Pertanika J. Sci. & Technol. 21 (1): 239 - 246 (2013)

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Using SVMs for Classification of Cross-Document Relationships

Yogan Jaya Kumar[1,2]*, Naomie Salim[2], Ahmed Hamza Osman[2] and Albaraa Abuobieda[2]

[1]Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, 76100 Durian Tunggal, Melaka, Malaysia

[2]Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia

ABSTRACT

Cross-document Structure Theory (CST) has recently been proposed to facilitate tasks related to multi-document analysis. Classifying and identifying the CST relationships between sentences across topically related documents have since been proven as necessary. However, there have not been sufficient studies presented in literature to automatically identify these CST relationships. In this study, a supervised machine learning technique, i.e. Support Vector Machines (SVMs), was applied to identify four types of CST relationships, namely "Identity", "Overlap", "Subsumption", and "Description" on the datasets obtained from CSTBank corpus. The performance of the SVMs classification was measured using Precision, Recall and F-measure. In addition, the results obtained using SVMs were also compared with those from the previous literature using boosting classification algorithm. It was found that SVMs yielded better results in classifying the four CST relationships.

Keywords: CST relation, multi-document, rhetorical relation, SVMs

INTRODUCTION

Discourse analysis in texts has nowadays become very prominent, especially when it involves multiple texts such as documents. The idea of cross-document structural relationship is to investigate the existence of inter-document rhetorical relationships. These rhetorical relations are based on the CST model (Cross-document Structure Theory) (Radev, 2000). Documents which are related to the same topic usually contain semantically related textual units. These textual units can be words, phrases, sentences, or the documents itself. The general schema of CST is shown in Fig.1. In the current work, only the semantic relations between sentences were taken into consideration. Some examples of such semantic connections are "Identity",

Article history:
Received: 31 March 2012
Accepted: 31 August 2012

E-mail addresses:
yogan@utem.edu.my (Yogan Jaya Kumar),
naomie@utm.my (Naomie Salim),
ahmedagraa@hotmail.com (Ahmed Hamza Osman),
albarraa@hotmail.com (Albaraa Abuobieda)
*Corresponding Author

ISSN: 0128-7680 © 2013 Universiti Putra Malaysia Press.

"Contradiction", "Description", and "Historical background". Table 1 shows some examples of the sentence pairs that hold CST relationship. Full descriptions of the CST relations are given in Radev (2000).

The work on CST can be put in line with Rhetorical Structure Theory (RST) (Taboada & Mann, 2006). The difference between these two theories is that RST aims to capture the rhetorical relation between span of adjacent text units, while CST goes across topically related documents to describe its rhetorical relation. In topically related documents, especially news articles, the information contents are closely connected even though the news story comes from various sources. By referring to the description of CST relations shown in Table 1, it can be seen that these types of relations are essential for the analysis of redundancy, complementarity and contradiction among different information sources. Thus, the ability to automatically identify the types of CST relationship will definitely be handy for tasks related to multi-document analysis. A number of research works have addressed the benefits of CST for summarization task (see for instance, Zhang *et al*., 2002; Jorge *et al*., 2010). Nonetheless, a major limitation of these works is that the CST relationships need to be manually annotated by human expert. Human annotation is not only expensive, but it also consumes a lot of time.

There have been efforts put to learn the CST relationships in texts. Zhang *et al*. (2003) used boosting, i.e. a classification algorithm, to identify the presence of CST relationships between sentences. It is an adaptive algorithm which works by iteratively learning previous weak classifiers and adding them to a final strong classifier. The authors experimented with CST annotated article cluster that built the CSTBank corpus. Hence, lexical, syntactic and semantic features were used for data representation. Their classifier was able to identify sentence pairs with no relationship very well, but showed a rather poor performance in classifying the other types of CST relationship.

