Exploiting Discourse Relations between Sentences for Text Clustering

*Nik Adilah Hanin Zahri**
Fumiyo Fukumoto
Suguru Matsuyoshi
Interdisciplinary Graduate School of Medicine and Engineering,
University of Yamanashi, Japan
g09dhl03*,fukumoto,sugurum@yamanashi.ac.jp

ABSTRACT

Over the years, the usage of discourse relations has been proven to enhance many applications such as text summarization, question answering and natural language generation. This paper proposes an approach that expands the benefit of discourse relations for natural language processing from a different aspect. We exploit the discourse relations existing between sentences to generate clusters of similar sentences from document sets. We first examined and defined the type of discourse relations that useful to retrieve sentences with identical content. We then assigned these relations to each sentence pair using a machine learning method. Finally we performed discourse relation-based clustering algorithm to generate clusters of similar sentences. We evaluated our method by measuring the cohesion and separation of the clusters and compared to a well recognized clustering method. The experimental result shows that our method performed significantly well, which demonstrated that discourse relation between sentences can be exploited for text clustering.

KEYWORDS : discourse relation, rhetorical relation, text clustering, SVMs, cluster validation

Proceedings of the Workshop on Advances in Discourse Analysis and its Computational Aspects (ADACA), pages 17–32,
COLING 2012, Mumbai, December 2012.

1 Introduction

The massive amount of data growth each day has become motivation for many researchers to develop text processing system with the ability to comprehend and process data effectively. The interpretation of how the phrases, clauses, and texts relate to each other is crucial to retrieve relevant information from texts. Therefore, the knowledge of discourse relation is prominent for natural language processing.

Many discourse coherent structures have been proposed over the years, such as Rhetorical Structure Theory (RST) (Mann and Thompson, 1987), RST Treebank (Carlson *et al*., 2002), Lexicalized Tree-Adjoining Grammar based discourse (Webber *et al.*, 2003), Cross-document Structure Theory (CST) (Radev *et al.*, 2004), and Discourse GraphBank (Wolf and Gibson, 2005, 2006). Discourse GraphBank represents discourse relation as graph structure, while other works represent them as hierarchical structure between textual units. Each work proposed different kind of methods to distinguish how events in text are related by identifying the transition point of a relation from one text span to another. Here, similar to the TDT project, an event refers to something that occurs at a specific place and time associated with some specific actions. This gives system abilities to detect important information or content within the text spans. For instance, the following example describes "*Evidence*" relation between texts proposed by RST.

Example 1:
S_1: Smokes billows from the Pirelli building.
S_2: The Pirelli Building in Milan, Italy, was hit by a small plane.

S1 describes an event (claim), while *S2* describes the information to increase the reader's belief, which is the evidence of why the event occurred. This relation indicates that information in *S2* is necessary for *S1* to take place. Consider another example of discourse relation from different structure. The following sentences describe "*Subsumption*" relation defined by CST.

Example 2:
S_3: Police were trying to keep people away, and many ambulances were at the scene.
S_4: Police and ambulance were at the scene.

CST defined sentences with *Subsumption* relation as having the same content along with additional facts in one sentence compared to another. From this example, *Subsumption* indicates that the content conveyed by *S4* is alternatively can be expressed in *S3* with more information.

We found that discourse relation between sentences not only indicates how two sentences are connected to each other, but also shows the amount of similar contents in both sentences. Relations such as *Identical* (defined in many discourse structures), *Subsumption* (CST), and *Generalization* (RST), links two text span in different way, however, provides identical information regarding the corresponding event. For instance, we observed that the same information can be extracted from *Subsumption* in *Example 2*, where both sentences indicate that police and ambulance were at the scene.

