PAPER

Link Analysis Based on Rhetorical Relations for Multi-Document Summarization

Nik Adilah Hanin BINTI ZAHRI[†], ***Nonmember***, **Fumiyo FUKUMOTO**[††a)], ***Member***, ***and* Suguru MATSUYOSHI**[††], ***Nonmember***

SUMMARY This paper presents link analysis based on rhetorical relations with the aim of performing extractive summarization for multiple documents. We first extracted sentences with salient terms from individual document using statistical model. We then ranked the extracted sentences by measuring their relative importance according to their connectivity among the sentences in the document set using PageRank based on the rhetorical relations. The rhetorical relations were examined beforehand to determine which relations are crucial to this task, and the relations among sentences from documents were automatically identified by SVMs. We used the relations to emphasize important sentences during sentence ranking by PageRank and eliminate redundancy from the summary candidates. Our framework omits fully annotated sentences by humans and the evaluation results show that the combination of PageRank along with rhetorical relations does help to improve the quality of extractive summarization.

key words: *probability model, N-grams, link-based analysis, support vector machine, rhetorical relation*

1. Introduction

Due to the accelerating rate of data growth on the Internet, retrieving vital information from a huge amount of data is crucial, and it requires a lot of time and efforts. As a result, automatic text summarization has become an important task. Text summarization limits the need for user to access the original documents and improves the efficiency of the information search.

The general approach of summarizing documents is extractive or abstractive summarization. Extractive summarization focuses on finding the most salient sentences from the original document, while abstractive summarization focuses on generating summary by selecting only important terms from documents and may not contain original phrase or word. Our work focuses on extractive summarization for multiple documents. Various techniques have been proposed such as centroid-based summarization method [1], automated document indexing based on statistical latent model [2], and Cross-document Structure Theory (CST) based summarizer [3], [4].

In accordance with the information overload nowadays, previous works show many attempts to improve summary generation for multiple documents. A multi-document summarization system is much more complex than a single document summarization. The system should be able to generate a short summary with concise information of multiple topics and provide a wide diversity of views on the topic. 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 order to solve this problem, our work focuses on evaluating the connections, *i.e.* the rhetorical relations among the sentences, and we proposed a rhetorical relation-based link analysis to evaluate the complementarity and redundancy of the summary candidates for multi-document summarization. We first performed content selection by taking into account the news headlines to retrieve relevant candidate summary from individual documents, hereafter referred to as local context. We observed that since news headline usually gives an overview of the overall written events, the utilization of news headline can benefit and improve the local context extraction. We then ranked these local contexts using link analysis to retrieve the most salient sentence from the whole document, hereafter referred to as the global context. Our method shares the same concept as CST based text summarization. The difference is that we applied the rhetorical relations to link analysis, PageRank and modified the sentences ranking process. The rhetorical relations here were used to emphasize or disregard the connectivity among the sentences in the document sets. For most ranking algorithm including PageRank, redundancy is the most common problem. Therefore, we also utilized the identified rhetorical relations between sentences to eliminate redundancy from the ranked sentences.

Our method consists of two steps, (i) the extraction of local context using statistical model, and (ii) the retrieval of global contexts using rhetorical relation-based PageRank. The rest of the paper is organized as follows. The next section provides an overview of the existing techniques. Section 3 describes our summarizer system. Finally, we report experiments and conclude our discussion with some direction for further works.

2. Related Work

Since large scale machine readable textual corpus has be-

Manuscript received March 28, 2012.
Manuscript revised October 15, 2012.

[†]The author is with the Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi, Kofu-shi, 400–8511 Japan.
[††]The authors are with the Faculty of Medicine and Engineering, Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi, Kofu-shi, 400–8511 Japan.
a) E-mail: fukumoto@yamanashi.ac.jp

DOI: 10.1587/transinf.E96.D.1182

come available, there are many techniques have been proposed to improve text summarization using either surface characteristics or deep linguistics knowledge between the sentences. The most popular method is analyzing the surface characteristics of each sentence using statistical models. One of them is proposed by [5]. This method applies statistical models for content selection and term ordering process to generate summaries. The system builds a model of the relationship of the surface features appearing in both headline and document content to select the most appropriate context. Another work, text summarization using Probabilistic Latent Semantic Indexing (PLSI) [2], proposed an algorithm that extracts more than one topic from a document by analyzing surface features such as term frequency and graph structure for a single document.

Link based analysis technique also have been successfully used in text summarization. The most common algorithms is PageRank [6], designed for ranking Web Pages. PageRank considers the impact of both coming and outgoing links into one single model. It has shown that PageRank provided great performance during unsupervised sentence extraction in the context of text summarization task [7].

As for deep linguistic knowledge approach, one of the earliest works is CST based text summarization [3]. This method proposes the enhancement of text summarization by replacing low-salience sentences with sentences that have maximum numbers of CST relationship. The most recent work is a deep knowledge approach system [4], which ranks input sentences according to the number of CST relations exist between sentences. Most of the CST-based works observed the effects of individual CST relationships to the summary generation, and focused on the user preference-based summarization. Text summarization using deep linguistic knowledge mainly requires manually annotated corpus.