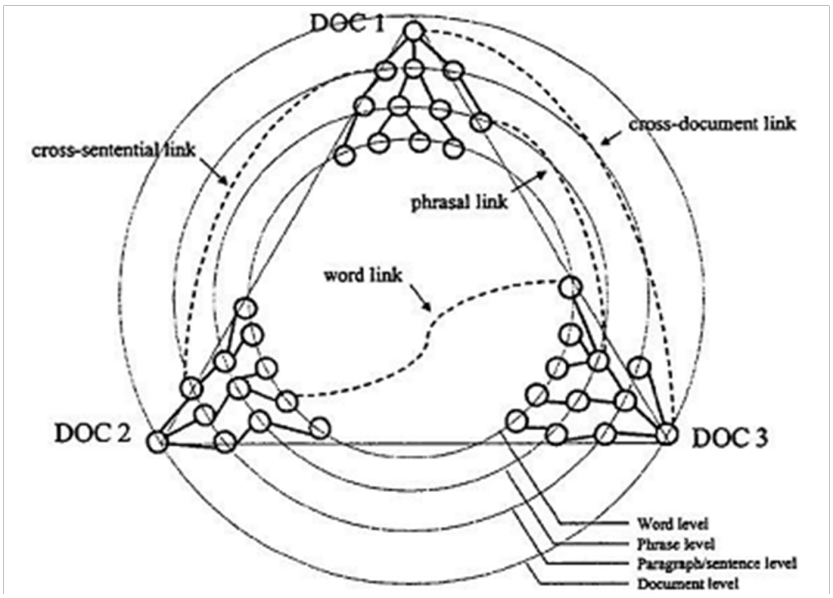

Fig.1: CST general schema (Radev, 2000)

TABLE 1: Some examples of the CST relationship between sentences (source: Zhang *et al*., 2002)

Relationship	Description	Text span 1 (S1)	Text span 2 (S2)
Identity	The same text appears in more than one location	Tony Blair was elected for a second term today.	Tony Blair was elected for a second term today.
Equivalence	Two text spans have the same information content	Derek Bell is experiencing resurgen0ce in his career.	Derek Bell is having a "comeback year."
Translation	The same information content in different languages	Shouts of "Viva la revolucion!" echoed through the night.	The rebels could be heard shouting, "Long live the revolution".
Subsumption	S1 contains all information in S2, plus additional information not in S2	With 3 wins this year, Green Bay has the best record in the NFL.	Green Bay has 3 wins this year.
Contradiction	Conflicting information	There were 122 people on the downed plane.	126 people were aboard the plane.
Historical Background	S1 gives historical context to information in S2	This was the fourth time a member of the Royal Family has gotten divorced.	The Duke of Windsor was divorced from the Duchess of Windsor yesterday.

In another related work, Miyabe *et al*. (2008) investigated on the identification of CST relationship types by using cluster-wise classification with SVM classifier. They used a Japanese cross-document relation corpus annotated with CST relationships. The authors proposed using the detected "Equivalence" relations to address the task of "Transition" identification. In particular, similarity through the variable noun phrases was used for transition identification. They obtained F-measure of 75.50% for equivalence and 45.64% for transition. However, their approach is only limited to the two aforementioned relations.

Closely related to our work is the approach by Zahri and Fukumoto (2011). The authors determined five types of CST relation between sentences using SVMs. The authors computed the lexical features between sentence pairs using the dataset from the CSTBank corpus. Then, they used the identified CST relations to determine the directionality between the sentences for PageRank (Erkan & Radev, 2004) computation. However there were no experimental results specifically shown on the performance of their CST relationship classification. This is essential because the performance of the classification has direct implication on the final results of the system.

METHODS

Relying on manually annotated text for CST relationship identification consumes a lot of time and resources. Thus, it is favourable to have a system which can automatically identify the existence of the CST relations between pairs of sentences. However, at this point of time, we are only considering four types of CST relations, namely "Identity", "Overlap", "Subsumption", and "Description". More details of these relations are given in Table 2. Meanwhile, further details with examples can be found in Zhang *et al*. (2002).