Therefore, we are motivated to explore the potential of discourse relation further more. By exploiting discourse relation between text spans, we believe that clusters of similar sentences can be constructed. We propose a method that establishes the benefit of discourse relation in generating cluster of similar sentences. Our main objective is to expand the usage of discourse relation to data mining in natural language processing. In addition, we also hope to explore the

construction of text clustering based on user preference, where users can determine how much similarity of information allowed in a text cluster according to the type of discourse relations used during clustering, which is difficult to achieve only with lexical and syntactic features of the sentences. For instance, clustering of sentences with *Identity* relation would only allow sentences with the exact same information within a cluster, while sentences with *Overlap* would include sentences with partial overlapping information within a cluster.

Our method consists of three main steps. We first define discourse relations which are useful for text clustering. Then, we identify these relations using Support Vector Machine (SVMs) (Vapnik, 1995). Finally, we performed a discourse relation based clustering algorithm to create clusters of similar sentences. Next section provides an overview of the existing works regarding discourse relation. Section 3 describes the framework of our system. In Section 4, we report experimental results and conclude our discussion with some direction for further works.

2 Previous Work

Since large scale machine readable textual corpus has become available, many techniques have been proposed to harvest vital information from documents using discourse relations analysis. Up until now, discourse relations have benefit various NLP applications such as text summarization ((Marcu, 1997), (Zhang *et al.*, 2002), (Radev *et al.*, 2004), (Uzêda *et al.*, 2009), (Louis *et al.*, 2012)), question answering ((Litkowski, 2002), (Verbe and Oostdijk, 2007)) and natural language generation ((Theune, 2002), (Piwek *et al.*, 2010)).

In text summarization, discourse relations are used to produce optimum ordering of sentences in a document, and remove redundancy from generated summaries. One of the well known works is CST based text summarization (Zhang *et al.*, 2002). In this work, sentences with most relations in the documents are considered to be important. They proposed an enhancement of text summarization by replacing low-salience sentences with sentences having maximum numbers of CST relations. Another work, (Uzêda *et al.*, 2009) presents comparative evaluation of RST-based text summarization methods. Besides informativeness, they also examined the effect of summary characteristics such as coherence and cohesion against each RST methods. One of the most recent work is a deep knowledge summarizer system (Jorge, 2010), which ranks input sentences according to the number of CST relations existing between sentences in accordance with user preference. They also demonstrated the effectiveness of redundancy elimination in summary using discourse relations. Most of the CST-based work observed the effects of individual CST relationships to the summary generation, and focused on the user preference based summarization, which requires manually annotated corpus.

The relevance of discourse analysis in QA application is pointed out by (Litkowski, 2002). This approach makes use of structural information of sentences, *e.g.*, discourse entities, semantic relation to generate database for question answering system. Another work, (Verbene *et al.*, 2007) suggested that the propositions of a question topic and answer are both represented by a text span in document, where the connection between text spans are described by RST relation. The topic of text span that matches RST tree will be the answer to the wh*y*-question.

Many of the previous works mentioned in the above show that the information obtained by discourse relation can improve single or multi-document summarization and QA application. In contrast, our work has different objective and approach. We investigated the potential of discourse relation in retrieving similar sentences, *i.e*. text clustering for data mining.

3 Framework

3.1 Redefinition of Discourse Relations

Different work proposed different types and definitions of discourse relations. Since our objective is to retrieve sentences with similar content using discourse relation, discourse structure that defines discourse relation between two text spans is mostly appropriate. Therefore, in this paper, we adopted the definition of rhetorical relation by CST (Radev *et al.*, 2004). We examined the definition of 18 types of CST relations in order to select relevant rhetorical relations for this work. According to the definition by CST, some of the relationship presents similar surface characteristics. Except for different version of event description, relations such as *Paraphrase*, *Modality* and *Attribution* share similar characteristic of information content with *Identity*. Consider the following example:

Example 3:
S_5: *RAI state TV reported that the pilot said the SOS was because of engine trouble.*
S_6: *RAI state TV reported that the pilot said he was experiencing engine trouble.*

Both sentences demonstrate an example of sentence pair that can represent *Identity, Paraphrase*, *Modality* and *Attribution* relations. The quality and 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 4:
S_7: *The crash put a hole in the 25th floor of the Pirelli building, and smoke was seen pouring from the opening.*
S_8: *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 can see from *Example 5* that *Subsumption* and *Elaboration* also shares some similar characteristics.

