On the use of a schema-based framework to improve relevance and discourse coherence in blog summarization

Shamima Mithun [1], **Leila Kosseim** [2]

1. *Concordia University*
   Department of Computer Science and Software Engineering
   Montreal, Quebec, Canada
   s_mithun@encs.concordia.ca

2. *Concordia University*
   Department of Computer Science and Software Engineering
   Montreal, Quebec, Canada
   kosseim@encs.concordia.ca

ABSTRACT. *Question irrelevance and discourse incoherence are important and typical problems in multi-document summarization especially when dealing with informal and opinionated texts. To address these two issues, we propose a domain-independent query-based summarization approach for opinionated documents that uses intra-sentential discourse structures in the framework of schemata. We have developed a generic domain-independent schema-based approach that selects the most appropriate text schema to answer specific types of questions. The schemata define the content and the organization of summaries based on the discourse relations present in candidate sentences. To decide which candidate sentences should be included in the final summary and where, each sentence is automatically tagged with the rhetorical predicates it conveys and allowed to fill a slot of the schema. Finally post-schema heuristics that work at the inter-sentence level are used to improve coherence further.*

To validate our approach, we have built a system named BlogSum and have evaluated its performance for question relevance and coherence using two datasets: blogs and reviews. ROUGE scores show that our approach is effective at reducing question irrelevant sentences and a manual evaluation shows a significant improvement in question relevance and coherence compared to the original candidate list. These results indicate that the use of discourse relations combined with text schemas can effectively reduce question irrelevance and discourse incoherence even with informal and opinionated documents.

RÉSUMÉ. *Les problèmes de pertinence et de cohérence discursive sont des enjeux importants et usuels dans le cadre du résumé automatique à partir de documents multiples, en particulier, lorsque les documents sources sont informels et sont fondés sur des opinions plutôt que des faits. Pour faire face à ces problèmes, nous proposons une approche pour le résumé automatique à partir de requêtes (query-based summarisation) qui utilise les relations de discours intra-phrastiques jumelées à des schémas textuels. Nous avons développé une approche*

Document numérique – n° 2/2012, 91-120

générique et indépendante du domaine qui sélectionne le schéma textuel le plus approprié pour répondre à certains types de questions. Les schémas définissent le contenu et l'organisation des résumés en se basant sur les relations de discours présentes dans les phrases candidates. Pour sélectionner quelles phrases devraient être incluses dans le résumé et où, chaque phrase est automatiquement étiquetée par les relations rhétoriques qu'elle contient permettant ainsi de remplir des positions spécifiques dans le schéma textuel. Finalement, des heuristiques post-schéma œuvrant au niveau inter-phrastique sont utilisées pour améliorer la cohérence.

Pour valider notre approche, nous avons développé un système nommé BlogSum et avons évalué ses performances vis-à-vis de la pertinence et de la cohérence textuelle en utilisant deux types de documents : des blogs et des critiques. Les scores ROUGE démontrent que notre approche est efficace pour réduire les phrases non pertinentes et une évaluation manuelle démontre une nette amélioration de la pertinence et de la cohérence textuelle comparé à la liste des phrases candidates originale. Ces résultats indiquent que l'utilisation de relations discursives combinées à des schémas textuels peut améliorer la pertinence et la cohérence des résumés même dans le cas de documents informels et critiques.

KEYWORDS: *automatic summarization, discourse relations, schemata, question relevance, discourse coherence.*

MOTS-CLÉS : *résumé automatique, relations de discours, schémas textuels, pertinence, cohérence discursive.*

DOI:10.3166/DN.15.2.91-120

1. Introduction

With the rapid growth of the Social Web, a large amount of informal opinionated texts are available on numerous topics. Natural language tools for automatically analyzing these opinions become necessary to help individuals, organizations, and governments in making timely decisions. For example, businesses and organizations are interested to know consumers' opinions and sentiments as part of their product and service evaluations; individuals are interested to know others' opinions when they intend to purchase a product or service... Various natural language tools to process and utilize information from texts have already been developed. Question answering systems (e.g. (Yang *et al.*, 2003)) and summarization systems (e.g. (Liu *et al.*, 2007)) are only a few examples. However, most of these systems have been developed to process factual information, for example news articles or scientific papers. As more and more people use the Web to express their opinions, natural language tools to automatically analyze opinionated information has quickly become a necessity. A query-based opinion summarizer from opinionated documents, as introduced in 2008 at the Text Analysis Conference (TAC)[1], can address this need. Query-based opinion summarizers present what people think or feel on a given topic in a condensed

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

manner to analyze others' opinions regarding a specific question (e.g. *Why do people like Starbucks better than Dunkin Donuts?*[2]). This kind of topic-oriented, informative multi-document summarization is similar to complex questions answering (Chali *et al.*, 2009; Harabagiu *et al.*, 2006) that requires inferencing and synthesizing information from multiple documents. This research interest motivated us to develop an effective query-based extractive multi-document opinion summarization approach for blogs.

Over the years, results of the Document Understanding Conference (DUC) and Text Analysis Conference (TAC) have shown that system-generated summaries are by no means comparable to human-generated summaries (Conroy, Dang, 2008; Dang, Owczarzak, 2008). Recently, (Genest *et al.*, 2009) empirically demonstrated that there is still much space to improve coherence of summaries even for pure extractive summaries. All these results indicate that extractive summaries can be improved for both content and coherence.

Literature (e.g. (Otterbacher *et al.*, 2002; Ku *et al.*, 2006)) as well as our own study (Mithun, Kosseim, 2009) show that *Question Irrelevance*, *Topic Irrelevance*, and *Discourse Incoherence* are the most frequently occurring errors in blog summaries and that these errors occur more frequently in blog summaries compared to news summaries. Indeed, sentences in blogs do not have a predictable discourse structure (e.g. in formal writing, the first and the last sentences of a paragraph usually contain important information) which can be used to rank sentence during summarization. As a result, it is much more difficult to rank blog sentences compared to news article sentences. Opinion (sentiment) information is typically used to rank blog sentences for summarization, but this task can possibly add more noise to the blog sentence ranking process if not done properly. Moreover, unlike focused news articles, blogs are quite unfocused. The above mentioned errors decrease the overall quality of a summary; *Question Irrelevance* and *Topic Irrelevance* errors weaken the summary content and *Discourse Incoherence* reduces the summary coherence.

In our research, we targeted to reduce *Question Irrelevance* and *Discourse Incoherence* with the goal of improving the summary content and its coherence. In addition, we believe that our sentence selection method also reduces *Topic Irrelevance*. The heart of our approach is based on discourse relations and text schemata. Text schemata were first introduced by (McKeown, 1985) and used by other researchers (e.g. (Cline, Nutter, 1994)) in the context of question answering and text generation. However, schema-based approaches are usually domain-dependent. Discourse relations have been found useful in natural language generation (McKeown, 1985) and in news summarization (Blair-Goldensohn, McKeown, 2006; Bosma, 2004). However, to the best of our knowledge, discourse relations have never been used for blog summarization. We propose a domain-independent query-based blog summarization approach using intra-sentential discourse relations as opposed to inter-sentential discourse relations in the framework of schemata. To verify our approach, we have devel-

2. This question originates from the TAC 2008 dataset: http://www.nist.gov/tac/

oped a system called BlogSum and evaluated its performance using the Text Analysis Conference (TAC 2008) opinion summarization track data. The evaluation results show the effectiveness of our approach in reducing question irrelevant sentences by about 18% using the ROUGE scores and in significantly improving question relevance and coherence with a p-value of 0.0028 in a t-test and p-value of 0.0223 in a t-test using a manual likert scale of 1 to 5 compared to the original candidate list. We have also evaluated BlogSum-generated summaries using the OpinRank dataset and (Jindal, Liu, 2006)'s dataset of reviews for question relevance and coherence. The t-test results of this experiment show that in a two-tailed test, BlogSum also performs significantly better than the original candidate list with a p-value of 0.0023 and a p-value of 0.0371 for question relevance and coherence, respectively.

This paper is organized as follows: Section 2 defines and describes the problems of *Question Irrelevance* and *Discourse Incoherence*; Section 3 provides related work; Section 4 describes the heart of our approach; and finally Sections 5 and 6 contain the evaluation results, and conclusion and future work.

2. Question Irrelevance and Discourse Incoherence

Summaries generated automatically can contain many types of problems that must be addressed in order to make them more useful and natural. A query-focused summary is produced from a document or a set of documents to satisfy a request for information expressed by a question. In a query-focused summary, if the sentences are not relevant to the question, then the summary exhibits a *Question Irrelevance* error. A question irrelevant summary does not fulfil the user's information need as it does not relate to the original question. On the other hand, a summary will exhibit a *Discourse Incoherence* error if the reader cannot identify the communicative intentions of the writer from the sentences or if the sentences do not seem to be interrelated. A summary with poor coherence confuses the readers and degrades the quality and readability of the summary.

