Measuring the Effect of Discourse Relations on Blog Summarization

Shamima Mithun
Concordia University
Montreal, Quebec, Canada
shamima.mithun@gmail.com

Leila Kosseim
Concordia University
Montreal, Quebec, Canada
kosseim@encs.concordia.ca

Abstract

The work presented in this paper attempts to evaluate and quantify the use of discourse relations in the context of blog summarization and compare their use to more traditional and factual texts. Specifically, we measured the usefulness of 6 discourse relations - namely *comparison*, *contingency*, *illustration*, *attribution*, *topic-opinion*, and *attributive* for the task of text summarization from blogs. We have evaluated the effect of each relation using the TAC 2008 opinion summarization dataset and compared them with the results with the DUC 2007 dataset. The results show that in both textual genres, *contingency*, *comparison*, and *illustration* relations provide a significant improvement on summarization content; while *attribution*, *topic-opinion*, and *attributive* relations do not provide a consistent and significant improvement. These results indicate that, at least for summarization, discourse relations are just as useful for informal and affective texts as for more traditional news articles.

1 Introduction

It is widely accepted that in a coherent text, units should not be understood in isolation but in relation with each other through discourse relations that may or may not be explicitly marked. A text is not a linear combination of textual units but a hierarchial organized group of units placed together based on informational and intentional relations to one another. According to (Taboada, 2006), "Discourse relations - relations that hold together different parts (i.e. proposition, sentence, or paragraph) of the discourse - are partly responsible for the perceived coherence of a text". For example, in the sentence *"If you want the full Vista experience, you'll want a heavy system and graphics hardware, and lots of memory"*, the first and second clauses do not bear much meaning independently; but become more meaningful when we realize that they are related through the discourse relation *condition*.

Discourse relations have been found useful in many NLP applications such as natural language generation (e.g. (McKeown, 1985)) and news summarization (e.g. (Blair-Goldensohn and McKeown, 2006; Bosma, 2004)) to improve coherence and better simulate human writing. However, most of these work have been developed for formal, well-written and factual documents. Text available in the social media are typically written in a more casual style, are opinionated and speculative (Andreevskaia et al., 2007). Because of this, techniques developed for formal texts, such as news articles, often do not behave as well when dealing with informal documents. In particular, news articles are more uniform in style and structure; whereas blogs often do not exhibit a stereotypical discourse structure. As a result, for blogs, it is usually more difficult to identify and rank relevant units for summarization compared to news articles.

Several work have shown that discourse relations can improve the results of summarization in the case of factual texts or news articles (e.g. (Otterbacher et al., 2002)). However, to our knowledge no work has evaluated the usefulness of discourse relations for the summarization of informal and opinionated texts, as those found in the social media. In this paper, we consider the most frequent discourse relations found in blogs: namely *comparison*, *contingency*, *illustration*, *attribution*, *topic-opinion*, and *attributive* and evaluate the effect of each relation on informal text summarization using the Text Analysis Conference (TAC)

International Joint Conference on Natural Language Processing, pages 1401–1409,
Nagoya, Japan, 14-18 October 2013.

2008 opinion summarization dataset[1]. We then compare these results to those found with the news articles of the Document Understanding Conference (DUC) 2007 Main task dataset[2]. The results show that in both types of texts, discourse relations seem to be as useful: *contingency*, *comparison*, and *illustration* relations provide a statistically significant improvement on the summary content; while the *attribution*, *topic-opinion*, and *attributive* relations do not provide a consistent and significant improvement.

2 Related Work on Discourse Relations for Summarization

The use of discourse relations for text summarization is not new. Most notably, (Marcu, 1997) used discourse relations for single document summarization and proposed a discourse relation identification parsing algorithm. In some work (e.g. (Bosma, 2004; Blair-Goldensohn and McKeown, 2006)), discourse relations have been exploited successfully for multi-document summarization. In particular, (Otterbacher et al., 2002) experimentally showed that discourse relations can improve the coherence of multi-document summaries. (Bosma, 2004) showed how discourse relations can be used effectively to incorporate additional contextual information for a given question in a query-based summarization. (Blair-Goldensohn and McKeown, 2006) used discourse relations for content selection and organization of automatic summaries and achieved an improvement in both cases. Discourse relations were also used successfully by (Zahri and Fukumoto, 2011) for news summarization.