Example 5:
S_9: *The building houses government offices and is next to the city's central train station.*
S_{10}: *The building houses the regional government offices, authorities said.*

Thus, 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.* Here, *Subsumption* denotes S_2 as the subset of S_1, but as for *Elaboration*, S_2 is not necessary a subset of S_1. Therefore, we keep *Subsumption* and *Elaboration* as two different relations so that we can precisely perform the automated identification of discourse relation by using SVMs.

We redefined the definition of relations from CST by combining the relations types that resemble each other as described in *Example 3*, *4* and *5*. *Fulfillment* by CST refers to sentence pair which asserts the occurrence of predicted event, where overlapped information present in both sentences. Therefore, we combined *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. The combination of rhetorical relations in this paper is concluded in Table 1. We modified the definition of each relation in accordance with the combination of relations shown in Table 1. The taxonomy for rhetorical relations we used in the system is described in Table 2. By definition, although *Change of Topics*

Relations by CST	Relations by System
Identity, Paraphrase, Modality, Attribution	Identity
Subsumption, Indirect Speech, Citation	Subsumption
Elaboration, Follow-up	Elaboration
Overlap, Fulfillment	Overlap
Change of Perspective, Contradiction, Reader Profile	Change of Topics
Description, Historical Background,	Description
Translation, Summary	-
-	No Relations

TABLE 1 – Combination of CST relations

Relations	Definition
Identity	S_1 and S_2 contain the same information
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 and S_2 provides partial overlapping information
Change of Topics	S_1 and S_2 provide different facts about the same entity.
Description	S_1 gives historical or present description about any entity mentioned in S_2.
No Relations	No relation exits between S_1 and S_2.

TABLE 2 – Redefinition of discourse relations

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 discourse relation in text clustering.

3.2 Determining Discourse Relations Using SVMs

To identify discourse relations, we used a machine learning approach, Support Vector Machine (SVMs) (Vapnik, 1995). We used CST-annotated sentences pair obtained from CST Bank (Radev *et al.*, 2004) 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 specific information, *e.g.* location and time of the event. We provided the 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

The amount of overlapping information among sentences is important to determine the type of discourse relations exist between them. Here, we used a few similarity measurements to compute the similarity between word content in both sentences from different aspects. We defined nouns, verbs and adjectives as word content in the experiment.

1. Cosine Similarity

   We compute the similarity of both sentences using cosine similarity measurement, defined as follows:

$$\cos(S_1, S_2) = \frac{\sum s_{1,i} * s_{2,i}}{\sqrt{\sum (s_{1,i})^2} * \sqrt{\sum (s_{2,i})^2}} \quad (1)$$

   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. We observed the following 5 types of similarity in this experiment:

   i) Similarity between word contents
   ii) Similarity of *nouns* tokens
   iii) Similarity of *verbs* tokens
   iv) Similarity of *adjectives* tokens
   v) Similarity of bigram words

   We not only measure the similarity value of words, but also consider the similarity value of word sequence in (v). We found that different word sequence sometimes provides different meaning. For example, the word "*test driving*" and "*driving test*". The word "*test driving*" refers to the action of driving a vehicle in order to evaluate its performance, meanwhile "*driving test*" refers to procedure designed to test a person's ability to drive a motor vehicle. The words ordering indirectly determine the semantic meaning in sentences. Therefore, we included the similarity of bigram words in the measurement.

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:

$$wol(S_1) = \frac{\#commonwords(S_1, S_2)}{\#words(S_1)} \times 2 \quad (2)$$

$$wol(S_2) = \frac{\#commonwords(S_1, S_2)}{\#words(S_2)} \times 2 \quad (3)$$

   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.