It has been shown in the previous works that the information obtained from CST can improve multi-document summarization. We followed this idea, but we applied it to the different approach. Previous works observed the effect of more than 10 types of relation individually against the quality of generated summary. Some of these relations, such as *Identity* and *Paraphrase*, shared similar characteristics and presented positive effect in summarization, while relations such as *Reader Profile* and *Fulfillment* showed negative effect on summary [3] . Regardless the positive or negative impact they might be to summarization, we found that it is redundant to determine similar relation with the same effect as different category. Relations with slight difference in terms of definition cause difficulty even during annotation by humans. Since that most summarization task is aimed to produce generic summaries, we examined relations which is crucial to this task and simplified the categories of the rhetorical relations. The modification of the relation is necessary so that each relation can be defined automatically using machine learning while sustaining the original effect of the relations against text summarization. This will limit the need to classify all rhetorical relations proposed by CST and omit fully annotation by humans. Our method considers connection between two sentences during the retrieval of the global context, and uses these connections to eliminate redundancy in summary.

3. System Overview

Our summarizer system is illustrated in Fig. 1. It consists of two tasks, which are the extraction of local context, and the retrieval of global contexts. The retrieval of global contexts task includes i) the identification of rhetorical relations between sentences, ii) sentences ranking using rhetorical relations based PageRank, and iii) redundancy elimination.

3.1 Extraction of Local Contexts

Document headline usually describes an overview of the entire events written in a document. Hence, we made use of the document headline to retrieve the local context, which are the salient sentences from individual document. We used statistical model [5] to learn the connection between the features occurred both in the headlines and the documents. The probability of the terms for local context can be computed as the product of the probability of the terms in the sentences, assuming that the likelihood of a word in the summary is independent with each other. Hence, the overall probability is computed as the product of the likelihood of (i) the selected term from the document, (ii) the term length and (iii) the most likely sequencing of the terms in the document sets, as shown in Eq. (1).

$$\begin{aligned} P(w_i, \ldots, w_n) = &\prod_{i=1}^{n} P(w_i \in H | w_i \in D) \\ &\cdot P(len(H) = n) \\ &\cdot \prod_{i=2}^{n} P(w_i | w_1, \ldots, w_{i-1}) \end{aligned} \tag{1}$$

where $P(w_i \in H | w_i \in D)$ is derived from conditional proba-

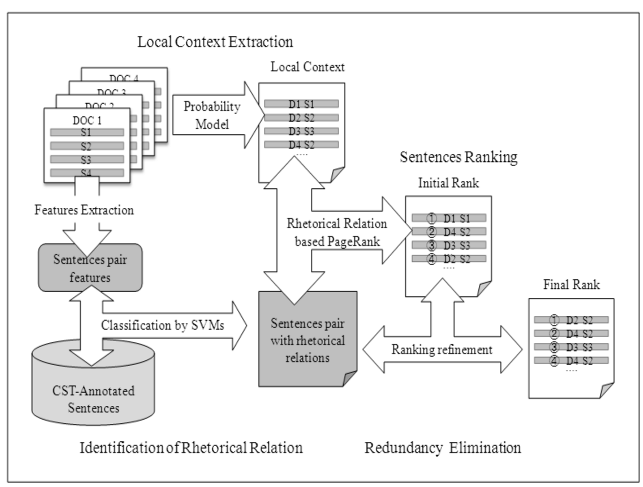


Fig. 1 Overview of the summarizer system.

bility of word occurred in the headline and documents, estimated as follows;

$$P(w_i \in H | w_i \in D) = \frac{P(w_i \in D | w_i \in H) \cdot P(w_i \in H)}{P(w_i \in D)} \quad (2)$$

H and D in Eq. (2) represent the list of words contained in headline, and document content, respectively. The conditional probability is computed for each word and used to compute the overall probability score for n-length terms of each sentence.

We observed that the average length of headlines in training data is 7 words with 49% of word content†. Therefore, we set the range of the term length†† from minimum 4 words array to maximum 10 words array with word content of 40% and above. Since the length of each sentence is different to each other, we measured the probability score for terms with n-length of 10% to 50% from the length of the sentences. For example, from 10 words sentences, *"The two countries have a military alliance with United State"*, the 40% from the length will be a 4 words array. Thus, we computed the probability score for each 4 consecutive words, for example, *"The two countries have"*, *"two countries have a"*, *"countries have a military"*, ... The probability of any word ordering for every term (iii) is computed by the probability model of the word sequence. Here, we used the simplest language model, the bigram model. The probability of a word sequence in a term is estimated by the product of the probabilities of words appear at the immediate left context. Meanwhile, the probability of the unseen word sequence in the training data are estimated by using back-off weight [8]. We built statistical models for categories where the document sets belong and measure the overall probability score according to these statistical models. Finally, the sentences include terms with probability score were extracted as local context for each document.

3.2 Retrieval of Global Contexts

3.2.1 Identification of Rhetorical Relations

When dealing with multi-document phenomena, our first step is to identify the rhetorical relationship between sentences and then pinpoints the relations that cause the redundancy. We later applied these relations to determine the connectivity types between sentences, which affected the retrieval of global context by PageRank as described in the next section. We also used these rhetorical relations for redundancy elimination during summary generation.