However, the work described above have been developed for formal, well-written and factual documents. Most of these work show how discourse relations can be used in text summarization and show their overall usefulness. To the best of our knowledge, our work is the first to measure the effect of specific relations on the summarization of informal and opinionated text.

3 Tagging Discourse Relations

To evaluate the effect of discourse relations on a large scale, sentences need to be tagged automatically with discourse relations. For example, the sentence "*Yesterday, I stayed at home because it was raining*." needs to be tagged as containing a *cause* relation. One sentence can convey zero or several discourse relations. For example, the sentence "*Starbucks has contributed to the popularity of good tasting coffee*" does not contain any discourse relations of interest to us. On the other hand, the sentence "*While I like the Zillow interface and agree it's an easy way to find data, I'd prefer my readers used their own brain to perform a basic valuation of a property instead of relying on zestimates.*" contains 5 relations of interest: one *comparison*, three *illustrations*, and one *attribution*.

3.1 Most Frequent Discourse Relations

Since our work is performed within the framework of blog summarization; we have only considered the discourse relations that are most useful to this application. To find the set of the relations needed for this task, we have first manually analyzed 50 summaries randomly selected from participating systems at the TAC 2008 opinion summarization track and 50 randomly selected blogs from BLOG06 corpus[3]. In building our relation taxonomy, we considered all main discourse relations listed in the taxonomy of Mann and Thompson's Rhetorical Structure Theory (RST) (Mann and Thompson, 1988). These discourse relations are also considered in Grimes' (Grimes, 1975) and Williams' predicate lists. From our corpus analysis, we have identified the six most prevalent discourse relations in this blog dataset, namely *comparison*, *contingency*, *illustration*, *attribution*, *topic-opinion*, and *attributive*. The *comparison*, *contingency*, and *illustration* relations are also considered by most of the work in the field of discourse analysis such as the PDTB: Penn Discourse TreeBank research group (Prasad et al., 2008) and the RST Discourse Treebank research group (Carlson and Marcu, 2001). We considered three additional classes of relations: *attributive*, *attribution*, and *topic-opinion*. These discourse relations are summarized in Figure 1 while a description of these relations is given below.

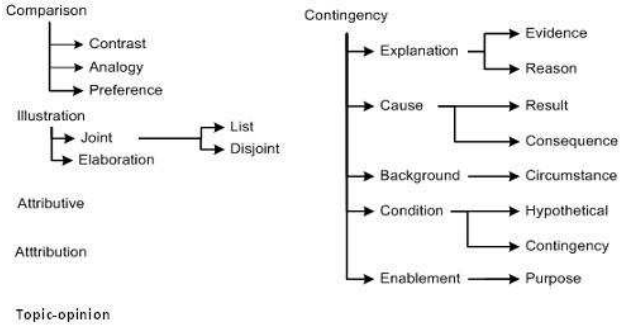
Illustration: Is used to provide additional information or detail about a situation. For example: *"Allied Capital is a closed-end management investment company that will operate as a business development concern."*

[1] http://www.nist.gov/tac/

[2] http://www-nlpir.nist.gov/projects/duc/guidelines/2007.html

[3] http://ir.dcs.gla.ac.uk/test_collections/blog06info.html

Figure 1: Most Frequent Discourse Relations in Blogs and their Sub-relations



As shown in Figure 1, *illustration* relations can be sub-divided into sub-categories: *joint*, *list*, *disjoint*, and *elaboration* relations according to the RST Discourse Treebank (Carlson and Marcu, 2001) and the Penn Discourse TreeBank (Prasad et al., 2008).

Contingency: Provides cause, condition, reason or evidence for a situation, result or claim. For example: *"The meat is good because they slice it right in front of you."*

As shown in Figure 1, the *contingency* relation subsumes several more specific relations: *explanation*, *evidence*, *reason*, *cause*, *result*, *consequence*, *background*, *condition*, *hypothetical*, *enablement*, and *purpose* relations according to the Penn Discourse TreeBank (Prasad et al., 2008).

