method only achieved good ROUGE-2 score, we considered that rhetorical relation-based clustering shows a great potential since that our clustering method is at initial stage yet already outperformed some of the well-established clustering method. Clearly, rhetorical relation-based cluster need some further improvement in future in order to produce better result. However, the result we obtained from this experiment shows that rhetorical relation-based clustering can enhance the cluster-based summary generation.

Table 11. Comparison of ROUGE score for DUC'2001 and DUC'2002

Method	DUC'2001		DUC'2002	
	ROUGE-1	ROUGE-2	ROUGE-1	ROUGE-2
Agglomerative	0.3571	0.0655	0.3854	0.0865
Divisive	0.3555	0.0607	0.3799	0.0839
K-Means	0.3582	0.0646	0.3822	0.0832
RRCluster2	0.3359	0.0650	0.3591	0.0753
RRCluster1	0.3602	0.0736	0.3693	0.0873

5. CONCLUSIONS

This paper investigated the relevance and benefits of the rhetorical relation for summary generation. We proposed the application of rhetorical relations exist between sentences to text clustering which improved extractive summarization for multiple documents. This work focused on the extraction of candidate summaries from generated clusters and redundancy elimination. We examined the rhetorical relations from Cross-document Theory Structure (CST), then selected and redefined the relations that benefits text summarization. We extracted surfaces features from annotated sentences obtained from CST Bank and performed identification of 8 types of rhetorical relations using SVMs. Then we performed similar text clustering by exploiting rhetorical relations among sentences. We used ranking algorithm that include the cluster-level information, Cluster-based Conditional Markov Random Walk (Cluster-based CMRW) to measure the saliency score of candidates summary extracted from generated clusters. For DUC'2001, our proposed method, *RRCluster1* performed significantly well for ROUGE-1 and ROUGE-2 score with highest score of 0.3602 and 0.0736, respectively. Meanwhile, *RRCluster1* gained the best score of ROUGE-2 with 0.0873 for DUC'2002. This work has proved our theory that rhetorical relations can benefit the similar text clustering which enhanced text summarization. From the evaluation results, we concluded that the rhetorical relations are effective to construct theme clusters of common information and eliminate redundant sentences. Furthermore, our system does not rely on fully annotated corpus and does not require deep linguistic knowledge.

ACKNOWLEDGEMENTS

This research is supported by many individuals from multiple organization of University of Yamanashi, Japan and University of Perlis, Malaysia.

REFERENCES

[1] Mann, W.C. and Thompson, S.A.,``Rhetorical Structure Theory: Towards a Functional Theory of Text Organization", Text, 8(3), pp.243-281, 1988.

[2] Carlson, L., Marcu, D. and Okurowski, M.E., ``RST Discourse Treebank'', Linguistic Data Consortium 1-58563-223-6, 2002.

[3] Webber, B.L., Knott, A., Stone, M. and Joshi, A., ``Anaphora and Discourse Structure", Computational Linguistics 29 (4), pp. 545–588, 2003.

[4] Radev, D.R., ``A Common Theory of Information Fusion from Multiple Text Source Step One: Cross-Document", In Proc. of 1st ACL SIGDIAL Workshop on Discourse and Dialogue, Hong Kong, 2000.

[5] Radev, D.R., Otterbacher, J. and Zhang, Z., CSTBank: Cross-document Structure Theory Bank, http://tangra.si.umich.edu/clair/CSTBank/phase1.htm, 2003.

[6]Wolf, F., Gibson, E., Fisher, A. and Knight, M.,``Discourse Graphbank", Linguistic Data Consortium, Philadelphia, 2005.

[7] Marcu, D., ``From Discourse Structures to Text Summaries", In Proc. of the Association for Computational Linguistics (ACL) on Intelligent Scalable Text Summarization, pp. 82-88, 1997.