The types of the rhetorical relations determined here were based on types proposed by CST. We examined the definition of each CST relationship [3], [11], and we observed that 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 different version of event description. Refer the following example:

Example 1:

S_1 *RAI state TV reported that the pilot said the SOS was because of engine trouble.*

S_2 *RAI state TV reported that the pilot said he was experiencing engine trouble.*

Both sentences demonstrate the example of sentence pair that can represent *Identity*, *Paraphrase*, *Modality* and *Attribution* relations. The quality and the amount of the information in both sentences are the same. Another example of sentence pair that can represent similar relations is shown in the following example:

Example 2:

S_3 *The crash put a hole in the 25th floor of the Pirelli building, and smoke was seen pouring from the opening.*

S_4 *A small plane crashed into the 25th floor of a skyscraper in downtown Milan today.*

Both sentences can be categorized as *Elaboration* and *Follow-up*. We also found that *Subsumption* and *Elaboration* shares some similar characteristics, as shown in Example 3.

Example 3:

S_5 *The building houses government offices and is next to the city's central train station.*

S_6 *The building houses the regional government offices, authorities said.*

Sentence pair connected as *Subsumption* can also be defined as *Elaboration*. However, sentence pair belongs to *Elaboration* in Example 2 cannot be defined as *Subsumption*. In some cases, surface characteristics of *Elaboration* are different from *Subsumption*. Moreover, we considered *Subsumption* as one of rhetorical relations that cause redundancy in summary generation. Therefore, we kept *Subsumption* and *Elaboration* as two different relations so that we can precisely perform the automated identification using SVMs.

In this paper, we simplified and combined the similar relationship types proposed by CST. The combination of rhetorical relations in this paper is shown in Table 1. We also modified the definition of each relation in accordance with the combination of relationship type shown in Table 1. The taxonomy for rhetorical relations we used in the system is described in Table 2. Considering this approach does not require deep linguistic knowledge, we assumed that these 7 types of relations are enough to evaluate the complementarity and redundancy of candidate summary.

To identify these relations, we used a machine learning approach, Support Vector Machine (SVMs) [10]. Our purpose here is to identify these relations automatically and minimize the usage of fully CST-annotated sentences. We used CST-annotated sentence pairs obtained from CST-Bank [11] as training data for the SVMs. We classified each data into one of two classes, where we defined the value of feature to be 0 or 1. Features with more than 2 values is normalized into [0,1] range. This value is represented by 10

†We defined the nouns contained in headlines as word content.

††We observed 3 word array ranges within the average headline length in training data.

Table 1 Combination of rhetorical relations.

Relations by CST	Relations by the System
Identity Paraphrase Modality Attribution	Identity
Indirect Speech Citation	Citation
Subsumption	Subsumption
Elaboration Follow-up	Elaboration
Overlap Contradiction Fulfillment	Overlap
Change of Perspective Reader Profile	Change Of Topics
Description Historical Background	Description
Translation	-
Summary	-

Table 2 Taxonomy of rhetorical relation used in the system.

Relations	Description
Identity	Two text spans have the same information content.
Citation	S_1 cites information in S_2.
Subsumption	S_1 contains all information in S_2, plus other additional information not in S_2
Elaboration	S_1 elaborates or provide more information given generally in S_2
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 Topics	S_1 and S_2 provide different facts about the same entity.
Description	S_1 gives historical context or describes an entity mentioned in S_2.

dimensional space, where the value is divided into 10 values 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 feature vector is set to 0001000000.

We provided the following surface features to SVMs for learning and identified the relationship of the text span S_1 according to the given text span S_2.

1. Cosine similarity value between S_1 and S_2
We computed the similarity of two sentences using Eq. (3):
$$cosine(S_1, S_2) = \frac{\sum_i (s_{1,i} \times s_{2,i})}{\sqrt{\sum_i (s_{1,i})^2} \times \sqrt{\sum (s_{2,i})^2}} \quad (3)$$
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 both sentences. For example, sentence pair of *Identity* demonstrates similarity close to 1.0.

2. Word overlap between S_1 and S_2
We also used another similarity measurement, word overlap metric to compute the ratio of the same words appears in both sentences, defined as follows:
$$WordOve(S_1, S_2) = \frac{commonwords(S_1, S_2)}{words(S_1) + words(S_2)} \times 2 \quad (4)$$
The metric finds the number of common words in S_1 and S_2. For example, sentence pair of *Identity* or *Citation* demonstrate similarity close to 1.0.

3. Common overlap S_1 in S_2 , and vice versa
This metric is used to measure the occurrence ratio of the words from S_2 appear in S_1, and vice versa. The feature with higher overlap ratio is set as 1, and 0 for lower value.
$$Ove(S_1) = \frac{commonwords(S_1, S_2)}{words(S_1)} \times 2 \quad (5)$$
For instance, given the sentence pair with relation of *Subsumption*, the ratio of words from S_2 appear in S_1 is higher than the ratio of words from S_1 appear in S_2.

4. Longest common substring for S_1
Given two text span, S_1 and S_2, the metric finds the longest common substring of S_1 in S_2.
$$LCS(S_1) = \frac{len(MaxComSubstring(S_1, S_2))}{length(S_1)} \quad (6)$$
The ratio value describes the maximum consecutive words sequences appear in both sentences, which determines the characteristic of relation *i.e. Subsumption* or *Overlap*.

5. Comparison of entities
Stanford NER (CRF Classifier) of Named Entity Recognizer [12] is used to labels sequence of words which indicate the name of person, organization, or location in the text span. We computed the number of entities appear in both S_1 and S_2. For instance, S_1 from *Elaboration* presents more information compared with S_2, therefore, S_1 is more likely to have higher number of entities.

6. Comparison 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 compared which sentence contains more conjunctions. The higher the number of conjunctions, the more information is provided in the corresponding text span. The comparison of the number of conjunctions helps to determine relation *i.e. Elaboration* and *Subsumption*.