Comparison: Gives a comparison and contrast among different situations. For example, *"Its fast-forward and rewind work much more smoothly and consistently than those of other models I've had."*

The *comparison* relation subsumes the *contrast* relation according to the Penn Discourse TreeBank (Prasad et al., 2008) and the *analogy* and *preference* relations according to the RST Discourse Treebank (Carlson and Marcu, 2001).

Attributive: Relation provides details about an entity or an event - e.g. *"Mary has a pink coat."*. It can be used to illustrate a particular feature about a concept or an entity - e.g. *"Picasa makes sure your pictures are always organized."*. The *attributive* relation, also included in Grimes' predicates (Grimes, 1975), is considered because it describes attributes or features of an object or event and is often used in query-based summarization and question answering.

Topic-opinion: We introduced topic-opinion relations to represent opinions which are not expressed by reported speech. This relation can be used to express an opinion: an internal feeling or belief towards an object or an event. For example: *"Cage is a wonderfully versatile actor."*

Attribution: These relations are instances of reported speech both direct and indirect which may express feelings, thoughts, or hopes. For example: *"The legendary GM chairman declared that his company would make "a car for every purse and purpose.""*

3.2 Automatic Discourse Tagging

Once the manual analysis identified the most prevalent set of relations, we tried to measure their frequency by tagging them automatically within a larger corpus. Only recently, the HILDA (Hernault et al., 2010) and (Feng and Hirst, 2012)'s discourse parser were made publicly available. Both of these parsers work at the text-level, as opposed to the sentence-level, and hence currently achieve the highest tagging performance when compared to the state of the art. (Feng and Hirst, 2012)'s work showed a significant improvement on the performance of HILDA by enhancing its original feature set. However, at the time this research was done, the only publicly available discourse parser was SPADE (Soricut and Marcu, 2003) which operates on individual sentences. To identify *illustration*, *contingency*, *comparison*, and *attribution* relations, we have used SPADE discourse parser. However, we have complemented this parser with three other approaches: (Jindal and Liu, 2006)'s approach is used to identify intra-sentence *comparison* relations; we have designed a tagger based on (Fei et al., 2008)'s approach to identify *topic-opinion* relations; and we have proposed a new approach to tag *attributive* relations (Mithun, 2012). A description and evaluation of these approaches can be found in (Mithun, 2012). By combining these approaches, a sentence is tagged with all possible discourse relations that it contains.

3.3 Distribution of Discourse Relations

To find the most prevalent discourse relations for opinion summarization, we have used the TAC 2008 opinion summarization track input document set (collection) which is a subset of BLOG06 and the answer nuggets provided by TAC 2008 as the reference summary (or model summaries), which

had been created to evaluate participants' summaries at the TAC 2008 opinion summarization track. The collection consists of 600 blogs on 28 different topics. The dataset of the model summaries consists of 693 sentences.

Using the discourse parsers presented in Section 3.2, we computed the distribution of discourse relations within the TAC 2008 opinion summarization collection and the model summaries. *Illustration*, *contingency*, *comparison*, *attributive*, *topic-opinion*, and *attribution* are the most frequently occuring relations in our data sets. The distribution is shown in Table 1 [4].

Table 1: Distribution of Discourse Relations in the TAC-2008 and DUC-2007 Datasets

Discourse Relation	TAC 2008		DUC 2007	
	Coll.	Model	Coll.	Model
Illustration	52%	46%	42%	38%
Contingency	31%	37%	34%	29%
Comparison	23%	18%	15%	12%
Attributive	12%	28%	3%	4%
Topic-opinion	14%	15%	4%	5%
Attribution	11%	9%	2%	3%
other	13%	9%	28%	31%
none	14%	10%	8%	7%