[8] Zhang, Z., Blair-Goldensohn, S. and Radev, D.R., ``Towards CST-enhanced Summarization", In Proc. of the 18th National Conference on Artificial Intelligence (AAAI) , 2002.
[9] Radev, D.R., Jing, H., Stys, M., Tam, D., ``Centroid-based Summarization of Multiple Documents", Information Processing and Management 40, pp. 919–938, 2004.
[10] Uzeda, V.R., Pardo, T.A.S., Nunes, M.G.V.,"A Comprehensive Summary Informativeness Evaluation for RST-based Summarization Methods", International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM) ISSN: 2150-7988 Vol.1, pp.188-196, 2009.
[11] Jorge, M.L.C and Pardo, T.S., ``Experiments with CST-based Multi-document Summarization", Workshop on Graph-based Methods for Natural Language Processing, Association for Computational Linguistics (ACL), pp. 74-82, 2010.
[12] Louis, A., Joshi, A., and Nenkova, A., ``Discourse Indicators for Content Selection in Summarization", In Proc. of 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), pp. 147-156, 2010.
[13] Otterbacher, J., Radev, D. and Luo, A.,``Revisions that Improve Cohesion in Multidocument Summaries: A Preliminary Study",In Proc. of Conference on Association of Computer Linguistics (ACL), Workshop on Automatic Summarization, pp. 27-36,2002.
[14] Teufel, S. and Moens, M.,``Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Structure", Computational Linguistics 28(4): 409-445, 2002.
[15] Pardo, T.A.S. and Machado Rino, L.H., ``DMSumm: Review and Assessment", In Proc. of Advances in Natural Language Processing, 3rd International Conference (PorTAL 2002), pp. 263-274,2002.
[16] McKeown,K., Klavans,J., Hatzivassiloglou,V., Barzilay,R. and Eskin, E., `` Towards Multi-document Summarization by Reformulation: Progress and prospects", In Proc. of the 16th National Conference of the American Association for Artificial Intelligence (AAAI), pp. 453-460, 1999.
[17] Marcu,D., and Gerber, L.,`` An Inquiry into the Nature of Multidocument Abstracts, Extracts, and their Evaluation", In Proc. of Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), Workshop on Automatic Summarization, pp. 1-8, 2001.
[18] Hardy,H., Shimizu,N., Strzalkowski,T., Ting,L., Wise,G.B., and Zhang,X.,`` Cross-document Summarization by Concept Classification", In Proc. of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 121-128, 2002.
[19] Radev, D.R., Jing,H., and Budzikowska, M., ``Centroid-based Summarization of Multiple Documents: Sentence extraction, Utility-based Evaluation, and User Studies", In ANLP/NAACL Workshop on Summarization, 2000.
[20] Carbonell,J.G. and Goldstein, J., `` The Use of MMR, Diversity-based Re-ranking for Reordering Documents and Producing Summaries," In Proc. of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 335-336, 1998.
[21] Hatzivassiloglou,V., Klavans,J., and Eskin, E.,`` Detecting Test Similarity Over Short Passages: Exploring Linguistic Feature Combinations via Machine Learning", In Proc. of Conference on Empirical Methods in Natural Language Processing (EMNLP),1999.
[22] Hatzivassiloglou,V., Klavans,J., Holcombe,M.L., Barzilay,R., Kan,M-Y., and McKeown, K.R.,`` SimFinder: A Flexible Clustering Tool for Summarization", In Proc. of Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), Workshop on Automatic Summarization, 2001.
[23] Cohen, W., `` Learning Trees and Rules with Set-valued Features", In Proc. of the 14th National Conference on Artificial Intelligence (AAAI), 1996.
[24] Barzilay,R., Elhadad,N., and McKeown, R.K., ``Sentence Ordering in Multi-document Summarization", In Proc. of the Human Language Technology Sarkar, K., ``Sentence Clustering-based Summarization of Multiple Text Documents", TECHNIA - International Journal of Computing Science and Communication Technologies, VOL. 2, NO. 1, (ISSN 0974-3375), pp. 325-335,2009.
[25] Sarkar, K., ``Sentence Clustering-based Summarization of Multiple Text Documents", TECHNIA - International Journal of Computing Science and Communication Technologies, VOL. 2, NO. 1, (ISSN 0974-3375), pp. 325-335,2009.
[26] Wan, X. and Yang, J., ``Multi-Document Summarization Using Cluster-Based Link Analysis" , In Proc. of the 31st Annual International Conference on Research and Development in Information Retrieval (ACM SIGIR) Conference, pp. 299-306, 2008.