7. Comparison of lengths
We defined the length of S_j by the number of word occurs in the corresponding text span [13], and compared the length of both sentences. The length of both text spans shows whether both text span are *Identity*, where the length is the same, or one of the text spans presents

more information than another, where S_1 is longer, *i.e. Subsumption*.

8. Type of Speech
We determined the type of speech, whether the text span, S_j cites another sentences by detecting the occurrence of quotation marks. Type of speech helps to identify *Citation* between two text span.

9. Ratio overlap of grammatical relationship for S_1
We used a broad-coverage parser of English language, MINIPAR [14] to parse S_1 and S_2, and then extracted the grammatical relationship between words in the text spans. Here we extracted the number of *surface subject* (*Subj1*), the *subject of verb* (*Subj2*) and *object of verbs* (*Obj*). We then compared the grammatical relationship between the text span as follows:

$$SubjOve(S_1) = \frac{commonSubj1(S_1, S_2)}{Subj1(S_1)} \tag{7}$$

$$ObjOve(S_1) = \frac{commonObj(S_1, S_2)}{Obj(S_1)} \tag{8}$$

The ratio value describes whether S_2 provide information regarding the same entity of S_1, *i.e. Change of Topics*.

10. Comparison of entities and the surface subject for S_1
We computed the ratio of entities in S_2 which are the surface subject of S_1.

$$SubjEnt(S_1) = \frac{common(Subj1(S_1), Ent(S_2))}{Subj1(S_1)} \tag{9}$$

This ratio value determines the features of *i.e. Description*, where the S_1 describes any entities mentioned in S_2.

The value of features (1)~(4) is the value of each similarity metrics. Meanwhile, features (5)~(11) are measured for both sentences, and the value is set to 1 for higher, and 0 for lower feature.

3.2.2 Sentences Ranking

The ranking score of the extracted local contexts were measured according to their relative importance within the document set using PageRank. The link between sentences is considered as a vote of support. Therefore, the more links connected to the sentence, the more important the sentence become. In this model, we assumed that one sentence is linked to another sentences if there is a similarity value exists between them. Here, a sentence connectivity matrix is constructed based on the word overlap ratio (defined in Eq. (4)) between two sentences. We then assigned identified rhetorical relation to every connections/links between the sentences and modified the directionality of the links based on the type of the rhetorical relations. The rhetorical relations here were used to emphasize relevant connection and disregard the irrelevant connection between sentences in the document sets. We assigned two-way connection for *Identity* and *Citation* due to the similar amount and quality of information within both text spans. *Subsumption* and *Elaboration* describing text span S_1 provides more information compared to S_2, are set as a 1-way direction, which is from S_2 to S_1. *Description* sets a 1-way direction from sentence contained background information to the main sentence, which is also from S_2 to S_1. Others were considered having a 2-way direction since the information contained is independent to each other. Refer Fig. 2 for rhetorical relations assignment and link modification made in the system.

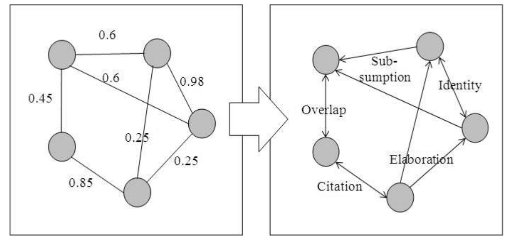


Fig. 2 Rhetorical relations assignment and link modification.

Let the overlap ratio value of both sentences be the value of each link. For a given sentence S_i, let $In(S_j)$ be the sentences that linked towards S_i. The PageRank score for sentence, S_i is defined as follows:

$$PageRank(S_i) = \frac{1-d}{N} + d \sum_{S_j \in In(S_i)} In(S_j) \tag{10}$$

where, d is the optimum damping factor, set as 0.85 [6] and N is the number of sentences in the document set. Although PageRank score was only computed against the extracted local contexts, the score was determined by assigned links from the entire sentences in the document set.

Finally, we sorted the ranking score of the local contexts in decreasing order. Here, we refer to the sentences with high value of PageRank as the global contexts. The high PageRank score indicates high amount of information content and high level of relevance in the entire document set. We defined the ranking of the global context by PageRank as initial rank.

3.2.3 Redundancy Elimination

The rhetorical relation based PageRank is proposed to emphasize the most salient sentences and to disregard lower priority sentences during global context retrieval. However, global contexts in the initial ranks might contain redundant sentences which affect the quality of information in final summary. Therefore, we applied redundancy elimination against initial rank to produce a better and improved final rank before summary extraction.

We utilized the rhetorical relations classified by SVMs to identify redundant sentence in the initial ranks. *Identity*,

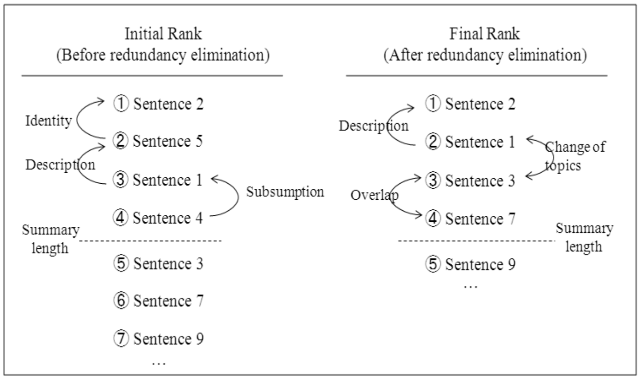


Fig. 3 Redundancy elimination.