Table 1 shows that in the TAC 2008 input document set, the *illustration* relation occurs in 52% of the sentences; while *attribution* is the least frequently occurring relation. In this dataset, other relations, such as *antithesis* and *temporal* relations, occur in about 13% of the sentences and about 14% of the sentences did not receive any relation tag. As indicated in Table 1, the TAC model summaries have a similar distribution as the collection as a whole. The *attributive* relation seems, however, to be more frequent in the summaries (28%) than in the original texts (12%). We suspect that the reason for this is due to the question types of this track. To successfully generate query-relevant summaries that answer the questions of this track, candidate sentences need to contain *attributive* relations. For example, to answer the questions from this track *"Why do people like Picasa?"* or *"What features do people like about Windows Vista?"*, the summary needs to provide details about these entities or illustrate a particular feature about them. As a result, the summary will be composed of many *attributive* relations since *attributive* relations help to model the required information.

To compare the distribution of discourse relations within more formal types of texts such as news articles, we used the Document Understanding Conference (DUC) 2007 Main Task input document set (collection) and their associated model summaries. The DUC 2007 dataset is a news article based dataset from the AQUAINT corpus. The DUC 2007 input document set contains 1125 news articles on 45 different topics. The model summaries were used to evaluate the DUC 2007 participants' summaries. The dataset of the model summaries contains 180 summaries generated by the National Institute of Standards and Technology (NIST) assessors with a summary length of about 250 words. The distribution of relations in this dataset are shown in Table 1.

Table 1 shows that the most frequently occurring relation in the DUC 2007 document collection and in the model summaries is *illustration*; while the *attribution* relation is the least frequently occurring relation. Here again, it is interesting to note that the distribution of the discourse relations in the document collection and in the model summaries is generally comparable.

The distribution of the *illustration*, *contingency*, and *comparison* relations in the DUC 2007 dataset is comparable to those in the TAC 2008 opinion summarization dataset. Indeed, Table 1 shows that *illustration*, *contingency*, and *comparison* relations occur quite frequently irrespective of the textual genre. However, in contrast to the TAC dataset, *attributive*, *topic-opinion*, and *attribution* relations occur very rarely in DUC 2007. We suspect that this is mostly due to the opinionated nature of blogs. Another observation is that *temporal* relations (included in "other") occurred very frequently (30%) in the DUC 2007 dataset whereas this relation occurs rarely in the blog dataset. This is inline with our intuition that news articles present events that inherently contain temporal information.

4 Evaluation of Discourse Relations

To measure the usefulness of discourse relations for the summarization of informal texts, we have tested the effect of each relation with four different summarizers: BlogSum (Mithun, 2012), MEAD (Radev et al., 2004), the best scoring sys-

[4] In Table 1, the percentages do not add up to 100 because a sentence may contain more than one relation.

tem at TAC 2008[5] and the best scoring system at DUC 2007[6]. We have evaluated the effect of each discourse relation on the summaries generated and compared the results. Let us first describe the BlogSum summarizer.

4.1 BlogSum

BlogSum is a domain-independent query-based extractive summarization system that uses intra-sentential discourse relations within the framework based on text schemata. The heart of BlogSum is based on discourse relations and text schemata.

BlogSum works in the following way: First candidate sentences are extracted and ranked using the topic and question similarity to give priority to topic and question relevant sentences. Since BlogSum has been designed for blogs, which are opinionated in nature, to rank a sentence, the sentence polarity (e.g. positive, negative or neutral) is calculated and used for sentence ranking. To extract and rank sentences, BlogSum thus calculates a score for each sentence using the features shown below:

$$\textit{Sentence Score} = w_1 \times \textit{Question Similarity} + w_2 \times \textit{Topic Similarity} + w_3 \times \textit{Subjectivity Score}$$

where, question similarity and topic similarity are calculated using the cosine similarity based on words *tf.idf* and the subjectivity score is calculated using a dictionary-based approach based on the MPQA lexicon[7]. Once sentences are ranked, they are categorized based on the discourse relations that they convey. This step is critical because the automatic identification of discourse relations renders BlogSum independent of the domain. This step also plays a key role in content selection and summary coherence as schemata are designed using these relations.