Figure 1 shows two sample summaries taken from the TAC 2008 opinion summarization track. Summary 1 contains a *Question Irrelevance* error because the second sentence is not relevant to the question. Currently, most of the automatic query-based summarization systems use extractive approaches. In general, these approaches work in two steps: first the most salient sentences are extracted from the source documents and then these sentences are ordered to create a summary. An inadequate content selection can result in *Question Irrelevance*.

To select sentences, current query-based summarization approaches typically compute the similarity between the question and candidate sentences (Murray *et al.*, 2008) using linguistic or statistical methods. Most of these approaches (e.g. (Radev, D. and Allison, T. and Blair-Goldensohn, S. *et al.*, 2004; Murray *et al.*, 2008)) utilize predefined features such as the sentence position in the document. However, many of these features might not be useful for unstructured domain like blogs because these do not have predictable discourse structures. Moreover, the semantic category of the ques-

tion, which is implicitly used in human writing to answer a specific type of question, is ignored in current approaches. As a result, current approaches often suffer from *Question Irrelevance* problems.

Topic: *Carmax*
Question: *What motivated positive opinions of Carmax from car buyers?*

Summary1:
(1) At Carmax, the price is the price and when you want a car you go get one. (2) Arthur Smith, 36, has been living in a van outside the Carmax lot, 24 hours a day, for more than a month. (3) Sometimes I wonder why all businesses can't be like Carmax. [...]

Summary2:
(1) It's like going to disney world for car buyers. (2) have to say that Carmax rocks. (3) We bought it at Carmax, and I continue to have nothing bad to say about that company. (4) After our last big car milestone, we've had an odyssey with cars. [...]

Figure 1. Sample Summaries

On the other hand, Summary 2 in Figure 1 shows a sample summary that contains a *Discourse Incoherence* error. Even though all the sentences are relevant to the question, improper sentence ordering degrades the coherence of this summary as the reader cannot deduce the discourse relations between sentences. For this summary a more coherent sentence order would be 4-3-1-2 or 4-3-2-1.

In extractive summarization, sentences may be selected from multiple documents or without consideration to their interdependency with other sentences. Moreover, in multi-document summarization, documents may be written by different writers who have different perspectives and writing styles; a strategy that deals with sentences on an individual basis can very well create *Discourse Incoherence* problems.

Two major types of sentence reordering approaches are used to address coherence: making use of chronological information (e.g. (Barzilay *et al.*, 2002; McKeown *et al.*, 2002)) and learning the natural order of sentences from large corpora (e.g. (Barzilay, Lee, 2004; Lapata, 2003)). However, in the first case, if the source documents are not event-based, the quality of the summaries will be degraded because temporal cues are missing. In the later case, probabilistic models of text structures are trained on a large corpus. If the genre of the training corpus and the source documents mismatch then they will perform poorly. In other work, (e.g. (Bosma, 2004; Blair-Goldensohn, McKeown, 2006; Marcu, 1997; Zahri, Fukumoto, 2011)) discourse relations are used to improve coherence in order to better simulate human writing where textual contents are typically connected to each other using various discourse relations. However, most of these work are developed for a particular domain or genre (e.g. news articles). Some schema and template based approaches have been used successfully in achieving

coherence (e.g. (Sauper, Barzilay, 2009; Jaidka *et al.*, 2010)); however, they are either domain dependent or applied to a very structured domain (e.g. Wikipedia pages).

The problem of question irrelevance and incoherence is not limited to text summarization but is also a concern in other applications such as natural language generation and question answering. Since question irrelevance and discourse incoherence errors are the outcomes of an inadequate content selection and content organization of the extractive summarization approach, if the summary contents are selected properly and the selected contents are organized properly then question irrelevance and discourse incoherence problems could be significantly reduced.

3. Related Work

Summarization for opinionated text is a recent endeavor. Such summarization work have been applied to diverse domains such as customer reviews (Hu, Liu, 2004) and conversations (Wang, Liu, 2011). Query-based blog summarization approaches have been first developed within the context of the TAC 2008 summarization track. Most of these summarization approaches designed at TAC 2008 (e.g. (Kim *et al.*, 2008; Murray *et al.*, 2008)) use feature-based sentence ranking for content selection where sentences with the highest scores are kept to produce the summary. These approaches mostly use question similarity, sentence position, polarity scores, and centroid as features. In our summarization approach, to extract initial candidate sentences, question similarity and polarity values are used along with subjectivity scores and topic similarity (described in Section 4). However, in our approach, the final summary content is selected using the schema. In TAC 2008, some systems (e.g. (Hendrickx, Bosma, 2008)) also use graph-based approaches, which is commonly used in news summarization, for sentence ranking. Most of the high performing systems for summary content at TAC 2008 used answer snippets[3] which were provided with the TAC 2008 dataset. In our summarization approach, we did not use answer snippets.

Most of the approaches at TAC 2008 used sentence scores to order final summaries. The highest ranked system for summary coherence at TAC 2008 (Conroy, Schlesinger, 2008) modeled the sentence ordering for outputs as a Traveling Salesman Problem, finding the shortest path among the sentences where term overlap was used to calculate sentence similarity (Conroy, Schlesinger, 2008). The second best ranked system at TAC 2008 for summary coherence (Bossard, Genereux, 2008) grouped sentences into three different categories positive, negative, and neutral for sentence ordering. In their approach, groups of sentences appeared in the same order as the question. In other words, if the first question was tagged as positive, the first sentences appearing in the summary were positive sentences. However, none of the top ranking systems at TAC 2008 used discourse relations to address summary coherence as we did.

3. Answer snippets were provided as optional, additional input in the form of answer-containing text snippets found by question answering systems that participated in the TAC 2008 question answering track and/or human assessors, along with a supporting document ID for each snippet. The answer-snippet need not appear literally in its associated document, but may be derived from information in the document.

Recently, (Paul *et al.*, 2010) developed a blog summarization approach to highlight contrast between multiple viewpoints expressed towards a topic by developing a model to jointly represent topic and viewpoints in the text. Some other researchers also attempted to add new dimensions to blog summarization such as usage of comment of blog posts (Potthast, Becker, 2010). However, to the best of our knowledge, text schemata and discourse relations have never been utilized in blog summarization.

In (McKeown, 1985), McKeown introduced a schema-based approach for text planning based on the observation that certain standard patterns of discourse organization (that she called schema) are more effective to achieve a particular discourse goal. She then demonstrated the usability of this schema-based approach for a domain-dependent question answering application. In this application, McKeown designed various schemata that incorporate discourse structures which are typically used in human writing to answer a specific question type (e.g. *identification*). Text schemata were later used by other researchers (e.g. (Cline, Nutter, 1994)) where specific schemata were designed according to the specific applications. In more recent work, (Jaidka *et al.*, 2010; Sauper, Barzilay, 2009) also tried to utilize discourse structures learned from domain relevant articles (e.g. scientific research paper) to design schemata (or templates) for summarization.

We also believe that regardless of the domain, for a particular type of question, certain types of sentences, if organized in a certain order, can meet the communicative goal more effectively to create a question-relevant and coherent text. For example, to take (McKeown, 1985)'s example, to define an entity or an event (e.g. "*what is a ship?*") it is natural to first include the identification of the item as a member of a generic class, then to describe the object's constituency or attributes followed by a specific example and so on. On the other hand, a comparison of two objects should use a different combination of sentences to be question-relevant and coherent.

Available schema-based approaches are typically domain-dependent and the domain knowledge is explicitly represented in knowledge bases which is used to identify discourse structures. As opposed to using this approach and target only a specific domain by tagging discourse relations in advance in a knowledge base, we have tried to develop a generic text schema-based approach applicable to any domain by identifying discourse relations automatically.

To describe discourse relations, different theories have been developed such as Rhetoric (Aristotle, 1954), Rhetorical Predicates (Grimes, 1975; Hobbs, 1985), Rhetorical Structure Theory (Mann, Thompson, 1988) and other theories by (Grosz, 1985; Hovy, 1993). Some theories are inclusive compared to others with respect to discourse structure definition and applicability. For example, Rhetorical Structure Theory (RST) (Mann, Thompson, 1988) is comprehensive compared to its predecessors because it provides extensive definitions of various discourse relations and showed that plan based approach can be used to apply these relations computationally (Mitkov, 1993). However, the set of discourse relations proposed by these theories are often comparable.

Discourse relations have been used in the past for text summarization. Most notably (Marcu, 1997) used discourse relations for single document summarization and proposed a discourse relation identification parsing algorithm. (Mani *et al.*, 1998; Otterbacher *et al.*, 2002) experimentally showed that discourse relations can improve the coherence of multi-document summaries. In some work (e.g. (Blair-Goldensohn, McKeown, 2006; Bosma, 2004)), discourse relations have been exploited successfully for multi-document summarization of news articles. In these work, discourse relations across sentences are utilized. (Bosma, 2004) shows the effectiveness of discourse relations to incorporate additional contextual information for the question. The evaluation was done on selected domains for which annotated discourse relations were available. (Blair-Goldensohn, McKeown, 2006) used discourse relations for content selection and organization and achieved improvement in both cases. However, due to the lack of availability of automatic approaches to identify discourse relations across sentences, they only covered two discourse relations: *cause* and *contrast*. Since this approach was developed for news articles, we could not compare our approach with theirs. Discourse relations were also used successfully by (Zahri, Fukumoto, 2011) for news summarization. In their work, they utilized 5 majors types of relations based on the cross-document relationship type between sentences proposed in Cross-document Structure Theory (CST). As opposed to using inter-sentential relations, we used intra-sentential relations and we utilized a large number (28) of discourse relations.