3. Longest Common Substring

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

$$lcs(S_1) = \frac{Length(MaxComSubstring(S_1, S_2))}{Length(S_1)} \quad (4)$$

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 (Lin, 1994) 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:

$$Subj\_ove(S_1) = \frac{\#comSubj(S_1, S_2)}{\#Subj(S_1)} \quad (5)$$

$$Obj\_ove(S_1) = \frac{\#comObj(S_1, S_2)}{\#Obj(S_1)} \quad (6)$$

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.

$$SubjNoun\_ove(S_1) = \frac{\#com\ Subj(S_1)Noun(S_2)}{\#Subj(S_1)} \quad (7)$$

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

NER Class	FrameNet	
	No. Frames	Frame Examples
PERSON	9	People (*e.g. person, lady, boy, man, woman*) People_by_vocation (*e.g. police_officer, journalist*)
ORGANIZATION	9	Bussiness (*e.g. company, corporation, firm*) Organization (*e.g. governent , agency, comittee*)
LOCATION	12	Locale (*e.g. earth, region, site, gzone, place*) Relational_natural_features (*e.g. lake, mountain*)
TIME	2	Calenderic_unit (*e.g. morning, evening, noon, eve)* Location _in _time (*e.g. time*)
DATE	2	Calenderic_unit (*e.g. winter, spring, summer*) Natural features (*e.g. spring, fall*)
MONEY	1	Money (*e.g. money, cash, funds*)
PERCENT	-	-

TABLE 3 – Information adopted from FrameNet

Recognizer (Finkel *et al.*, 2005) 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:

S1: *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 (Fillmore *et al*. 2003). 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 3 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, 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.

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. The feature with higher number of entities is set to 1, and 0 for lower value.

3. Lengths of sentences

We defined the length of S_j as follows:

$$Len(S_j) = \sum_{w_i \in S_j} w \quad (8)$$

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 Discourse Relations based Clustering Algorithm

Connections between two sentences can be represented by multiple discourse relations. For instance, in some cases, sentences defined as *Subsumption* can also be define as *Identity*. As we proposed a method of cluster generation of similar sentences, 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 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 discourse relations assignment can be concluded as follows:

Identity > Subsumption > Elaboration > Overlap > Change of Topics > Description

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

i) Assign the strongest relations determined by SVMs to each connection (refer to Figure 1(a)).

ii) Suppose each sentence is a centroid of its own cluster. Identify sentences connected to the centroid as *Identity* (*ID*), *Subsumption* (*SUB*), *Elaboration* (*ELA*) and *Overlap* (*OVE*) relations[1]. Sentences with these connections are evaluated as having similar content, and aggregated as one cluster (refer Figure 1(b)).

iii) Remove similar clusters by retrieving centroids connected as *Identity*, *Subsumption* or *Elaboration*.

iv) Merge the clusters from (iii) to minimize the occurrence of the same sentences in multiple clusters (refer Figure 1(c)).

v) Iterate step (iii) and (iv) until the number of clusters is convergence.

[1] We performed 2 types of text clustering, which includes and excludes *Overlap*

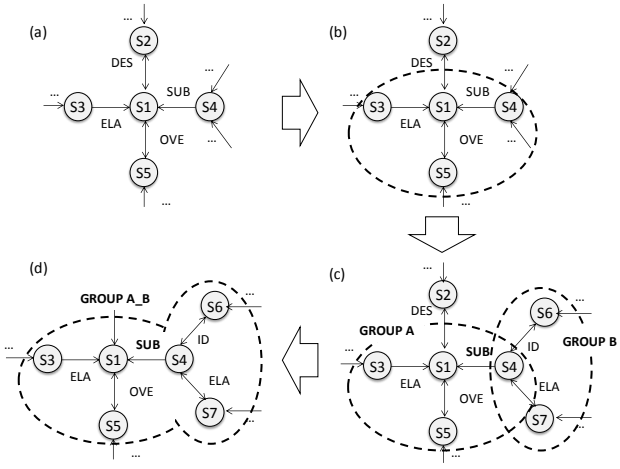

FIGURE 1 – Clustering algorithm based on discourse relations.