In order not to answer all questions the same way, BlogSum uses different schemata to generate a summary that answers specific types of questions. Each schema is designed to give priority to its associated question type and subjective sentences as summaries for opinionated texts are generated. Each schema specifies the types of discourse relations and the order in which they should appear in the output summary for a particular question type. Figure 2 shows a sample schema that is used to answer *reason* questions (e.g. "*Why do people like Picasa?*"). According to this schema[8], one or more sentences containing a *topic-opinion* or *attribution* relation followed by zero or many sentences containing a *contingency* or *comparison* relation followed by zero or many sentences containing a *attributive* relation should be used.

Figure 2: A Sample Discourse Schema used in BlogSum

Relations & Constraints
Relation: $\{Topic\text{-}opinion/Attribution\}^{+}$
Constraint: Sentence Polarity.

Relation: $\{Contingency/Comparison\}^{*}$
Constraint: Compared Objects, Sentence Focus.

Relation: $Attributive^{*}$
Constraint: Sentence Focus.

Finally the most appropriate schema is selected based on a given question type; and candidate sentences fill particular slots in the selected schema based on which discourse relations they contain in order to create the final summary (details of BlogSum can be found in (Mithun, 2012)).

4.2 Evaluation of Discourse Relations on Blogs

To evaluate the effect of each discourse relation for blog summarization, we performed several experiments. We used as a baseline the original ranked list of candidate sentences produced by BlogSum before applying the discourse schemata, and compared this to the BlogSum-generated summaries with and without each discourse relation. We used the TAC 2008 opinion summarization dataset which consists of 50 questions on 28 topics; on each topic one or two questions were asked and 9 to 39 relevant documents were given. For each question, one summary was generated with no regards to discourse relations and two summaries were produced by BlogSum: one using the discourse tagger and the other without using the specific discourse tagger. The maximum summary length was restricted to 250 words.

To measure the effect of each relation, we have automatically evaluated how BlogSum performs using the standard ROUGE-2 and ROUGE-SU4

[5] http://www.nist.gov/tac/

[6] http://www-nlpir.nist.gov/projects/duc/guidelines/2007.html

[7] MPQA: http://www.cs.pitt.edu/mpqa

[8] The notation / indicates an alternative, { } indicates optionality, * indicates that the item may appear 0 to n times and + indicates that the item may appear 1 to n times

measures. For comparative purposes, Table 2 shows the official ROUGE-2 (R-2) and ROUGE-SU4 (R-SU4) for all 36 submissions of the TAC 2008 opinion summarization track. In the table, "TAC Average" refers to the mean performance of all participant systems and "TAC-Best" refers to the best-scoring system at TAC 2008.

Table 2: Results of the TAC 2008 Opinion Summarization Track

System Name	R-2	R-SU4
TAC Average	0.069	0.086
TAC-Best	0.130	0.139

Table 3: Effect of Discourse Relations on ROUGE-2 with the TAC 2008 Dataset

System Name	BlogSum R-2	MEAD R-2	TAC-Best R-2
Baseline	0.102⇓	0.041⇓	0.130
w/o Illustration	0.107⇓	0.022⇓	0.112⇓
w/o Contingency	0.093⇓	0.025⇓	0.102⇓
w/o Comparison	0.103⇓	0.033⇓	0.113⇓
w/o Attributive	0.113⇓	0.050	0.124
w/o Topic-opinion	0.112⇓	0.049	0.123
w/o Attribution	0.118⇓	0.051⇓	0.128
with all Relations	**0.125**	**.053**	**0.138**

Table 4: Effect of Discourse Relations on ROUGE-SU4 with the TAC 2008 Dataset

System Name	BlogSum R-SU4	MEAD R-SU4	TAC-Best R-SU4
Baseline	0.107⇓	0.064⇓	0.139
w/o Illustration	0.110⇓	0.041⇓	0.120⇓
w/o Contingency	0.102⇓	0.046⇓	0.110⇓
w/o Comparison	0.108⇓	0.052⇓	0.122⇓
w/o Attributive	0.115⇓	0.072	0.130
w/o Topic-opinion	0.117	0.072	0.129
w/o Attribution	0.127⇓	0.073⇓	0.132
with all Relations	**0.128**	**0.075**	**0.151**