In news summarization, discourse relations across sentences were found useful. However, available approaches are domain-dependent or use only few discourse relations because of the unavailability of reliable automatic identification of inter-sentence relations. However, in extractive summarization, inter-sentential discourse relations are less likely be found. For example, in extractive summarization, question answering, information retrieval and in many other applications, individual sentences are extracted from different documents or from different positions of a document to build a candidate sentence list. As a result, it will be unlikely that inter-sentential relations will be present among candidate sentences. Instead, in these applications, it will be more advantageous to utilize intra-sentential relations. Intra-sentential relations have already been found useful to organize texts and select content by utilizing schema in question answering (McKeown, 1985; Blair-Goldensohn, 2007). Intra-sentential relations may enable to answer non-factoid questions such as "*Why do people like Picasa?*" by selecting text spans related through a causality. This was demonstrated by (Blair-Goldensohn, 2007) who showed that 95% of the time, causality occurred within sentences in the T corpus[4]. In addition, (Soricut, Marcu, 2003) notes that 95% of the sentences in the RST Discourse Treebank corpus contain intra-sentential relations. This is why in our research, we focused on intra-sentential discourse relations only to categorize sentences (see Section 4.2).

4. A gigaword newswire corpus of 4.7 million newswire documents.

4. Our Content Filtering and Organization Approach

Our approach to content filtering and organization assumes that we are given: 1) a topic, 2) an initial question on the topic, 3) a set of related input blogs on the topic, and 4) a ranked list of sentences extracted from the document set based on their extraction scores. To extract and rank sentences, our approach calculates a score for each sentence using the features shown below:

Sentence Score = Question Similarity + Topic Similarity + $|SubjectivityScore|$

where, question similarity and topic similarity are calculated using cosine similarity based on words *tf.idf* and subjectivity score is calculated using a dictionary-based approach using the MPQA lexicon[5]. However, any other sentence ranker could have been used.[6]

With the given inputs mentioned above, our approach tries to select a few most relevant sentences from the candidate sentence list and order them so as to produce a query relevant coherent summary. This is done by performing the following tasks (A.) Question Categorization (B.) Predicate Identification (C.) Schema Selection (D.) Sentence Selection and Ordering. In these tasks, questions need to be categorized based on their communicative goals (A). To include candidate sentences in the final summary, sentences need to be classified into predefined rhetorical predicates (B) to fill a slot of the matched schema. The most appropriate schema needs to be selected for the question categories (C) to incorporate the most relevant sentences into the summary from the candidate list. At the end of this process, sentences are reordered to improve coherence (D). Let us describe these steps in detail.

4.1. Question Categorization

Our content organization approach first categorizes questions to determine which schema will better convey the expected communicative goal of the answer for a particular question type. In our work, we have considered three categories of questions based on their communicative goals: *comparison*, *reason*, and *suggestion*. These question categories were determined by analyzing the TAC 2008 opinion summarization track questions.

1. *Comparison* questions ask about the differences between objects - e.g.
i) *What is the difference between iPod Touch and Zune HD?*
ii) *Why do people like Starbucks better than Dunkin Donuts?*

5. MPQA: http://www.cs.pitt.edu/mpqa

6. Our approach also removes redundant sentences from the candidate list using the cosine similarity. However, the novelty of our approach resides in the use of schemata and discourse relations to improve query relevance and coherence. Since our approach is extractive, currently it does not perform sentence compression.

2. *Reason* questions ask for reasons for some claim - e.g.
i) *Why do people like Mythbusters?*
ii) *What reasons are given for liking the services provided by Jiffy Lube?*

3. *Suggestion* questions ask for ideas to solve some problems - e.g.
i) *What do Canadian political parties want to happen regarding NAFTA?*
ii) *What steps are being suggested to correct this problem?*

Automatically classifying a new question into one of these 3 categories is a typical text classification task. Hence several approaches were available, notably based on machine learning approaches (e.g. (Jindal, Liu, 2006)). However, we have found that the use of simple lexico-syntactic patterns was sufficient, as the syntax and styles of the questions were rather standard and the number of classes was low. We have therefore designed lexico-syntactic patterns for each question type based on part of speech tags. For the *reason* questions, we have analyzed sample questions distributed for system development by the TAC 2008 opinion summarization track organizers. This sample set contains 16 questions and none of these questions appeared in the TAC 2008 opinion summarization track dataset. Since this sample set was only consisted of *reason* questions, to design the lexico-syntactic patterns for the *comparison* question, we used part of the dataset (50 randomly selected comparison questions) by (Jindal, Liu, 2006). For the *suggestion* question type, we have analyzed the same set of 3 questions (the TAC 2008 opinion summarization track questions) which we used to identify the question categories. By analyzing our development question set, we have designed 4 patterns for comparison questions, 4 patterns for suggestion questions, and 6 patterns for reason questions. Some of the lexico-syntactic patterns are shown in Figure 2.

Patterns
Comparison:
Pattern: [...]NP VB(opinionated terms) NP RBR(comparison terms) PP(containing topics)
Example: Why do people like Starbucks better than Dunkin Donuts?
Suggestion:
Pattern: What NNS VBP suggested/advised [...]
Example: What steps are being suggested to correct this problem?
Reason:
Pattern: Why do/don't NNS VB(opinionated terms) NP(containing topic)
Example: Why do people like Picasa?

Figure 2. Sample Lexical Patterns for Question Categorization

where, [...] refers to any lexical pattern; *NP*, *RBR*, *NNS*, ... refer to parts of speech categories. The topic is the target which is annotated in the corpus.[7]

7. In the question categorization task, we also need to know the word polarity (opinionated or not) and topic term information. The MPQA lexicon was used to know the word prior polarity and topic term information was extracted from the annotated dataset.

We have calculated the accuracy of the patterns for the *reason* and *comparison* question types. For the reason type question, we used the TAC 2008 opinion summarization track questions and we achieved an accuracy of 96%. Since the TAC 2008 opinion summarization track contains only 4% of *comparison* questions and 6% of *suggestion* questions, we could not use this dataset to evaluate the accuracy of patterns for *comparison* and *suggestion* questions. Therefore, we calculated the accuracy of the patterns for *comparison* questions using 100 comparison questions (different from the development set) from the (Jindal, Liu, 2006)'s dataset and achieved an accuracy of 97%. Due to the lack of data, we could not evaluate patterns of the *suggestion* question type. However, the evaluation results with the review dataset (Section 5.2.3) show that all 3 question types perform well.

4.2. Predicate Identification

In our schema-based approach, the basic units of a schema are *rhetorical predicates* which characterize the structural relations between propositions in a text where propositions can be clauses or sentences and describe different predicating acts a writer can use. Rhetorical predicates can also model discourse structures which are used to provide a definition or attributes of an object or a concept. For example, the sentence "*Mary has a pink coat*." provides details about an object.

In our work, we first defined the set of rhetorical predicates that are more useful for our application then we developed an automatic approach to identify these predicates. In our work, we used intra-sentential rhetorical predicates which occur between two clauses or within a clause. For example, the sentence "*[Although Mr. Freeman is retiring,] [he will continue to work as a consultant for American Express on a project basis.]*" shows a discourse structure where two clauses are held together with a relation called *contrast*. On the other hand, the sentence "*Its fast-forward and rewind work much more smoothly and consistently than those of other models I've had.*" shows a discourse structure where a *comparison* relation occurs within a clause.