TABLE 2: The CST relations used in this work

Relationship	Description
Identity	The same text appears in more than one location
Subsumption	S1 contains all information in S2, plus additional information not in S2
Description	S1 describes an entity mentioned in S2
Overlap (partial equivalence)	S1 provides facts X and Y while S2 provides facts X and Z; X, Y, and Z should all be non-trivial

In this study, the publically available CSTBank corpus was exploited (Radev *et al*., 2003) – , i.e. a corpus consisting clusters of English news articles annotated with the CST relationships. Using the datasets from CSTBank, we were able to obtain our training and testing data. Then, the training set comprising of the features between sentence pairs with its corresponding CST relationship was prepared. Each of these sentence pairs was represented using four lexical features which could be useful to differentiate the CST relationships between the sentences. After that 100 pairs of sentences that posed no CST relations were manually selected for the training and test data. The lexical features that were computed for each sentences pair are described in the following.

Cosine similarity – cosine similarity is used to measure how similar two sentences are. Here, the sentences are represented as word vectors having words with tf-idf as the element *value:*

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

Word overlap – this feature represents the measure on the numbers of words overlapping in the two sentences (after stemming process). This measure is not sensitive to the word order in the sentences:

$$overlap(S_1, S_2) = \frac{\#words(S_1 \cap S_2)}{\#words(S_1 \cup S_2)} \quad (2)$$

Length difference – length difference gives the measure of difference between the lengths of two sentences. It shows how long or how short a sentence is compared to the other:

$$lengdiff(S_1, S_2) = length(S_1) - length(S_2) \quad (3)$$

Length type of S_1 – this feature gives the length type of the first sentence when the lengths of two sentences are compared:

$$lengtype(S_1) = 1 \quad \text{if} \quad length(S_1) > length(S_2),$$
$$-1 \quad \text{if} \quad length(S_1) < length(S_2),$$
$$0 \quad \text{if} \quad length(S_1) = length(S_2) \tag{4}$$

Support Vector Machines (SVMs) (Vapnik, 1995), a supervised machine learning technique commonly used for classification and regression analysis, was employed in this work. SVMs are feature-based classifiers, where each instance from the datasets is usually represented as a feature vector which is then used as an input for machine learning. An excellent introduction to SVMs can be found in Cristianini and Taylor (2000). SVMs are basically two-class classifiers. A support vector machine builds a hyperplane that separates the instances of the two classes. Since ours is a multi-class problem, SVMs built a set of one-versus-one classifiers, and chose the class that is selected by most classifiers. The general flow of the classification process is shown in Fig.2.

Based on the dataset from CSTBank, a total of 477 sentence pairs were selected for the training and 205 sentence pairs were used for the testing. These included the sample of 100 sentence pairs with no CST relationship. First of all, the texts were preprocessed by stop-word filtering and word stemming. After computing each of the feature values for every sentence pair from the training set, they were used as inputs for the training of SVMs. The training data were then trained using the LibSVM tool (Chang & Lin, 2011) on MATLAB. The SVMs model best parameters were chosen after applying 5-folds cross validation. Once the training was completed, the resulting classifier model was tested by using the test data to measure its performance. The performance of SVMs classification was evaluated using Precision, Recall and F-measure.