Subsumption and *Citation* are relations that often cause redundancy. Therefore, we refined the initial rank by identifying these relations between the global contexts and eliminated sentences at the lower position in the initial ranks.

Figure 3 illustrates an example of hypothetical initial ranks of global context and the transformation to final rank after redundancy elimination process against *Identity* and *Subsumption*. According to this example, *Sentence 2* and *Sentences 5* share an *Identity* relation, and *Sentence 1* and *Sentences 4* share a *Subsumption* relation. Thus, we considered *Sentence 5* and *Sentence 4* contain redundant information and removed them from the initial rank. Sentences from refined rank were extracted as final summary.

4. Evaluation

4.1 Data

We used 1 year of Reuters'96 corpus to train and build statistical model for local context extraction. CST-annotated sentences were obtained from Cross-document Structure Theory Bank (CSTBank) [11]. Our system is evaluated using 3 data sets from Document Understanding Conference, which are DUC'2003, DUC'2004 and DUC'2007.

4.2 Extraction of Local Context

Using statistical model, we extracted sentences with salient term as relevant candidate summary, or the local context. We evaluated the salient terms from extracted local context by comparing them with the news headline from the individual document. We measured the precision, recall and F-measure score of the extracted salient terms for data sets DUC'2003 and DUC'2007. The DUC'2004 data set is omitted from this evaluation because no news headline available in the document set. The result is shown in Table 3. The best precision value for DUC'2003 and DUC'2007 are 0.397 and 0.449, respectively. The recall values, however, were quite low for both data set due to the different length of average salient terms and the average length of the document headlines in the test data. For instance, the average length of headlines in DUC'2003 is 8 words, where the average of extracted salient sentence is 5 words. As a result, we attained the best F-measure score of 0.251 for DUC'2003, and 0.293 for DUC'2007.

Table 3 The average of macro precision of extracted salient terms.

Len	DUC'2003			DUC'2007		
	Precision	Recall	F-measure	Precision	Recall	F-measure
10%	0.397	0.172	0.234	0.449	0.231	0.293
20%	0.344	0.206	0.251	0.338	0.242	0.271
30%	0.285	0.210	0.234	0.292	0.255	0.260
40%	0.257	0.203	0.222	0.278	0.252	0.250
50%	0.255	0.200	0.218	0.263	0.240	0.237

4.3 Classification of Rhetorical Relation

SVMs classified the type of 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 paper, we focused on the strength of the connection, rather than the number of the rhetorical relations belongs to each connection. Since that a sentence pair might contain multiple relations, we assigned the strongest relations to represent each connection. In order to determine which relation to be assigned, we examined the connection between two 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

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

Identity > Citation > Subsumption > Elaboration
Overlap > Change Of Topics > Description

According to training pattern by SVMs, we conducted analysis of 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 result is shown in Table 4. Table 4 demonstrates the top 10 of most significant features for each relation. For instance, *Identity* indicates that both sentences are the same type of speech, where both sentence have no quotation marks ($\text{Quo}(S_1)=0 \wedge \text{Quo}(S_2)=0$), while the cosine similarity and word overlap metrics indicates a value of 0.9 and above. Overall, the following features show significant characteristics during classification.

(i) similarity measurements
(ii) grammatical relationship
(iii) number of entities

Table 4 Top 10 most significant features of classified rhetorical relations.

Identitiy	Citation	Subsumption	Elaboration
Quo(S_1)=0∧Quo(S_2)=0	Quo(S_1)=0∧Quo(S_2)=1	Quo(S_1)=0∧Quo(S_2)=0	Len(S_1)≥Len(S_2)
0.9<cosine≤1.0	Len(S_1)≥Len(S_2)	Len(S_1)≥Len(S_2)	Quo(S_1)=0∧Quo(S_2)=0
0.9<Ove(S_1)≤1.0	0.3<Ove(S_1)≤0.4	0.9<Subj1Ove(S_1)≤1.0	Conj(S_1)≥Conj(S_2)
0.8<Ove(S_2)≤0.9	0.4<Ove(S_2)≤0.5	0.9<Subj2Ove(S_1)≤1.0	Ent(S_1)≥Ent(S_2)
0.8<WordOve≤1.0	Ent(S_1)=Ent(S_2)	Ent(S_1)≥Ent(S_2)	0.1<LCS(S_1)≤0.2
0.9<Subj2Ove(S_1)≤1.0	Conj(S_1)≥Conj(S_2)	0.7Ove(S_2)≤0.8	0.4<Ove(S_2)≤0.5
0.9<Ove(S_2)≤1.0	0.9<Subj1Ove(S_1)≤1.0	0.1LCS(S_1)≤0.2	0.4<WordOve≤0.5
0.9<Subj1Ove(S_1)≤1.0	0.3<LCS(S_1)≤0.4	0.8<Ove(S_2)≤0.9	0.9<Subj1Ove(S_1)≤1.0
0.9<Ove(S_1)≤1.0	0.7<Ove(S_2)≤0.8	Conj(S_1)≥Conj(S_2)	0.3<Ove(S_1)≤0.4
0.8<cosine≤0.9	0.5<cosine≤0.6	0.9<cosine≤1.0	0.4<cosine≤0.5