4 Experiment

4.1 Data

CST-annotated sentences pairs are obtained from publicly available data set from Cross-document Structure Theory Bank (Radev *et al.*, 2004) and were combined into relations according to Table 2. We used 218 sentence pairs of *Identity*, 317 pairs of *Subsumption*, 58 pairs of *Elaboration*, 157 pairs of *Overlap*, 348 pairs of *Change of Topics*, 70 pairs of *Description* and 120 pairs of *No Relations*. Our system is evaluated using 2 data sets from Document Understanding Conference, which are DUC'2001 and DUC'2002. DUC'2001 and DUC'2002 provided 30 and 59 document sets consisting 10,412 and 14,790 sentences, respectively. We used Brill's Tagger (Brill, 1992) to POS-tag the sentences, and extracted content words and lemmas of the words.

4.2 Result and Discussion

4.2.1 Discourse Relation Identification

The discourse relations assigned between sentences by SVMs is 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 (DUC'2001 and DUC'2002). The human judges performed manual annotation against sentence pairs, and assessed if SVMs assigned the correct discourse relation to each pair. The correct discourse relation refers to either one of the discourse relations assigned by human judges in case of multiple relations exist between the two sentences. We also assigned the most frequent relations

Relations	DUC'2001			DUC'2002		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Baseline	0.112	0.946	0.197	0.144	0.855	0.241
Identity	0.983	1.000	0.991	0.855	1.000	0.921
Subsumption	0.688	0.985	0.804	0.685	0.900	0.773
Elaboration	0.650	0.952	0.768	0.644	0.902	0.737
Overlap	0.776	0.652	0.703	0.740	0.694	0.715
Change of Topics	0.553	0.701	0.614	0.611	0.593	0.597
Description	0.797	0.947	0.853	0.818	0.856	0.828
No Relations	0.969	0.556	0.697	0.985	0.652	0.782

TABLE 4 – Evaluation result for classification of discourse relations

to all sentence pairs as a baseline method. We used the precision, recall and F-measure score as an evaluation measure.

Table 4 shows the macro average of precision, recall and F-measure for DUC'2001 and DUC'2002**.** Evaluation results from Table 4 indicates that SVMs works well for the classification of *Identity*, *Subsumption*, *Elaboration* and *Overlap*, where the F-measure values achieved are above 70% for both data sets. In contrast, the F-measure value of *Change of Topics* shows an average result due to lack of significant characteristics which caused false positive result for sentence pairs with no relation. The following sentence pair shows the example of false positive result of *Change of Topics.*

S_{11} *: Boston have skyline, 2 1/2 miles in the distance, can seem so far away.*
S_{12} *: Though an interpreter, Martinez said he started out running 5:15 or 5:20 miles.*

The examples show that the *subject of the verb* in both sentences is different and both sentences semantically represent no relation with each other. Consider another example:

S_{13} *: The eight day trip will leave from Chicago and will include sightseeing, guided runs and fun run from Malahide Castle to Swords.*
S_{14} *: I had to have patience and run from the back.*

Both sentences were identified as *Overlap* by SVMs while there is no relation present between the sentences. As a result, the low recall value affected the F-measure of *No Relations*. Overall, classification by SVMs shows that our method outperformed over the baseline method, where our system achieved more than 60% accuracy for most relations even though we only consider surface characteristics from sentence pairs during classification.