The results of our evaluation are shown in Tables 3 (ROUGE-2) and 4 (ROUGE-SU4). As the tables show, BlogSum's baseline is situated below the best scoring system at TAC-2008, but much higher than the average system (see Table 2); hence, it represents a fair baseline. The tables further show that using both the ROUGE-2 (R-2) and ROUGE-SU4 (R-SU4) metrics, with the TAC 2008 dataset, BlogSum performs better when taking discourse relations into account. Indeed, when ignoring discourse relations, BlogSum has a R2=0.102 and R-SU4=0.107 and misses many question relevant sentences; whereas the inclusion of these relations helps to incorporate those relevant sentences into the final summary and brings the R-2 score to 0.125 and R-SU4 to 0.128. In order to verify if these improvements were statistically significant, we performed a 2-tailed t-test. The results of this test are indicated with the ⇓ symbol in Tables 3 and 4. For example, the baseline setup of BlogSum performed significantly lower for both R-2 and R-SU4 compared to BlogSum with all relations. This result indicates that the use of discourse relations as a whole helps to include more question relevant sentences and improve the summary content.

To ensure that the results were not specific to our summarizer, we performed the same experiments with two other systems: the MEAD summarizer (Radev et al., 2004), a publicly available and a widely used summarizer, and with the output of the TAC best-scoring system. For MEAD, we first generated candidate sentences using MEAD, then these candidate sentences were tagged using discourse relation taggers used under BlogSum. Then these tagged sentences were filtered using BlogSum so that no sentence with a specific relation is used in summary generation for a particular experiment. We have calculated ROUGE scores using the original candidate sentences generated by MEAD and also using the filtered candidate sentences. As a baseline, we used the original candidate sentences generated by MEAD. As a best case scenario, we have passed these candidate sentences through the discourse schemata used by BlogSum (see Section 4.1). In Tables 3 and 4, this is referred to as "MEAD with all relations". We have applied the same approach with the output of the TAC best-scoring system. In the tables, "TAC-Best Baseline" refers to the original summaries generated by the TAC-Best system and "TAC-Best with all relations" refers to the summaries generated by applying discourse schemata using the summary sentences generated by the TAC-Best system.

When looking at individual relations, Tables 3 and 4 show that considering *illustrations*, *contingencies* and *comparisons* make a statistically significant improvement in all scenarios, and with all summarisers. For example, if TAC-Best does not consider *illustration* relations, then the R-2 score decreases from 0.138 to 0.112, 0.102 and 0.113, respectively. On the other hand, the relations of *topic-opinion*, *attribution*, and *attributive* do not consistently lead to a statistically significant improvement on ROUGE scores.

It is interesting to note that although informal texts may not exhibit a clear discourse structure, the use of individual discourse relations such as *illustration*, *contingency* and *comparison* is nonetheless useful in the analysis of informal documents such as those found in the social media.

4.3 Effect of Discourse Relations on News

To compare the results found with blogs with more formal types of texts, we have performed the same experiments but, this time with the DUC 2007 Main Task dataset. In this task, given a topic (title) and a set of 25 relevant documents, participants had to create an automatic summary of length 250 words from the input documents. In the dataset, there were 45 topics and thirty teams participated to this shared task. Table 5 shows the official ROUGE-2 (R-2) and ROUGE-SU4 (R-SU4) scores of the DUC 2007 main task summarization track. In Table 5, "DUC Average" refers to the mean parformance of all participant systems and "DUC-Best" refers to the best scoring system at DUC 2007.