4.2.1. Our Set of Rhetorical Predicates

Twenty-eight rhetorical predicates, which have been found most useful for our blog summarization application, were considered. These predicates are shown in Figure 3. Six of them are top level predicates:

1. Attributive: Provides details about an entity or event or can be used to illustrate a particular feature about a concept - e.g. *Picasa makes sure your pictures are always organized.*

2. Comparison: Gives a comparison and contrast among different situations - e.g. *Perhaps that's why for my European taste Starbucks makes great espresso while Dunkin's stinks.*

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

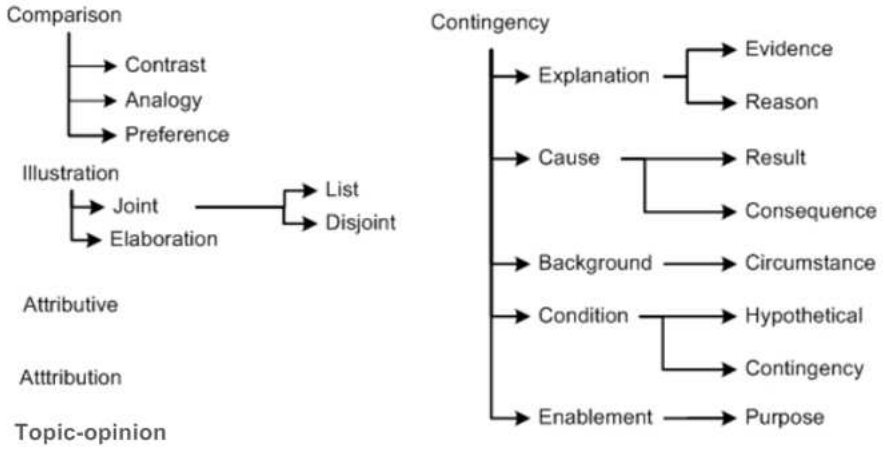

Figure 3. Rhetorical Predicates that we Considered

4. Illustration: Is used to provide additional information or detail about a situation - e.g. *The Xbox 360 and Vista both will use a new technology that makes games run at the fastest speed possible.*

5. Attribution: Provides instances of reported speech both direct and indirect which may express feelings, thoughts, or hopes - e.g. *The legendary GM chairman declared that his company would make "a car for every purse and purpose."*

6. Topic-opinion: We introduced topic-opinion predicates to represent opinions which are not expressed by reported speech. It can be used to express an opinion; an agent can express internal feeling or belief towards an object or an event - e.g. *The thing that I love about their sandwiches is the bread.*

Comparison, contingency, and illustration predicates can be sub-divided into sub-categories as shown in Figure 3. All these 28 predicates are considered to tag a sentence (see Section 4.3).

4.2.2. Predicate Identification Approaches

In our approach, candidate sentences need to be tagged with these rhetorical predicates based on what discourse relations they contain. Candidate sentences are therefore classified as containing zero or many rhetorical predicates to fill the various slots of the matched schema. For example, the sentence "*Yesterday, I stayed at home because it was raining*." will be tagged as a cause predicate as it contains the discourse relation cause. One sentence can convey zero or more rhetorical predicates. For example, the sentence "*Starbucks has contributed to the popularity of good tasting coffee*" does not contain any rhetorical predicate 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 pre-*

fer my readers used their own brain to perform a basic valuation of a property instead of relying on zestimates." contains 4 predicates of interest: contrast (sub-category of comparison), joint (sub-category of illustration), attribution, and elaboration (sub-category of illustration).

In (Mithun, Kosseim, 2011), we have presented four domain-independent predicate identification approaches. As specified earlier, predicates can describe a single clause or a relation between clauses. To identify the predicates that exist between clauses - e.g. *evidence*, we have used the SPADE discourse parser (Soricut, Marcu, 2003). On the other hand, in order to identify predicates within a single clause (not covered by SPADE) - e.g. *attributive*, we have used three other approaches. (Jindal, Liu, 2006)'s approach is used to identify intra-clause comparison predicates; we have designed a tagger based on (Fei *et al.*, 2006)'s approach to identify topic-opinion predicates; and we have proposed an approach to tag attributive predicates (Mithun, Kosseim, 2011). A description of these approaches can be found in (Mithun, Kosseim, 2011). By combining these approaches, a sentence is tagged with all possible predicates that it may contain and is ready to be used in a schema. If a sentence does not contain any rhetorical predicate of our interest, that sentence will not be part of the final summary.

4.3. Schema Selection

In human writing, writers often convey different content and organize it differently in order to answer specific types of question and make the answer more relevant to the question and its communicative goal. Based on this observation, we have designed three text schemata that specify which set of rhetorical predicates are more effective to answer each question type we have considered, 1) *comparison*, 2) *suggestion*, and 3) *reason*. To following (McKeown, 1985)'s work, we have studied 15 articles of each type written by different authors. To design these schemata, we have studied compare/contrast essays and comparison review articles found on the web to design the comparison schema; and argumentative essays and problem-solution essays to design the reason and the suggestion schemata, respectively. In this analysis, we have annotated sentences with predicates. From our analysis, we have derived which question types should contain which type of predicates and where. Each schema is designed based on giving priority to its associated question type and subjective sentences as we are generating summaries for opinionated texts. For each type of schema, we have also defined constraints on the predicates based on their semantic content in order to improve question relevance. As part of the schema selection, our approach selects the associated schema for a specific question category to select and order sentences for the final summary.

Figure 4 shows the comparison schema used to answer a comparison question. According to this schema, a sentence to be included at the beginning of the summary needs to contain either a comparison predicate or a contingency predicate followed by sentences containing a topic-opinion or attribution predicate then by illustration

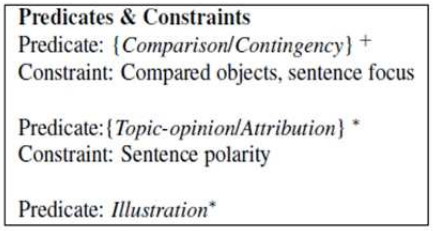

Figure 4. The Comparison Schema

predicates[8]. From Figure 4, we can see that constraints are also defined on predicates based on their semantic content to make summaries question relevant. In the example, the comparison and contingency predicates must contain all objects or events which are being compared and the topic of the sentence needs to be the focus of the sentence (meaning the topic needs to be the subject or object of the sentence) and topic-opinion and attribution predicates must have the same polarity as the question.

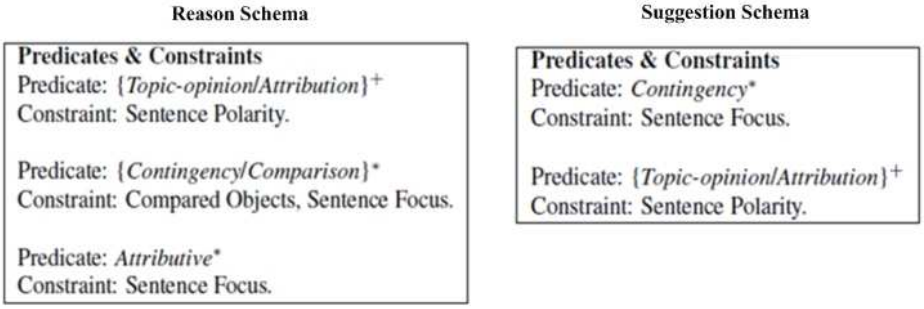

Figure 5. The Reason and the Suggestion Schemas

Figure 5 shows the reason and the suggestion schemas. In Figure 5, the constraints "sentence polarity", "compared objects", and "sentence focus" indicate that the sentence needs to have the same polarity as the question, the sentence needs to contain all objects which are being compared, and the topic of the sentence needs to be the focus of the sentence, respectively.

It must be noted that schemata can be designed in different ways. However, our current content organization approach allows the generation of different summaries for particular question types by providing flexible sentence selection and reordering strategies.

8. Following (McKeown, 1985)'s notations, the symbol / indicates an alternative, * indicates that the item may appear 0 to n times, + indicates that the item may appear 1 to n times.

4.4. Sentence Selection and Ordering

In our approach, sentence selection and ordering is accomplished by selecting candidate sentences to fill particular slots in the selected schema based on which rhetorical predicate they convey and whether they satisfy the semantic constraints. This process is performed for each candidate sentence based on their extraction scores until the maximum summary length is reached. Since the use of schemata alone is not sufficient to achieve a linear sentence order (for example, several sentences may fill a particular slot) we have used post-schemata heuristics to improve this partial order and coherence. These heuristics include: topical similarity, explicit discourse markers and aggregation, and context.

1. **Topical Similarity**: This heuristic was inspired by the work of (Barzilay *et al.*, 2002) who demonstrated that even if human written summaries may have different discourse structures, topically similar sentences tend to stay together. Based on this observation, when selecting sentences for a particular predicate type (e.g. *attributive*) for a selected schema (e.g. *reason*), we tried to group together sentences that describe the same topic. In principle, this should prevent the summary from going back and forth on various topics and hence improve its coherence further.

2. **Explicit Discourse Markers and Aggregation**: To further improve discourse coherence, we added explicit discourse markers, a strategy which has also been used by other researchers (e.g. (Grote, Stede, 1998; Knott, Dale, 1993)) successfully. The choice of the discourse marker from a predefined set is based on the sentences' topical similarity and polarity value.

For example, if two sentences are topically similar and have identical polarity, our approach will place them next to each other and make a single sentence out of them using a discourse marker (e.g. *and*) even though these sentences may not be adjacent in the original candidate list. If our approach finds another sentence on this topic, it will position that sentence together using another discourse marker. In our work, we used the MPQA subjectivity lexicon[9], which contains more than 8000 entries of polarity words, to identify the polarity class of sentences.

3. **Context:** To improve discourse coherence further, if a potential sentence starts with a pronoun without having a potential antecedent, we include its previous sentence from the source document in the final summary. More sophisticated approaches, such as probabilistic models (Barzilay *et al.*, 2002; Lapata, 2003) could be used, but we have found that this simple heuristic increased coherence with very little processing cost.