4.2.2 Discourse Relation-based Clustering

We evaluated our method by measuring the cohesion and separation of the constructed clusters (Raskutti and Leckie, 1999) (IBM SPSS Statistics, 2011). The cluster cohesion refers to how closely the sentences are related within a cluster, measured using *Sum of Squared Errors* (*SSE)*;

$$AverageSSE = \frac{1}{N}\sum_{i}\sum_{x \in C_i} sim(x, m_i)^2 \quad (9)$$

where *sim (x,m$_i$)* refers to the similarity of sentence *x* with other members in the same cluster, *m$_i$* and *N* denotes the number of clusters. The smaller value of *SSE* indicates that the sentences in clusters are closer to each other. Meanwhile, cluster separation refers to how distinct or well-separated a cluster from others, measured using *Sum of Squares Between* (*SSB*);

$$AverageSSB = \frac{1}{N}\sum_i |C_i| sim(m, m_i)^2 \quad (10)$$

where *sim(m,m$_i$)* refers the similarity between sentences from the corresponding cluster with sentences outside the cluster, $|C_i|$ is the size of cluster and *N* is the number of clusters. 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*) to measure the harmonic mean of both cohesion and separation of the clusters (Kaufman and Rousseeuw, 1990) (IBM SPSS Statistics, 2011) by using Equation (11);

$$AverageSC = \frac{1}{N}(1-\frac{a}{b}) \quad if\ a < b \ \ or,$$
$$= \frac{1}{N}(\frac{b}{a}-1) \quad if\ a \geq b \quad (11)$$

where *a* is the average similarity of sentence *i* with other members in the cluster, and *b* is the minimum average distance of sentence *i* with sentences outside the cluster and *N* is the number of clusters. The value range of the *Silhouette Coefficient* is between 0 and 1, where the value closer to 1 is the better.

Table 5 shows the evaluation results of text clustering. *Method1* refers to the clusters constructed by *Identity*, *Subsumption* and *Elaboration*, while *Method2* refers to the clusters constructed by *Identity*, *Subsumption*, *Elaboration* and *Overlap*. We also used *K-Means* clustering for comparison. *K-means* iteratively reassigns sentences to the closest clusters until a convergence criterion is met (McQueen, 1967). Evaluation results indicate that *Method1,* 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, *Method2* 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). *Method2* generated bigger clusters, therefore resulted wider separation from other clusters. Overall, the average of *Silhouette Coefficient* shows that our method, *Method1* (0.628 for DUC'2001 and 0.639 for DUC'2002) and *Method2* (0.652 for DUC'2001 and 0.636 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 clustered sentences by using a pair-wise evaluation measure, where we sampled 5 sets of data consisting 100 sentences pairs and evaluated if both sentences are actually belong to the same clusters. Table 6 shows the macro average Precision, Recall and F-measure for pair-wise evaluation. *Method1*, which excludes *Overlap* relation during clustering, demonstrated a lower Recall value compared to *Method2* and *K-Means*. However, the Precision score of *Method1* indicates better performance compared to *K-Means*. Overall, *Method2* obtained the best value for all measurement compared to *Method1* and *K-Means* for both data sets. We achieved optimum pair-wise results by including *Overlap* during clustering, where the F-measure

Clustering Method	DUC'2001			DUC'2002		
	Average SSE	Average SSB	Average SC	Average SSE	Average BSS	Average SC
K-Means	7.271	209.111	0.512	6.991	154.511	0.510
Method1 (ID, SUB, ELA)	4.181	308.153	0.628	3.624	214.762	0.639
Method2 (ID, SUB, ELA,OVE)	4.599	397.237	0.652	3.927	257.118	0.636

TABLE 5 –Evaluation result for cohesion and separation of the clusters

Clustering Method	DUC'2001			DUC'2002		
	Precision	Recall	F-measure	Precision	Recall	F-measure
K-Means	0.577	0.898	0.702	0.603	0.885	0.716
Method1 (ID, SUB, ELA)	0.805	0.590	0.678	0.750	0.533	0.623
Method2 (ID, SUB, ELA,OVE)	0.783	0.758	0.770	0.779	0.752	0.766

TABLE 6 – Evaluation result for pair-wise

obtained for DUC'2001 and DUC'2002 are 0.770 and 0.766, respectively.