Table 5: DUC 2007 Main Task Summarization Results

System Name	R-2	R-SU4
DUC Average	0.095	0.157
DUC-Best	0.124	0.177

Table 6: Effect of Discourse Relations on ROUGE-2 with the DUC 2007 Dataset

System Name	BlogSum R-2	MEAD R-2	DUC-Best R-2
Baseline	0.089	0.099	0.124⇓
w/o Illustration	0.079⇓	0.061⇓	0.103⇓
w/o Contingency	0.074⇓	0.060⇓	0.097⇓
w/o Comparison	0.086⇓	0.078⇓	0.114⇓
w/o Attributive	0.092	0.099	0.119⇓
w/o Topic-opinion	0.092	0.099	0.115⇓
w/o Attribution	0.093	0.099	0.120⇓
with all Relations	**0.093**	**0.110**	**0.157**

Tables 6 and 7 show the results with this dataset with respect to ROUGE-2 and ROUGE-SU4, respectively. As the tables show, BlogSum's performance with all discourse relations (R2=0.093 and R-SU4=0.132) is similar to the DUC average performance shown in Table 5 (R2=00.095 and R-SU4=0.157) which is much lower than the DUC-Best performance (R2=0.124, R-SU4=0.177) shown in Table 5). However, these results show that even though BlogSum was designed for informal texts, it still performs relatively well with formal documents. Tables 6 and 7 further show that with the news dataset, the same relations have the most effect as with blogs. Indeed BlogSum generated summaries also benefit most from the *contingency*, *illustration*, and *comparison* relations; and all three relations bring a statistically significant contribution to the summary content.

Table 7: Effect of Discourse Relations on ROUGE SU-4 with the DUC 2007 Dataset

System Name	BlogSum R-SU4	MEAD R-SU4	DUC-Best R-SU4
Baseline	0.110⇓	0.142⇓	0.177⇓
w/o Illustration	0.117⇓	0.118⇓	0.138⇓
w/o Contingency	0.113⇓	0.118⇓	0.123⇓
w/o Comparison	0.122⇓	0.130⇓	0.144⇓
w/o Attributive	0.131	0.141⇓	0.159⇓
w/o Topic-opinion	0.130	0.141⇓	0.153⇓
w/o Attribution	0.131	0.142⇓	0.164⇓
with all Relations	**0.132**	**0.168**	**0.196**

Here again, as shown in Tables 6 and 7, we performed the same experiments with two other systems: the MEAD summarizer and the output of the DUC-Best system. Again, for the DUC 2007 dataset, each discourse relation has the same effect on summarization with all systems as with the blog dataset: *contingency*, *illustration*, and *comparison* provide a statistically significant improvement in content; while *attributive*, *topic-opinion* and *attribution* do not reduce the content, but do not see to bring a systematic and significant improvement.

5 Conclusion and Future Work

In this paper, we have evaluated the effect of discourse relations on summarization. We have considered the six most frequent relations in blogs - namely *comparison*, *contingency*, *illustration*, *attribution*, *topic-opinion*, and *attributive*. First, we have measured the distribution of discourse relations on blogs and on news articles and show that the prevalence of these six relations is not genre dependent. For example, the relations of *illustration*, *contingency*, and *comparison* occur frequently in both textual genres. We have then evaluated the effect of these six relations on summa-

rization with the TAC 2008 opinion summarization dataset and the DUC 2007 dataset. We have conducted these evaluations with our summarization system called BlogSum, the TAC best-scoring system, the DUC best-scoring system, and the MEAD summarizer. The results show that for both textual genres, some relations have more effect on summarization compared to others. In both types of texts, the *contingency*, *illustration*, and *comparison* relations provide a significant improvement on summary content; while the *attribution*, *topic-opinion*, and *attributive* relations do not provide a systematic and statistically significant improvement. These results seem to indicate that, at least for summarization, discourse relations are just as useful for informal and affective texts as for more traditional news articles. This is interesting, because although informal texts may not exhibit a clear discourse structure, the use of individual discourse relations is nonetheless useful in the analysis of informal documents.

In the future, it would be interesting to evaluate the effect of other relations such as the *temporal* relation. Indeed, *temporal* relations occur infrequently in blogs but are very frequent in news articles. Such an analysis would allow us to tailor the type of discourse relations to include in the final summary as a function of the textual genre being considered. In the future, it would also be interesting to use other types of texts such as reviews and evaluate the effect of discourse relations using other measures than ROUGE-2 and ROUGE-SU4. Finally, we would like to validate this work again with the newly available discourse parsers of (Hernault et al., 2010) and (Feng and Hirst, 2012).