At the end of the sentence ordering process, to create the final summary, we finally use the rank of the sentences in the original list of candidates.

9. MPQA: http://www.cs.pitt.edu/mpqa

4.5. An Example to Describe Content Organization

To better illustrate the entire content organization process (Sections 4.1 to 4.4), let us take the example shown in Figure 6.

Topic: *Carmax*
Question: *What motivated positive opinions of Carmax from car buyers?*

Candidate Sentences	Score	Rhetorical Predicate
(1) With Carmax you will generally always pay more than from going to a good used car dealer.	0.536	Comparison, contingency
(2) We bought it at Carmax, and I continue to have nothing bad to say about that company.	0.449	Topic-opinion, illustration
(3) Carmax did split the bill (which made me happy).	0.416	Topic-opinion
(4) Not sure if you have a Carmax near you, but I've had 2 good buying experiences from them.	0.381	Topic-opinion, illustration
(5) have to say that Carmax rocks.	0.368	Topic-opinion
(6) At Carmax, the price is the price and when you want a car you go get one.	0.299	Attributive, illustration
(7) Sometimes I wonder why all businesses can't be like Carmax.	0.278	Comparison
(8) Arthur Smith, 36, has been living in a van outside the Carmax lot, 24 hours a day, for more than a month.	0.250	Illustration

Figure 6. Candidate Sentences along with Rhetorical Predicates

Here, the topic is "*Carmax*" and the question is "*What motivated positive opinions of Carmax from car buyers*?", the 8 most relevant sentences along with their scores (out of 1) are shown in Figure 6[10]. The question categorization module classifies the above question as a reason type based on the question pattern matching (discussed in Section 4.1). Then the predicate identification module tags each of the candidate sentence with the rhetorical predicates they contain. For this question, the *Reason* schema shown in Figure 5 is used to filter and order the sentences. The *Reason* schema and the final order of the sentences are shown in Figure 7.

In this sample, we can see that the schema did not include sentences 1 and 8 in the final summary. This is because these sentences did not fit within the *Reason* schema.

10. In order to test our approach, we used items 1, 2, and 3 of Section, that were provided by the TAC 2008 dataset and our own sentence extractor. Further tests have also been done with the OpinRank dataset and (Jindal, Liu, 2006)'s dataset of reviews (see Section 5.2.3).

Though sentence 1 was classified as containing a *comparison* predicate, it did not fulfil the semantic constraint (shown in Figure 7) that the topic of the sentence (Carmax) be the focus of the sentence[11]. On the other hand, sentence 8 was not included, because it did not contain any of the rhetorical predicates which can fill the slots of this schema. This scenario shows that schemata help remove question irrelevant sentences, but Section 5 will provide a more formal evaluation.

Schema	Summary
Predicate: {*Topic-opinion*/*Attribution*}$^{+}$ Constraint: Sentence Polarity.	(2-1) After our last big car milestone, we've had an odyssey with cars. (2, 4) We bought it at Carmax, and I continue to have nothing bad to say about that company; not sure if you have a Carmax near you, but I've had good experiences from them. (3) Moreover, Carmax did split the bill (which made me happy). (5) have to say that Carmax rocks.
Predicate: {*Contingency*/*Comparison*}* Constraint: Sentence Focus, Compared Objects.	(7) Sometimes I wonder why all businesses can't be like Carmax.
Predicate: *Attributive** Constraint: Sentence Focus.	(6) At Carmax, the price is the price and when you want a car you go get one.

Figure 7. Summary Generated using the Reason Schema

We can see that since for sentence 2, the antecedent of the pronoun *it* is missing, our approach added the preceding sentence (2-1) of sentence 2 from the source document. Our approach placed sentences 2 and 4 next to each other because of their topical similarity and also merged them using a semi-colon as a conjunctive marker. We can also see that the system added the discourse marker "Moreover" in sentence 3. In the summary, sentences 6 and 7 are also reordered compared to the original candidate list based on the rhetorical predicates that they contained. This example shows that schema can help to reduce question irrelevant sentences and to improve discourse coherence; however, the next section will provide a more formal evaluation.

11. To identify this, we test if the topic is the subject or object of the sentence.

5. Evaluation

To evaluate our summarization approach, we have built a system called BlogSum. BlogSum is implemented using java and using third-party tools such as the Stanford parser[12], the SPADE parser[13], WordNet lemmatizer[14], and uses the Weka toolbox[15].

BlogSum-generated summaries have been evaluated for content and discourse coherence[16]. The content evaluation gives an indication of the question relevance of the summary as well as the usefulness of our approach and the evaluation of discourse coherence gives an indication of the coherence of the summary. The evaluation of the content was done both automatically and manually and the evaluation of the coherence was done manually for lack of such an automatic tool.

5.1. Baseline

In our evaluation, BlogSum-generated summaries were compared with the original candidate list (OList) generated by our approach without the discourse re-ordering. In order to verify how OList compared with other possible baselines, we have compared it to MEAD (Radev, D. and Allison, T. and Blair-Goldensohn, S. *et al.*, 2004), a widely used publicly available summarizer[17]. In this evaluation, we have generated summaries using MEAD with centroid, query title, and query narrative features. In MEAD, query title and query narrative features are implemented using cosine similarity based on the *tf-idf* value. In this evaluation, we used the TAC 2008 opinion summarization dataset (described later in this section) and summaries were evaluated using the ROUGE-2 and ROUGE-SU4 scores. The ROUGE-2 score is based on the overlap of word bi-grams between the automatically generated summaries and gold standard summaries (Lin, 2004). The ROUGE-SU4 score is also based on the overlap of bi-grams between summaries but allows a maximum gap of 4 tokens between the two tokens in a bi-gram (skip-bi-gram), and includes uni-gram co-occurrence statistics as well (Lin, 2004). Table 1 shows the results of the automatic evaluation using ROUGE based on summary content.

Table 1 shows that MEAD-generated summaries achieved weaker ROUGE scores compared to that of our candidate list (OList). The table also shows that MEAD

12. The Stanford parser: http://nlp.stanford.edu/software/lex-parser.shtml
13. The SPADE parser: http://www.isi.edu/licensed-sw/spade
14. WordNet: http://wordnet.princeton.edu
15. Weka: http://www.cs.waikato.ac.nz/ml/weka/
16. We did not evaluate readability or overall responsiveness of BlogSum-generated summaries because to evaluate readability or overall responsiveness manual evaluation need to be performed and these results cannot be compared across the actual TAC results because of the different group of assessors. There will be a chance that people might use these results for comparison. For the same reason we only calculated ROUGE score instead of the Pyramid scores.
17. http://www.summarization.com/mead

Table 1. Comparison of Possible Baselines on TAC 2008

System	ROUGE-2 (F-Measure)	ROUGE-SU4 (F-Measure)
MEAD	0.0407	0.0642
Average	0.0690	0.0860
OList	0.1020	0.1070

performs weaker than the average performance of the participants of TAC 2008 (Average). On the other hand, the performance of OList was better than the average performance of the participants of TAC 2008. For this reason, the rest of the evaluation was performed using OList as a baseline in order to be more strict on our approach.

5.2. Evaluation of Content

5.2.1. Automatic Evaluation of Content

First, we have automatically evaluated the summaries generated by our approach for content. As mentioned earlier, we used the OList as a baseline and compared them to the final summaries after the discourse structuring. We have used the data from the TAC 2008 opinion summarization track for the evaluation. The data set consists of 50 questions on 28 topics; on each topic one or two questions are asked and 10 to 50 relevant documents are given. In this experiment, we used the ROUGE metric, which is a standard automatic summary content evaluation metric, using answer nuggets (provided by TAC), which had been created to evaluate participants' summaries at TAC, as gold standard summaries. F-scores are calculated for BlogSum and OList using ROUGE-2 and ROUGE-SU4. In this experiment, ROUGE scores are also calculated for all 36 submissions in the TAC 2008 opinion summarization track.

The evaluation results are shown in Table 2. Note that in the table *Rank* refers to the rank of the system compared to the other 36 systems. Table 2 shows that BlogSum achieved a better F-Measure for ROUGE-2 and ROUGE-SU4 compared to OList. BlogSum gained 18% and 16% in F-Measure over OList using ROUGE-2 and ROUGE-SU4, respectively. Compared to the other systems, BlogSum (based on OList) performed very competitively; it ranked third and its F-Measure score difference from the best system is very small. Both BlogSum and OList performed better than the average systems.