We can see from Table 5 and Table 6 that the connection between sentences can allow text clustering according to the user preference. For instance, sentences with *Identity*, *Subsumption* and *Elaboration* were classified into a small group without overlapping with other clusters. In contrast, sentences with *Identity*, *Subsumption*, *Elaboration* and *Overlap* allow minimum information overlapping between clusters. Thus, the experimental results demonstrate that the utilization of discourse relation can be another alternative of cluster construction other than observing word distribution in corpus.

Conclusion and perspectives

This paper explored the benefits of discourse relation in data mining. The evaluation results showed that the discourse relation-based method has promising potential as a novel approach for text clustering. Our method is capable to offer various kind of text clustering, such as clustering of only identical or overlapping sentences. In future, addition of other types of relations, *e.g.*, *Attribution* (from CST) can be used to perform clustering of attributed information from corpus. Previously, discourse relation has been used to remove redundancy from generated summaries, thus, sentence clustering based on discourse relations will definitely benefits text summarization for multiple documents. Our future works will include (i) the investigation of more discourse relations for text clustering, (ii) to improve the classification of discourse relations, and (iii) the application of discourse relation-based clustering to text summarization.

References

Mann, W.C. and Thompson, S.A., "Discourse Structure Theory: A Theory of Text Organization", Technical Report ISI/RS-87-190, ISI, Los Angeles, California, 1987.

Carlson, L., Marcu, D. and Okurowski, M.E., "RST Discourse Treebank", Linguistic Data Consortium 1-58563-223-6, 2002.

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

Radev, D.R., Ottherbacher, J. and Zhang, Z., "CSTBank: A Corpus for the Study of Cross-document Structural Relationship", In Proc. of Language Resource and Evaluation Conference (LREC), 2004.

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

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

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.

Radev, D.R., Jing, H., Sty,M., and Tam, D.,"Centroid-based Summarization of Multiple Documents", Inf. Process.Manage.(40), pp. 919-938, 2004.

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.

Uzêda, 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.

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.

Litkowski, K., "CL Research Experiments in TREC-10 Question Answering", The 10th Text Retrieval Conference (TREC 2001). NIST Special Publication, pp. 200-250, 2002.

Verberne, S., Boves, L., and Oostdijk, N., "Discourse-based Answering of *Why*-Questions", Traitement Automatique des Langues, special issue on Computational Approaches to Discourse and Document Processing, pp. 21-41, 2007.

Theune, M., "Contrast in Concept-to-speech Generation", Computer Speech & Language,16(3-4), ISSN 0885-2308, pp. 491-530, 2002.

Piwek, P. and Stoyanchev, S., " Generating Expository Dialogue from Monologue Motivation, Corpus and Preliminary Rules", In Proc. of 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2010.

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.

Lin, D., "PRINCIPAR- An Efficient, Broad-coverage, Principle-based Parser", In Proc. of 15th International Conference on Computational Linguistics (COLING), pg.482-488, 1994.

Finkel, J.R., Grenager, T. and Manning, C., "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling", In Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL), pp. 363-370, 2005.

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.

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

Brill, E., "A Simple Rule-based Part-of-Speech Tagger", In Proc. of 3rd Conference on Applied Natural Language Processing, pp. 152-155, 1992.

Raskutti, B. and C. Leckie, "An Evaluation of Criteria for Measuring the Quality of Clusters", In Proc. of the 16th International Joint Conference on Artificial Intelligence, ISBN:1-55860-613-0, pp: 905-910,1999.

IBM SPSS Statistic Database, "Cluster Evaluation Algorithm" http://publib.boulder.ibm.com/ , 2011

Kaufman, L. and Rousseeuw, P., "Finding Groups in Data: An Introduction to Cluster Analysis", John Wiley and Sons, London. ISBN: 10: 0471878766, 1990.

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.