Acknowledgement

The authors would like to thank the anonymous referees for their valuable comments on an earlier version of the paper. This work was financially supported by an NSERC grant.

References

Andreevskaia, A., Bergler, S., Urseanu, M.: All Blogs are Not Made Equal: Exploring Genre Differences in Sentiment Tagging of Blogs. *In Proceedings of the International Conference on Weblogs and Social Media (ICWSM-2007)*,(2007), Boulder, Colorado.

Blair-Goldensohn, S.J., McKeown, K.: Integrating Rhetorical-Semantic Relation Models for Query-Focused Summarization. *In Proceedings of the Document Understanding Conference (DUC) Workshop at NAACL-HLT 2006*, (2006), New York, USA.

Bosma, W.: Query-Based Summarization using Rhetorical Structure Theory. *In Proceedings of the 15th Meeting of Computational Linguistics in the Netherlands CLIN*, (2004), Leiden, Netherlands.

Carlson, L., Marcu, D.: Discourse Tagging Reference Manual. University of Southern California Information Sciences Institute, ISI-TR-545, 2001.

Fei, Z., Huang, X., Wu, L.: Mining the Relation between Sentiment Expression and Target Using Dependency of Words. *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation*, 257–264 (2008), Wuhan, China.

Feng, V. W., Hirst, G.: Text-level Discourse Parsing with Rich Linguistic Features. *In Proceedings of ACL-2012*, 60–68 (2012), Stroudsburg, USA.

Grimes, J. E.: The Thread of Discourse. Technical report No. NSF-TR-1, NSF-GS-3180. Cornell University, Ithaca, New York, 1975.

Hernault, H., Prendinger. H., duVerle, D. A., Ishizuka, M.: HILDA: A discourse parser using support vector machine classification. *J. Dialogue and Discourse*, 1(3):1–33, 2010.

Jindal, N., Liu, B.: Identifying Comparative Sentences in Text Documents. *In Proceedings of SIGIR-2006*, 244-251 (2006), Washington, USA.

Mann, W.C., Thompson, S. A.: Rhetorical Structure Theory: Toward a Functional Theory of Text Organisation. *J. Text*, 3(8):234–281, 1988.

Marcu, D.: From Discourse Structures to Text Summaries. *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization.* 1997, 82–88, Madrid, Spain.

McKeown, K.R.: Discourse Strategies for Generating Natural-Language Text. *J. Artificial Intelligence*, 27(1):1–41, 1985.

Mithun, S.: Exploiting Rhetorical Relations in Blog Summarization. *PhD Thesis*, Department of Computer Science and Software Engineering, Concordia University, Montreal, Canada, 2012.

Otterbacher, J. C., Radev, D. R., Luo, A.: Revisions that Improve Cohesion in Multi-document Summaries: A Preliminary Study. *In Proceedings of the ACL-2002 Workshop on Automatic Summarization*, 27–36 (2002), Philadelphia, USA.

Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A., Robaldo, L., Webber, B.: The Penn Discourse Treebank 2.0. Annotation Manual. University of Pennsylvania, IRCS-08-01, 2008.

Radev, D. et al.: MEAD -A Platform for Multidocument Multilingual Text Summarization. *In Proceedings of LREC-2004*, 1–4 (2004), Lisbon, Portugal.

Soricut, R., Marcu, D.: Sentence Level Discourse Parsing using Syntactic and Lexical Information. *In Proceedings of NAACL/HLT 2003*, 149–156 (2003), Edmonton, Canada.

Taboada, M.: Discourse Markers as Signals (or not) of Rhetorical Relations. *J. Pragmatics*, 38(4):567–592, 2006.

Zahri, N. A. H. B., Fukumoto, F.: Multi-document Summarization Using Link Analysis Based on Rhetorical Relations between Sentences. *In Proceedings of CICLing*, 328–338 (2011), Tokyo, Japan.