A further manual analysis shows that BlogSum reduced the number of question irrelevant sentences from OList by 21%. However, BlogSum still contains a number of question irrelevant sentences. We need to investigate the reasons for the presence of these sentences; it could be that incorrect results of other intermediate tasks such as predicate identification, polarity identification, or design of the schema result in these irrelevant sentences. We have also found that BlogSum missed many relevant sentences. A further investigation has revealed that since BlogSum does not perform anaphora resolution, it misses question relevant sentences occasionally. For example, the sentence "*It systematically singles out Israel for discriminatory treatment*." is a

Table 2. Automatic Evaluation of BlogSum based on Content

System Name	ROUGE-2 (F-Measure)	ROUGE-SU4 (F-Measure)	Rank
MEAD	0.041	0.064	
TAC Average	0.069	0.086	
OList - Baseline	0.102	0.107	10
TAC Best	0.130	0.139	1
BlogSum based on OList	0.125	0.128	3
BlogSum based on TAC Best	0.138	0.151	<1

relevant sentence for the question "*What reasons are given as examples of UN commission's ineffectiveness?*". But BlogSum missed the sentence because it does not attempt to identify the referent for the pronoun "*it*".

We have also tried to verify if we feed the summaries of the best participant at the TAC 2008 opinion summarization track to BlogSum as the candidate set (instead of OList) can BlogSum improve those summaries further. The results of this evaluation, shown in Table 2, indicate that BlogSum can indeed further improve the output of a high performing summarizer, hence the schemas do improve the state of the art.

However, a further analysis of the results of Table 2 shows that there is no significant difference between BlogSum-generated summaries and OList summaries using the t-test with a *p-value* of 0.228 and 0.464 for ROUGE-2 and ROUGE-SU4, respectively. However, based on DUC and TAC evaluation results, (Conroy, Dang, 2008; Dang, Owczarzak, 2008) showed that the performance gap between humans-generated summaries and system-generated summaries at DUC and TAC is clearly visible in a manual evaluation, but is often not reflected in automated evaluations using ROUGE scores. Based on these findings, we suspected that there might be a performance difference between BlogSum-generated summaries and OList which is not reflected in ROUGE scores. To verify our suspicion, we have conducted manual evaluations for content. We have conducted two manual evaluations using two different datasets to better quantify BlogSum-generated summary content.

5.2.2. Manual Evaluation of Content using the Blog Dataset

In the first evaluation, we have again used the TAC 2008 opinion summarization track data. For each question, one summary was generated by OList and one by BlogSum (based on OList) and the maximum summary length was again restricted to 250 words. To evaluate coherence, 3 participants manually rated 50 summaries from OList and 50 summaries from BlogSum using a blind evaluation. These summaries were rated on a likert scale of 1 to 5 where 1 refers to "very poor" and 5 refers to "very good". Evaluators rated each summary with respect to the question for which it was generated and against the reference summary. In this experiment, we have used the answer nuggets provided by TAC as the reference summary, which had been created to evaluate participants' summaries at TAC.

In this evaluation, we have calculated the average scores of all 3 annotators' ratings to a particular question to compute the score of BlogSum for a particular question. Table 3 shows the performance comparison between BlogSum and OList. The results show that 58% of the time BlogSum summaries were rated better than OList summaries which implies that 58% of the time, our approach has improved the question relevance compared to that of the original candidate list (OList).

Table 3. Comparison of OList and BlogSum based on the Manual Evaluation of Summary Content on TAC 2008

Comparison	**%**
BlogSum Score > OList Score	58%
BlogSum Score = OList Score	30%
BlogSum Score < OList Score	12%

Table 4 shows the performance of BlogSum versus OList on each likert scale; where Δ shows the difference in performance. Table 4 demonstrates that 52% of the times, BlogSum summaries were rated as "very good" or "good", 26% of the times they were rated as "barely acceptable" and 22% of the times they were rated as "poor" or "very poor". From Table 4, we can also see that BlogSum outperformed OList in the scale of "very good" and "good" by 8% and 22%, respectively; and improved the performance in "barely acceptable", "poor", and "very poor" categories by 12%, 8%, and 10%, respectively.

Table 4. Manual Evaluation of BlogSum and OList based on Summary Content on TAC 2008

Category	**OList**	**BlogSum**	Δ
Very Good	6%	14%	8%
Good	16%	38%	22%
Barely Acceptable	38%	26%	-12%
Poor	26%	18%	-8%
Very Poor	14%	4%	-10%

We have also calculated whether there is any performance gap between BlogSum and OList. The *t*-test results show that in a two-tailed test, BlogSum performed significantly better than OList with a *p*-value of 0.00281.

Whenever human performance is computed by more than one person, it is important to compute inter-annotator agreement. This ensures that the agreement between annotators did not simply occur by chance. In this experiment, we have also calculated the inter-annotator agreement using Cohen's kappa coefficient to verify the annotation subjectivity. We have found that the average pair-wise inter-annotator agreement is moderate according to (Landis, Koch, 1977) with the kappa-value of 0.58.

5.2.3. Manual Evaluation of Content using the Review Dataset

Although our approach is generic, some of our development was based on the TAC 2008 dataset. Therefore, we wanted to make sure that it behaved well with other datasets. To verify this, we have conducted a second evaluation using the OpinRank dataset[18] and (Jindal, Liu, 2006)'s dataset on reviews (together referred as the Review dataset) to evaluate BlogSum-generated summary content. These datasets have been chosen because they are also informal and opinionated in nature. The OpinRank dataset contains reviews on cars and hotels collected from Tripadvisor (about 259,000 reviews) and Edmunds (about 42,230 reviews). The OpinRank dataset contains 42,230 reviews on cars for different model-years and 259,000 reviews on different hotels in 10 different cities. For this dataset, we created a total of 21 questions including 12 reason questions and 9 suggestions. For each question, 1500 to 2500 reviews were provided as input documents to create the summary.

(Jindal, Liu, 2006)'s dataset consists of 905 comparison and 4985 non-comparison sentences. Four human annotators labeled these data manually. This dataset consists of reviews, forum, and news articles on different topics from different sources. We have created 9 comparison questions for this dataset. For each question, 700 to 1900 reviews were provided as input documents to create the summary. For each question, one summary was generated by OList and one by BlogSum and the maximum summary length was restricted to 250 words. To evaluate question relevance, 3 participants manually rated 30 summaries from OList and 30 summaries from BlogSum using a blind evaluation. These summaries were rated on a likert scale of 1 to 5 again. Evaluators rated each summary with respect to the question for which it was generated.

Table 5 shows the performance comparison between BlogSum and OList. The results show that 67% of the time BlogSum summaries were rated better than OList summaries. The table also shows that 30% of the time both approaches performed equally well and 3% of the time BlogSum was weaker than OList.

Table 5. Comparison of OList and BlogSum based on the Manual Evaluation of Summary Content on the Review Dataset

Comparison	**%**
BlogSum Score > OList Score	67%
BlogSum Score = OList Score	30%
BlogSum Score < OList Score	3%

Table 6 demonstrates that 44% of the time BlogSum summaries were rated as "very good", 33% of the time rated as "good", 13% of the time they were rated as "barely acceptable" and 10% of the time they were rated as "poor" or "very poor". From Table 6, we can also see that BlogSum outperformed OList in the scale of "very

18. OpinRank Dataset: http://kavita-ganesan.com/entity-ranking-data

good" by 34% and improved the performance in "poor" and "very poor" categories by 23% and 10%, respectively.

Table 6. Manual Evaluation of BlogSum and OList based on Summary Content on the Review Dataset

Category	**OList**	**BlogSum**	**Δ**
Very Good	10%	44%	34%
Good	37%	33%	-4%
Barely Acceptable	10%	13%	3%
Poor	23%	0%	-23%
Very Poor	20%	10%	-10%

In this evaluation, we have also calculated whether there the performance gap between BlogSum and OList statistically significant. The *t*-test results show that in a two-tailed test, BlogSum performed significantly very better than OList with a *p*-value of 0.00236. In addition, we have found that the average pair-wise inter-annotator agreement is substantial according to (Landis, Koch, 1977) with the kappa-value of 0.77.

Figure 8 compares the results of the two manual experiments for content using the TAC 2008 dataset and the Review dataset. In the experiment with the review dataset, 44% of the time BlogSum-generated summaries were rated as "very good" whereas 14% of the time BlogSum-generated summaries were rated "very good" for the TAC 2008 dataset. For the review dataset, only 23% of the time BlogSum-generated summaries rated as "acceptable", "poor" and "very poor". On the other hand, for the TAC 2008 dataset, 48% of the time BlogSum-generated summaries rated as "acceptable", "poor" and "very poor". These results indicate that blogs contain more question irrelevant sentences compared to reviews.

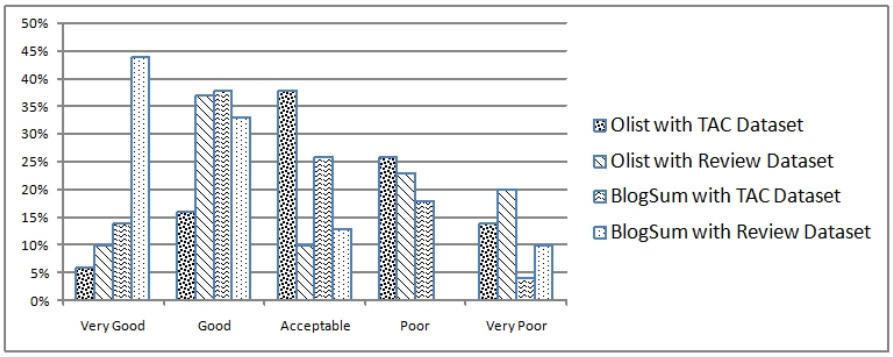

Figure 8. Comparison of the TAC and Review Dataset for Content Evaluation

In both manual evaluation for content, BlogSum performed significantly better than OList. We can see that even though there was not any significant performance gap between BlogSum and OList-generated summaries in the automatic evaluation,

the manual evaluation shows that BlogSum summaries are significantly better at the content level than OList.