Overlap	Change of Topics	Description
Quo(S_1)=0∧Quo(S_2)=0	Quo(S_1)=0∧Quo(S_2)=0	Quo(S_1)=0∧Quo(S_2)=0
Len(S_1)≥Len(S_2)	0.9<Subj1Ove(S_1)≤1.0	Len(S_1) ≥ Len(S_2)
0.9<Subj2Ove(S_1)≤1.0	0.9<Subj2Ove(S_1)≤1.0	0.3<WordOve(S_1)≤0.4
0.9<Subj1Ove(S_1)≤1.0	0.3<WordOve(S_1)≤0.4	0.1< LCS(S_1)≤ 0.2
0.1< LCS(S_1)≤ 0.2	0.1< LCS(S_1)≤ 0.2	0.9<SubjEnt(S_1)≤1.0
Conj(S_1) = Conj(S_2)	Len(S_1)≥Len(S_2)	0.9<Subj1Ove(S_1)≤1.0
Len(S_2)≥Len(S_1)	0.4<cosine≤0.5	0.4<cosine≤0.5
Ent(S_2)≥Ent(S_1)	0.3<Ove(S_2)≤0.4	Conj(S_1)≥Conj(S_2)
0.3<WordOve(S_1)≤0.4	Ent(S_2) = Ent(S_1)	0.4<Subj1Ove(S_1)≤0.5
0.4<cosine≤0.5	Conj(S_1) = Conj(S_2)	0.4<Subj2Ove(S_1)≤0.5

Table 5 The macro average precision for each rhetorical relation.

Type	DUC'2003		DUC'2004		DUC'2007	
	Max	Ave	Max	Ave	Max	Ave
Identity	1.000	1.000	1.000	1.000	1.000	1.000
Citation	0.719	0.719	0.680	0.647	0.770	0.628
Subsumption	0.940	0.910	0.940	0.888	0.900	0.830
Elaboration	0.730	0.686	0.750	0.674	0.650	0.626
Overlap	0.810	0.748	0.730	0.650	0.610	0.556
Change of Topics	0.850	0.812	0.780	0.678	0.750	0.650
Description	0.520	0.488	0.450	0.414	0.440	0.404

Table 6 The number of annotated sentences used in the experiment.

Type	Number of Annotated Sentences
Identity	218
Citation	12
Subsumption	317
Elaboration	58
Overlap	157
Change of Topics	348
Description	70

We performed unsupervised classification of rhetorical relations as there are no CST-annotated sentences available for DUC data sets. Previous CST-based works require fully annotated sentences which limit the methods to certain data sets. To overcome this limitation, we proposed the automated classification of rhetorical relation that can be carried out regardless the number of annotated sentences available. Our purpose is to expand the data sets that can be used in the method. Therefore, to assess the classification by SVMs, we manually evaluated the rhetorical relations assigned to each sentence pair. We conducted 5 times of random sampling consisting 100 sentence pairs for each rhetorical relation. We assessed if SVMs assigned the correct rhetorical relation to each pair and measure the precision score against the sampling data.

Table 5 shows the macro average of precision for DUC'2003, DUC'2004 and DUC'2007 data set. Column *Max* and *Ave* respectively refers to the maximum precision value yield in the document set, and the macro average precision of each document set. According to the evaluation result, SVMs performed well during the classification of *Identity* and *Subsumption*, where the precision values achieved are above 80% for all data set. 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. Therefore, compared to other relations, the *Identity* classification by SVMs performed the best as expected.

For identification of rhetorical relations, the evaluation result of *Citation*, *Elaboration*, *Overlap* and *Change of Topics* are average compared with Identity and Subsumption. However, most of the average precision of these relations exceeded 60%, which demonstrate more than half of the sentence pairs were correctly classified by SVMs. This shows a promising result considering the limited number of annotated sentences, especially for *Citation* and *Elaboration* (as shown in Table 6). From all of relations, we observed that *Description* classification resulted many false positive assignment. The automatic classification of *Description* is harder compared with others due to lack of significant surface characteristics. The following sentence pair shows the example of false positive result of *Description*.

S_7 *If the Klan wants to march, they may apply for permit.*

S_8 *Dee sought to destroy the corporate Klan.*

According to this example, S_7 doesn't provide background information or any description of entities mentioned in S_8. Both sentences semantically present no relation to each other. However, since an entity is mentioned in S_8, both sentences were classified as *Description*. We

Table 7 The macro average result for small sampling data set.

Type	DUC'2003			DUC'2004			DUC'2007		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
Identity	0.986	1.000	0.994	1.000	0.988	0.994	1.000	1.000	1.000
Citation	0.587	0.953	0.722	0.605	0.976	0.699	0.523	0.963	0.672
Subsumption	0.725	0.981	0.831	0.643	0.940	0.750	0.659	0.888	0.722
Elaboration	0.635	0.888	0.732	0.743	0.814	0.763	0.701	0.938	0.799
Overlap	0.703	0.818	0.784	0.674	0.704	0.671	0.561	0.747	0.617
Change of Topics	0.653	0.654	0.644	0.681	0.617	0.639	0.772	0.476	0.586
Description	0.617	0.883	0.685	0.551	0.880	0.633	0.560	0.811	0.638

also found that rhetorical relations of some sentence pairs were unidentified by SVMs. Refer the following example:

S_9 *The president, a saxophone player who played in his high school band, has long championed the cause of preserving music program in school.*

S_{10} *He said he volunteered to get involved when he heard about the VHI1 program.*

In this example, the word *He* in S_{10} refers to the *president* in S_9. The personal pronoun *i.e.* he or she, which are frequently used in article writing, causing difficulties to identify the reference subject in sentences. This is one of the main reasons why no relation is assigned to certain related sentence pair.