[27] Erkanand, G.and Radev,D.R., ``LexPageRank:Graph-based Lexical Centrality as Salience in Text Summarization", Journal of Artificial Intelligence Research 22, pp.457-479, 2004.
[28] Mihalcea, R., and Tarau. P.,`` A language Independent Algorithm for Single and Multiple Document Summarization", In Proc. of International Joint Conference on Natural Language Processing (IJCNLP), 2005.
[29] Wan, X. and Yang. J.,`` Improved Affinity Graph based Multi-document Summarization", In Proc. of Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL), 2006.
[30] Nik Adilah Hanin Binti Zahri, Fumiyo Fukumoto, Suguru Matsuyoshi, "Exploiting Discourse Relations between Sentences for Text Clustering", In Proc. of 24th International Conference on Computational Linguistics (COLING 2012), Advances in Discourse Analysis and its Computational Aspects (ADACA) Workshop, pp. 17-31, December 2012, Mumbai, India.
[31] Vapnik, V. : The Nature of Statistical Learning Theory, Springer, 1995.
[32] Lin, D., ``PRINCIPAR- An Efficient, Broad-coverage, Principle-based Parser", In Proc. of 15th International Conference on Computational Linguistics (COLING), pp.482-488, 1994.
[33] Fillmore1998} Fillmore, C.J., Baker, C.F., and Lowe, J.,B., ``FrameNet and Software Tools", In Proc. of 17th International Conference on Computational Linguistics (COLING), 36th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 86-90, 1998.
[34] Buckland,L. & Dang, H.,Document Understanding Conference Website, http://duc.nist.gov/
[35] Kohavi, R. and Provost, F.,``Glossary of Terms", Machine Learning 30, No.2-3, pp. 271-274, 1998.
[36] IBM SPSS Statistic Database, ``Cluster Evaluation Algorithm" http://publib.boulder.ibm.com , 2011.
[37] Kaufman, L. and Rousseeuw, P., ``Finding Groups in Data: An Introduction to Cluster Analysis", John Wiley and Sons, London. ISBN: 10: 0471878766, 1990
[38] McQueen, J., ``Some Methods for Classification and Analysis of Multivariate Observations", In Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297, 1967.

Authors

N. Adilah Hanin Zahri graduated from Computer Science and Media Engineering, University of Yamanashi in 2006. She received MSc in 2009 and PhD in Human Environmental Medical Engineering in 2013 from Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi, Japan. Currently, she is working at Department of Computer Engineering, School of Computer and Communication Engineering in University of Malaysia Perlis, Malaysia.

Fumiyo Fukumoto graduated from Department of Mathematics in the faculty of Sciences, Gakushuin University, 1986. From 1986 to 1988, she joined R&D Department of Oki Electric Industry Co., Ltd. From 1988 to 1992, she joined Institute for New Generation Computer Technology (ICOT). She was at Centre for Computational Linguistics of UMIST (University of Manchester Institute of Science and Technology), England as a student and a visiting researcher, from 1992 to 1994, and awarded MSc. Since 1994, she has been working at University of Yamanashi, Japan. She is a member of ANLP, ACL, ACM, IPSJ and IEICE.

Suguru Matsuyoshi received the B.S. degree from Kyoto University in 2003, and the M.S. and Ph.D. degrees in informatics from Kyoto University, Japan, in 2005 and 2008, respectively. Prior to 2011, he was a Research Assistant Professor in Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Japan. Since 2011, he has been an Assistant Professor in Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi, Japan.

Ong Bi Lynn graduated with B. Eng. (Hons) Electrical and Electronics from Universiti Malaysia Sabah (UMS) in the year 2001. She received her Master of Business Administration from Universiti Utara Malaysia (UUM) in 2003. She obtained her Ph.D. in the field of Computer Network in the year 2008 from Universiti Utara Malaysia (UUM). Currently, she is working with Department of Computer Network Engineering, School of Computer and Communication Engineering in Universiti of Malaysia Perlis (UniMAP), Perlis, Malaysia.