5.3. Evaluation of Discourse Coherence

The second type of evaluation that we performed was geared at measuring automatically generated summaries for coherence. As a baseline, we used the original ranked list of candidate sentences again (OList), and we again compared it to the final summaries (BlogSum based on OList).

5.3.1. Automatic Evaluation of Discourse Coherence

To test the summary coherence, we did not use an automatic evaluation because, (Blair-Goldensohn, McKeown, 2006) found that the ordering of content within the summaries is an aspect which is not evaluated by ROUGE. Moreover, in the TAC 2008 opinion summarization track, on each topic, answer snippets were provided which had been used as summarization content units (SCUs) in pyramid evaluation to evaluate TAC 2008 participants summaries but no complete summaries is provided to which we can compare BlogSum-generated summaries for coherence. As a result, we only performed two manual evaluations using two different datasets again to see whether BlogSum performs significantly better than OList for summary coherence too.

5.3.2. Manual Evaluation of Discourse Coherence using the Blog Dataset

In this evaluation, we have again used the TAC 2008 opinion summarization track data. For each question, one summary was generated by OList and one by BlogSum and the maximum summary length was restricted to 250 words again. Four participants manually rated 50 summaries from OList and 50 summaries from BlogSum for coherence with respect to the question for which the summary is generated using a blind evaluation. These summaries were again rated on a likert scale of 1 to 5. To compute the score of BlogSum for a particular question, we calculated the average scores of all annotators' ratings to that question. Table 7 shows the performance comparison between BlogSum and OList. We can see that 52% of the time BlogSum summaries were rated better than OList summaries; 30% of the time both performed equally well; and 18% of the time BlogSum was weaker than OList. This means that 52% of the time, our approach has improved the coherence compared to that of the original candidate list (OList).

Table 7. Comparison of OList and BlogSum based on the Manual Evaluation of Discourse Coherence on TAC 2008

Comparison	**%**
BlogSum Score > OList Score	52%
BlogSum Score = OList Score	30%
BlogSum Score < OList Score	18%

From Table 8, we can see that BlogSum outperformed OList in the scale of "very good" and "good" by 16% and 8%, respectively; and improved the performance in "barely acceptable" and "poor" categories by 12% and 14%, respectively.

Table 8. Manual Evaluation of BlogSum and OList based on Discourse Coherence on TAC 2008

Category	**OList**	**BlogSum**	Δ
Very Good	8%	24%	16%
Good	22%	30%	8%
Barely Acceptable	36%	24%	-12%
Poor	22%	8%	-14%
Very Poor	12%	14%	2%

The *t*-test results show that in a two-tailed test, BlogSum performed significantly better than OList with a *p*-value of 0.0223. In addition, the average pair-wise inter-annotator agreement is substantial according to with the kappa-value of 0.76.

5.3.3. Manual Evaluation of Discourse Coherence using the Review Dataset

In this evaluation, we have again used the OpinRank dataset and (Jindal, Liu, 2006)'s dataset to conduct the second manual evaluation of content. In this evaluation, for each question, one summary was generated by OList and one by BlogSum and the maximum summary length was restricted to 250 words. Three participants manually rated 30 summaries from OList and 30 summaries from BlogSum for coherence with respect to the question for which the summary is generated.

To compute the score of BlogSum for a particular question, we calculated the average scores of all annotators' ratings to that question. Table 9 shows the performance comparison between BlogSum and OList. We can see that 57% of the time BlogSum

Table 9. Comparison of OList and BlogSum based on the Manual Evaluation of Discourse Coherence on the Review Dataset

Comparison	**%**
BlogSum Score $>$ OList Score	57%
BlogSum Score = OList Score	20%
BlogSum Score $<$ OList Score	23%

summaries were rated better than OList summaries; 20% of the time both performed equally well; and 23% of the time BlogSum was weaker than OList.

From Table 10, we can see that BlogSum outperformed OList in the scale of "very good" and "good" by 10% and 16%, respectively; and improved the performance in "barely acceptable" and "very poor" categories by 10% and 16%, respectively.

We have also evaluated if the difference in performance between BlogSum and OList was statistically significant. The *t*-test results show that in a two-tailed test,

Table 10. Manual Evaluation of BlogSum and OList based on Discourse Coherence on the Review Dataset

Category	**OList**	**BlogSum**	Δ
Very Good	13%	23%	10%
Good	27%	43%	16%
Barely Acceptable	27%	17%	-10%
Poor	10%	10%	0%
Very Poor	23%	7%	-16%

BlogSum performed significantly better than OList with a *p*-value of 0.0371. The average pair-wise inter-annotator agreement is substantial according to (Landis, Koch, 1977) with the kappa-value of 0.74.

Results from Table 9 show that in 23% of the time, our approach was weaker than OList. We believe that the reason behind these are wrong polarity identification, wrong predicate identification, and wrong results of the post-schemata heuristics.

Figure 9 compares the results of the two manual experiments for discourse coherence using the TAC 2008 dataset and the review dataset. The evaluation of coherence shows that for the blog dataset, BlogSum-generated summaries were rated as "good" and "very good" 50% of the time compared to 66% of the time for the review dataset. From the evaluation of content, we have seen that summaries generated from the blog dataset contain more question relevance compared to that of summaries generated for the review dataset. We suspect that for the blog dataset, question irrelevant sentences make the improvement of summary coherence a difficult task.

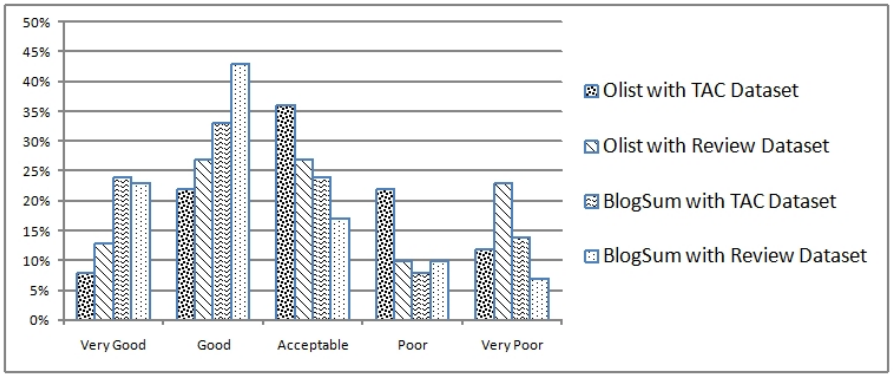

Figure 9. Results Comparison of the TAC and Review Dataset for Discourse Coherence Evaluation

Results of both manual evaluation of discourse coherence also show that BlogSum performs significantly better than OList.

6. Conclusion and Future Work

In this work, we have presented an approach to use discourse structures with the help of schema to improve question relevance and discourse coherence of query-based blog summaries. This approach is based on the automatic identification of rhetorical predicates within candidate sentences in order to instantiate the most appropriate discourse schema and filter and order candidate sentences in the most efficient way to achieve the communicative goal of the summary. We have developed a query-based blog summarization system called BlogSum to validate our approach and evaluated its performance for question relevance and coherence using two different datasets: the TAC 2008 opinion summarization dataset (a blog dataset) and the OpinRank dataset and (Jindal, Liu, 2006)'s dataset on reviews (a review dataset).

We have calculated BlogSum's performance for question relevance automatically with the TAC dataset using the ROUGE scores and also manually with the TAC dataset and the review dataset on a likert scale 1 to 5. We have evaluated BlogSum's performance for discourse coherence manually again with the TAC dataset and the review dataset on a likert scale 1 to 5.

The automatic evaluation of content shows that BlogSum achieved a performance gain of about 18% over the original ranked list in removing question irrelevant sentences. From the results we have also seen that our approach performed very competitively (positioned at rank 3) compared to all 36 participants in TAC-2008. The manual evaluation shows that our approach performs significantly better than the original candidate list for question relevance and coherence. These results show that our approach can effectively reduce question irrelevance and discourse incoherence of automated summaries even in the case of informal and opinionated documents. Evaluation results also demonstrate that our approach works well for blogs but also for any type of informal opinionated documents such as reviews.