Data obtained from Document Understanding Conference (DUC) are created for text summarization without CST annotation. In order to examine the fraction of how much rhetorical relations were retrieved by the system (recall), we randomly sampled 5 small data sets containing 100 sentence pairs each from document sets (DUC'2003, DUC'2004 and DUC'2007). We then performed manual annotation of rhetorical relations against these sentence pairs, and measure the precision, recall and F-measure value from classification result by SVMs. Refer Table 7 for the evaluation result of DUC'2003, DUC'2004 and DUC'2007 data.

Table 7 shows similar evaluation result for precision as Table 5. However, precision value for *Citation* performed worse compared to *Description* in DUC'2003 and DUC'2007. Evaluation result shows that sentence pairs with quotation marks mostly classified as *Citation*. Meanwhile, the recall values demonstrated significant result, where the values exceeded more than 70% for most rhetorical relations, except for *Change of Topics*. We found that SVMs was unable to identified *Change of Topics*, when multiple subjects (especially contained personal pronoun) occurred in a sentence. Overall, F-measure values for *Identity*, *Subsumption*, *Elaboration* and *Overlap* show significant result, where most of the accuracy exceeded 60%.

We observed that characteristics such as similarity between sentences, grammatical relationship and number of entities are enough to determine the type of rhetorical relation, except for *Description*. The best score of precision value for most relations show that the classification by SVMs is capable to exceed more than 70% of correct ratio for two out of three data sets (DUC'2003 and DUC'2004). Therefore, we considered the ratio of rhetorical relations except for Description, show a great potential for practical use. In future, the increment of annotated sentences with significant characteristics of each relation will improve the identification of rhetorical relation. Also, 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.

Table 8 ROUGE-1 score for DUC'2003 and DUC'2004.

Method	DUC'2003		DUC'2004	
	Max	Ave	Max	Ave
Statistical Model (20%)	0.367	0.287	0.401	0.303
PageRank	0.425	0.318	0.424	0.331
LexRank	0.373	0.367	0.381	0.374
Cont. LexRank	0.370	0.365	0.383	0.376
Rhetorical Relation PageRank (20%, w/o redundancy elimination)	0.461	0.361	0.454	0.357
Rhetorical Relation PageRank (20%) (20%, with redundancy elimination)	0.495	0.375	0.470	0.380

4.4 Summary Generation

We generated short summaries for DUC'2003 and DUC'2004, and long summaries for DUC'2007 to evaluate our summarizer system. The experimental results also include the evaluation of summaries generated by statistical model and global contexts extraction by PageRank as baseline. For DUC'2003 and DUC'2004 evaluation, we included the LexRank and continuous LexRank, other noble link analysis methods proposed by [15] as comparison. The evaluation results for DUC'2003 and DUC'2004 is shown in Table 8. We enclosed the best result, which are according to local context extracted according to 20% of salient terms. We also included the result of our method without redundancy elimination for comparison. Column *Max* refers to the maximum ROUGE-1 score from an individual document, and *Ave* refers to the average score measured from the whole document sets. Our proposed method (rhetorical relation based PageRank with redundancy elimination) outranked other system with the highest maximum and average score. Rhetorical relation based PageRank (without redundancy elimination) shows a quite high best score compared to PageRank and LexRank, however, the average score is lower compared to our system due to redundancy issue. At best case scenario, our system attained a ROUGE-1 score of 0.495 and 0.470 for DUC'2003 and DUC'2004, respectively. The modification of the directionality, and the links combination help to emphasis the most salient sentences, while the rhetorical relation between sentences help to deal with redundancy issues.

Table 9 ROUGE Score for DUC'2007.

Method	ROUGE-1	ROUGE-2
Statistical Model (20%)	0.355	0.064
PageRank	0.362	0.077
Hierarchical Pachinko Allocation Model (*hPAM*)	0.412	0.089
Two-Tiered Topic Model (*TTM*)	0.447	0.107
Rhetorical Relation PageRank (20%, w/o redundancy elimination)	0.396	0.095
Rhetorical Relation PageRank (20%) (20%, with redundancy elimination)	0.405	0.100

As for DUC'2007, we generated longer summaries, which contain 250 words. For comparison, we included the result of rhetorical relation based PageRank without redundancy elimination and other different approaches for extractive summarization, *hPAM* [16] and *TTM* [17] . Both *hPAM* and *TTM* and methods used hierarchical topic models to retrieve coherent sentences. *hPAM* considered both topic and hierarchy depth to characterize word distribution in every hierarchy model. Meanwhile, *TTM* observed word distribution in specific topics, and directly extract sentences include these high-level topics word as coherent sentence. The evaluation result for DUC'2007 is shown in Table 9. *TTM* performed the best compared to others. Rhetorical relation based PageRank without redundancy elimination shows average performance compared to TTM and proposed method, but outperformed other methods in both score. Our system (local context = 20%, with redundancy elimination) yield a 0.405 of ROUGE-1 and 0.100 of ROUGE-2. According to ROUGE-1, our system does not outperform other model except for baselines. On the other hand, ROUGE-2 shows better results, where our system performed better than *hPAM* and baselines, except for *TTM*.