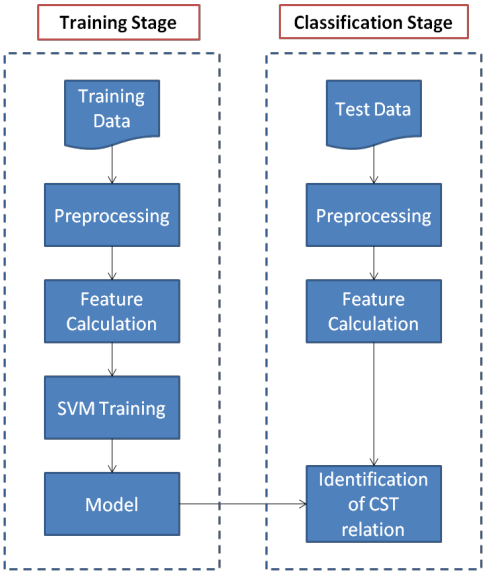

Fig.2: The training and classification processes

RESULTS AND DISCUSSION

Fig.3 shows the results of the classification in the present study. It can be observed that the SVM achieved a high precision and recall in determining the relation "Identity" as compared to the rest of the CST relationships. To detect "No relation" sentence pairs, the precision score was good but its recall was below average. The rest of the relations obtained average results. One possible reason for getting imbalance classification results was probably the features chosen for the SVM training. Since the features used in this work are only of lexical type, SVM might not well differentiate most of the relation types. Thus, it can be assumed that the performance of the classifier can be further improved by selecting better features for training.

With the motivation to observe the general performance of the SVM classification in identifying the CST relationship between sentences, the initial results retrieved were also compared with those obtained by Zhang *et al*. (2003) who had applied boosting classification algorithm (BCA). Table 3 gives the precision, recall and F-measure, while Fig.4 shows the F-measure comparison between the two methods. It was observed that BCA performed well in differentiating non-CST related sentence pairs. However, SVM was found to outperform BCA in classifying the other types of CST relationships.

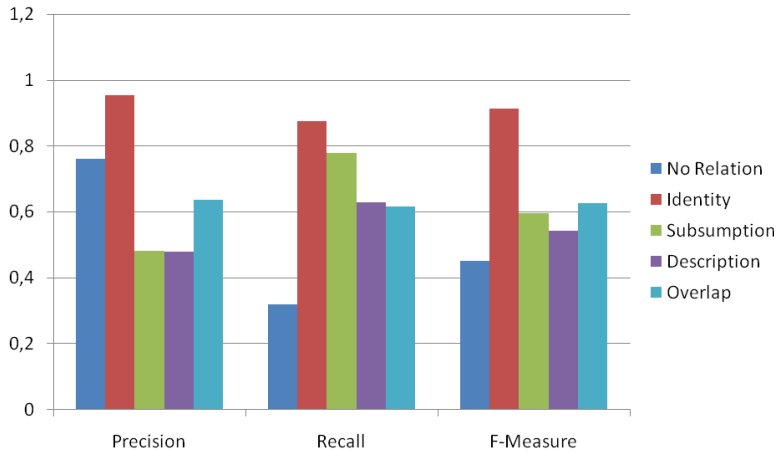

Fig.3: The classification by SVM

TABLE 3: A comparison of the results

Relationship	Precision		Recall		F-Measure	
	BCA	SVM	BCA	SVM	BCA	SVM
No relation	0.89	0.76	0.94	0.32	0.91	0.45
Identity	-	0.95	-	0.87	-	0.91
Subsumption	0.06	0.48	0.04	0.77	0.05	0.59
Description	0.26	0.47	0.21	0.62	0.23	0.54
Overlap	0.55	0.63	0.35	0.61	0.43	0.62