In the future, we plan to evaluate the individual contribution of intermediate steps such as the schemata design and the post-schema heuristics to the overall coherence of the summaries. It would be interesting to quantify how effective is a schema for a specific question type. Moreover, we would be curious to see how the design of a schema influences the quality of the summary. By doing that we want to make sure that the schema design is optimal. Another interesting research topic is to consider the automatic acquisition of schema from corpora. The problems of question irrelevance and incoherence are not limited to text summarization, but are also a concern in other applications such as natural language generation and question answering. Another research avenue would be to apply our schema-based approach for question answering or natural language generation.

Acknowledgment

The authors would like to thank the anonymous referees for their valuable comments. This work was financially supported by NSERC.

References

Aristotle. (1954). *The rhetoric. translation in the rhetoric and the poetics of aristotle* (W. R. Roberts, Ed.). New York, Random House.

Barzilay R., Elhadad N., McKeown K. R. (2002). Inferring Strategies for Sentence Ordering in Multidocument News Summarization. *Artificial Intelligence Research*, Vol. 17, pp. 35–55.

Barzilay R., Lee L. (2004). Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization. In *Proceedings of HLT-NAACL*, pp. 113–120. Boston.

Blair-Goldensohn S. J. (2007). *Long-answer Question Answering and Rhetorical-semantic Relations*. Unpublished doctoral dissertation, Dept. of Computer Science, Columbia University, USA.

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

Bosma W. (2004). Query-Based Summarization using Rhetorical Structure Theory. In 15^{th} *Meeting of Computational Linguistics in the Netherlands CLIN*, pp. 29–44. Netherlands.

Bossard A., Genereux M. (2008). Description of the LIPN Systems at TAC 2008: Summarizing Information and Opinions. In *Proceedings of TAC-2008.* Gaithersburg, USA.

Chali Y., Hasan S. A., Joty S. R. (2009). Complex Question Answering: Unsupervised Learning Approaches and Experiments. *Journal of Artificial Intelligence Research*, Vol. 35, pp. 1–47.

Cline B. E., Nutter J. T. (1994). Kalos - A System for Natural Language Generation with Revision. In *AAAI'94: Proceedings of the Twelfth National Conference on Artificial Intelligence (vol. 1)*, pp. 767–772. Seattle, Washington, USA.

Conroy J. M., Dang H. T. (2008). Mind the Gap: Dangers of Divorcing Evaluations of Summary Content from Linguistic Quality. In *COLING-2008*, pp. 145–152. Manchester, UK.

Conroy J. M., Schlesinger J. D. (2008). CLASSY and TAC 2008 Metrics. In *Proceedings of the Text Analysis Conference.* Gaithersburg, USA.

Dang H. T., Owczarzak K. (2008). Overview of the TAC 2008 Update Summarization Task. In *Proceedings of the Text Analysis Conference.* Gaithersburg, USA.

Fei Z., Huang X., Wu L. (2006). Mining the Relation between Sentiment Expression and Target Using Dependency of Words. In *PACLIC20: Coling 2008: Proceedings of the* 20^{th} *Pacific Asia Conference on Language, Information and Computation*, pp. 257–264. Wuhan, China.

Genest P., Lapalme G., Yousfi-Monod M. (2009). HEXTAC: the Creation of a Manual Extractive Run. In *Proceedings of the Text Analysis Conference.* Gaithersburg, USA.

Grimes J. E. (1975). *The Thread of Discourse*. Technical report Nos. NSF-TR-1, NSF-GS-3180. Cornell University, Ithaca, New York.

Grosz B. J. (1985). Discourse Structure and the Proper Treatment of Interruptions. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pp. 832–839.

Grote B., Stede M. (1998). Discourse Marker Choice in Sentence Planning. In *International Workshop on Natural Language Generation*, pp. 128–137. Niagara-on-the-Lake, Canada.

Harabagiu S. M., Lacatusu V. F., Hickl A. (2006). Answering complex questions with random walk models. In *Proceedings of SIGIR-2006*, pp. 220–227. Washington, USA.

Hendrickx I., Bosma W. (2008). Using Coreference Links and Sentence Compression in Graph-based Summarization. In *Proceedings of the Text Analysis Conference.* Gaithersburg, USA.

Hobbs J. R. (1985). *On the Coherence and Structure of Discourse*. Technical report No. CSLI-85-37. Center for the Study of Language and Information, Stanford University.

Hovy E. H. (1993). Automated Discourse Generation using Discourse Structure Relations. *Artificial Intelligence*, Vol. 63, No. 1-2, pp. 341–385.

Hu M., Liu B. (2004). Mining and Summarizing Customer Reviews. In *Proceedings of the* 10^{th} *ACM SIGKDD*, pp. 168–177.

Jaidka K., Khoo C. S. G., Na J. (2010). Imitating Human Literature Review Writing: An Approach to Multi-document Summarization. In *Proceedings of the* 12^{th} *International Conference on Asia-Pacific Digital Libraries*, pp. 116–119. Gold Coast, Australia.

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

Kim H. D., Park D. H., Vydiswaran V. G. V., Zhai C. (2008). Opinion Summarization using Entity Features and Probabilistic Sentence Coherence Optimization: UIUC at TAC 2008 Opinion Summarization Pilot. In *Proceedings of the Text Analysis Conference.* Gaithersburg, USA.

Knott A., Dale R. (1993). Choosing a Set of Coherence Relations for Text Generation: A Data-Driven Approach. In *Proceedings of the Fourth European Workshop on Trends in Natural Language Generation An Artificial Intelligence Perspective*, pp. 47–67. Pisa, Italy.

Ku L., Lee L., Chen H. (2006). Opinion Extraction, Summarization and Tracking in News and Blog Corpora. In *Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs.* Stanford, USA.

Landis R. J., Koch G. G. (1977). A One-way Components of Variance Model for Categorical Data. *Biometrics*, Vol. 33, No. 1, pp. 671–679.

Lapata M. (2003). Probabilistic Text Structuring: Experiments with Sentence Ordering. In *Proceedings of ACL*, pp. 545–552. Sapporo, Japan.

Lin C. Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pp. 74–81. Barcelona.

Liu M., Li W., Wu M., H. H. (2007). Summarization using Event Semantic Relevance from External Linguistic Resource. In *Proceedings of the Sixth International Conference on Advanced Language Processing and Web Information Technology, ALPIT*, pp. 117–122. Henan, China.

Mani I., Bloedorn E., B. G. (1998). Using Cohesion and Coherence Models for Text Summarization. In *Proceedings of the Spring Symposium on Intelligent Text Summarization (AAAI 98)*, pp. 69–76. Stanford, CA, USA.

Mann W. C., Thompson S. A. (1988). Rhetorical Structure Theory: Toward a Functional Theory of Text Organisation. *Text*, Vol. 3, No. 8, pp. 234–281.

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

McKeown K. (1985). Discourse Strategies for Generating Natural-Language Text. *Artificial Intelligence*, Vol. 27, No. 1, pp. 1-41.

McKeown K., Klavans J., Hatzivassiloglou V., Barzilay R., Eskin E. (2002). Towards Multidocument Summarization by Reformulation: Progress and prospects. In *Proceedings of AAAI/IAAI*, pp. 27–36. Edmonton, Canada.

Mithun S., Kosseim L. (2009). Summarizing Blog Entries versus News Texts. In *Proceedings of Events in Emerging Text Types. A Workshop of RANLP*, pp. 35–42. Borovets, Bulgaria.

Mithun S., Kosseim L. (2011). Comparing Approaches to Tag Discourse Relations. In *Proceedings of CICLing*, pp. 328–339. Tokyo, Japan.

Mitkov R. (1993). How Could Rhetorical Relations be used in Machine Translation? (And at Least Two Open Questions). In *Proceedings of the workshop on intentionality and structure in discource relations*, pp. 86–89. Columbus, Ohio.

Murray G., Joty S., Carenini G., Ng R. (2008). The University of British Columbia at TAC 2008. In *Proceedings of the Text Analysis Conference.* Gaithersburg, USA.

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

Paul M., C. Z., Girju R. (2010). Summarizing Contrastive Viewpoints in Opinionated Text. In *Proceedings of EMNLP 2010*, pp. 65–75. Massachusetts.

Potthast M., Becker S. (2010). Opinion Summarization of Web Comments. In *Proceedings of the* 32^{nd} *European Colloquium on IR Research*, pp. 668–669.

Radev, D. and Allison, T. and Blair-Goldensohn, S. *et al.* (2004). MEAD-A Platform for Multidocument Multilingual Text Summarization. In *Proc. of LREC-2004*, pp. 1–4. Lisbon.

Sauper C., Barzilay R. (2009). Automatically Generating Wikipedia Articles: A Structure-Aware Approach. In *Proceedings of ACL-AFNLP-2009*, pp. 208–216. Suntec, Singapore.

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

Wang D., Liu Y. (2011). Pilot Study of Opinion Summarization in Conversations. In *Proceedings of ACL-HLT*, pp. 331–339. Oregon, USA.

Yang H., Chua T. S., Wang S., Koh C. K. (2003). Structured use of External Knowledge for Event-based Open Domain Question Answering. In *Proceedings of SIGIR-2003*, pp. 33–40. Toronto, Canada.

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