Despite of the large number (about 60%) of false positive *Description* identified in DUC'2007 data set, the difference between our method and *TTM* (best method) for ROUGE-1 and ROUGE-2 score are small, which are 4.2% and 0.7%, respectively. From the methodology point of view, the *TTM* method requires synthetics and validation experiment with 2 additional data sets during the process of model development, while our method is a simple technique using SVMs with limited number of annotated sentence pairs from CSTBank. Plus, we only considered the connection between two sentences to enhance the performance of summary generation. Even with poor classification of certain relations, the ROUGE-2 score of our method outranked *hPAM*.

To relate one sentence to another, the subject and the object of the sentences are crucial information. Finding the reference of a personal pronouns or noun phrase with anaphora resolution certainly benefits the classification of rhetorical relation and text summarization. Tool such as JavaRAP [18], can be used to resolve third person pronouns, lexical anaphors, and identifies pleonastic pronouns in sentences. Acquiring this information is important especially in abstractive summarization. However, the extractive summarization in other hand focuses on finding the salient sentence from original documents. Therefore, to conduct a fair evaluation with other methods, we ignored the anaphora resolution in the experiment, and only make used of the original sentences. The purpose of utilizing SVMs as a tool is to simplify the classification of rhetorical relation. Most of previous works used fully annotated sentences, which limit the number of applicable data set. Thus, it is our main objective to achieve maximum performance regardless the number of annotated sentence so that our method can be applied to any data set in any language. Therefore, with further improvement mentioned above in classification of relations, we see that our system has promising potential to perform better summarization.

5. Conclusion

This paper presented a rhetorical relation based PageRank for multi-document summarization. Our system deals better with redundancy issue by modifying the connectivity of the sentences which successfully eliminates the redundancy problem. The most important feature is our system does not rely on fully annotated corpus and does not require deep linguistic knowledge. Future work will include (i) the improvement of rhetorical relations identification process, and (ii) expending the scope of summary generation to Web data.

References

[1] D.R. Radev, H. Jing, M. Sty, and D. Tam, "Centroid-based summarization of multiple documents," Inf. Process. Manage., vol.40, pp.919–938, 2004.

[2] H. Bhandari, M. Shimbo, T. Ito, and Y. Matsumoto, "Generic text summarization using probabilistic latent semantic indexing," 3rd International Joint Conference on Natural Language Processing, Hyderabad, India, Jan. 2008.

[3] Z. Zhang, S. Blair-Goldensohn, and D.R. Radev, "Towards CST-enhanced summarization," AAAI 2002, Aug. 2002.

[4] M.L.C. Jorge and T.S. Pardo, Experiments with CST-based Multi-document Summarization Workshop on Graph-based Methods for Natural Language Processing, ACL 2010, Uppsala, Sweden, pp.74–82, July 2010.

[5] Banko, Michele, Mittal, Vibhu O., and Witbrock, Michael J., "Headline generation based on statistical translation," ACL-2000, Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, Hong Kong, Oct. 2000.

[6] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," Computer Networks and ISDN Systems, vol.30, no.1-7, 1998.

[7] R. Mihalcea, "Graph-based ranking algorithms for sentence extraction, applied to text summarization," Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, companion volume (ACL 2004), Barcelona, Spain, July 2004.

[8] S. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," IEEE Trans. Acoust. Speech Signal Process., vol.35, no.3, 1987.

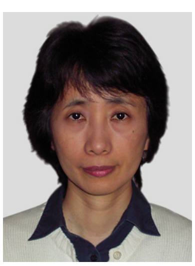
[9] D.R. Radev, J. Ottherbacher, and Z. Zhang, "CSTBank: A corpus for the study of cross-document structural relationship," Proceedings of Language Resource and Evaluation Conference (LREC'2004).

[10] V. Vapnik, "The Nature of Statistical Learning Theory," Springer, New York, 1995.

[11] D.R. Radev, J. Otterbacher, and Z. Zhang, CSTBank: Cross-document Structure Theory Bank,

http://tangra.si.umich.edu/clair/CSTBank/phase1.htm
[12] J.R. Finkel, T. Grenager, and C. Manning, "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling," Proc. 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp.363–370, 2005.
[13] J. Kupiec, J. Pedersen, and F. Chen, "A trainable document summarizer," Proc. 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, USA., pp.68–73, July 1995.
[14] D. Lin, "PRINCIPAR - An efficient, broad-coverage, principle-based parser," Proc. COLING'94, pp.482–488, 1994.
[15] G. Erkanand and D. Radev, "LexPageRank: Prestige in muli-document text summarization," Proc. EMNLP, 2004.
[16] D. Mimno, W. Li, and A. Mccallum, "Mixtures of hierarchical topics with pachinko allocation," Proc. 21st International Conference on Machine Learning, 2007.
[17] A. Celikyilmaz and D. Hakkani-tur, "Discovery of topically coherent sentences for extractive summarization," Proc. 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011), pp.491–499, 2011.
[18] L. Qiu, M.-Y. Kan, and T.-S. Chua, "A public reference implementation of the RAP anaphora resolution algorithm," Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004), vol.I, pp.291–294, 2004.

Nik Adilah Hanin Binti Zahri graduated from Computer Science and Media Engineering, University of Yamanashi in 2006. She received MSc from University of Yamanashi in 2009 and now doing doctoral course in the Human Environmental Medical Engineering, Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi, Japan.



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, and IPSJ.

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.