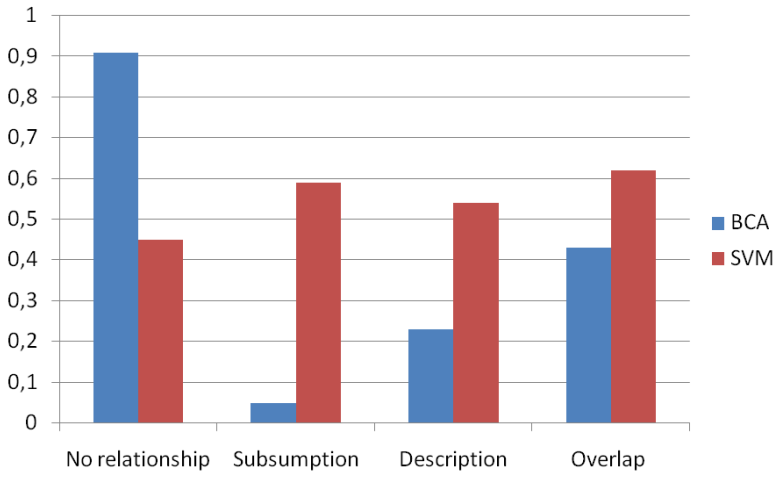

Fig.4: A comparison of the F-measures between BCA and SVM

CONCLUSION

The feasibility to identify the CST relationships between the sentences across topically related documents is a noteworthy advancement for multi-document analysis. However, there have not been sufficient studies presented in the literature to automatically identify these CST relationships. In this paper, an attempt to investigate the performance of SVMs classification technique for identifying four types of CST relations (namely, "Identity", "Overlap", "Subsumption", and "Description") was carried out. In order to achieve this, the publically available CSTBank corpus (i.e. the corpus with human annotated CST relationships) was exploited so as to obtain the required training and testing data. Each instance of these datasets was represented using the lexical features.

The performance of the SVMs classification was evaluated using Precision, Recall and F-measure. The experimental results showed that we were able to detect "Identity" relation very well and also produced average results for the other types of relations. Moreover, the results obtained using SVMs were also compared with those retrieved from the previous literature using boosting classification algorithm. It was observed that overall, the SVMs classification yields better results. Currently, the authors are investigating further into improving the performance of the classifier by proposing additional semantic features such as noun phrases, verb phrases, etc. for feature vector representation. By being able to produce better classification accuracy, it can be stated that many applications related to multi document analysis will gain benefit out of it.

ACKNOWLEDGEMENTS

The research was sponsored by IDF and the Ministry of Science Technology and Innovation, under the University's research grant vote number 01H74, Universiti Teknologi Malaysia.

REFERENCES

Chang, C. C., & Lin C. J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2*27(1-27), 27.

Cristianini, N., & Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge: Cambridge University Press.

Erkan, G., & Radev D. R. (2004). LexPageRank: Prestige in multi-document text summarization. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 365-371.

Jorge, M. L. C., & Pardo, T. S. (2010). Experiments with CST-based Multidocument Summarization. *Workshop on Graph-based Methods for Natural Language Processing, ACL*, 74–82.

Miyabe, Y., Takamura, H., & Okumura, M. (2008). Identifying cross-document relations between sentences. *In Proceedings of the 3rd International Joint Conference on Natural Language Processing*, 141–148.

Radev, D. R. (2000). *A Common Theory of Information Fusion from Multiple Text Sources Step One: Cross-Document Structure.* Paper presented at the 1st ACL SIGDIAL Workshop on Discourse and Dialogue, 10, pp.74-83.

Radev, D. R., & Otterbacher, J. (2003). *CSTBank Phase I*. Retrieved at http://tangra.si.umich.edu/clair/CSTBank/phase1.htm

Taboada, M., & Mann, W. C. (2006). Rhetorical Structure Theory: Looking Back and Moving Ahead. *Discourse Studies, 8*, 423-459.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York: Springer Verlag.

Zahri, N. A. H .B., & Fukumoto, F. (2011). *Multi-document Summarization Using Link Analysis Based on Rhetorical Relations between Sentences.* Paper presented at the 12th International Conference on Computational Linguistics and Intelligent Text Processing, 2, pp.328-338.

Zhang, Z., Blair-Goldensohn, S., & Radev, D. R. (2002). *Towards CST-Enhanced Summarization*. Paper presented at the 18th National Conference on Artificial Intelligence, pp.439-446.

Zhang, Z., Otterbacher, J., & Radev, D. R. (2003). *Learning cross-document structural relationships using boosting.* Paper presented at the 12th International Conference on Information and Knowledge Management, pp.124-130.