EXPLOITING RHETORICAL RELATIONS IN BLOG SUMMARIZATION

Shamima Mithun

A thesis

in

The Department

of

Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy

Concordia University

Montréal, Québec, Canada

September 2012

CONCORDIA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: Shamima Mithun

Entitled: Exploiting Rhetorical Relations in Blog Summarization

and submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY (Computer Science)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

Dr. C. Mulligan — Chair

Dr. Guy Lapalme — External Examiner

Dr. N. Bouguila — External to Program

Dr. S. Bergler — Examiner

Dr. G. Butler — Examiner

Dr. L. Kosseim — Thesis Supervisor

Approved by

Chair of Department or Graduate Program Director

Aug 17, 2012

Dean of Faculty

Abstract

Exploiting Rhetorical Relations in Blog Summarization

Shamima Mithun, Ph.D.

Concordia University, 2012

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. A query-based opinion summarizer from opinionated documents can address this need. Query-based opinion summarizers present what people think or feel on a given topic in a condensed manner to analyze others' opinions regarding a specific question (e.g. *Why do people like Chrome better than Firefox?*). This research interest motivated us to develop an effective query-based extractive multi-document opinion summarization approach for blogs.

The goal of this thesis is to design an effective extractive query-based summarization approach for blogs and evaluate it experimentally using current benchmarks. Since blog summarization is a more recent endeavor compared to news summarization, the current state of the art is typically weaker than for news summarizers. We first tried to identify and categorize problems which typically occur in opinion summarization through an error analysis of the state of the art blog summarizers. In this error analysis, we compared blog summaries with news summaries to assess whether there is any information processing difference needed for blogs. Evaluation results of various studies (e.g. [CD08, GLYM09]) as well as ours [MK09] show that *question irrelevance* and *discourse incoherence* are two major areas where automatic summaries need to be improved.

Question irrelevance and discourse incoherence are important and typical problems in multi-document summarization. These errors decrease the overall quality of a summary;

question irrelevance weakens the summary content and discourse incoherence reduces the summary coherence. To address these two issues, we have developed a schema-based summarization approach for query-based blog summaries that utilizes discourse structures. Our proposed approach is domain-independent and uses intra-sentential discourse structures in the framework of schemata. 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. To validate our approach, we have built a system named BlogSum and have evaluated its performance for question relevance and coherence using the TAC 2008 opinion summarization data. The evaluation results show the effectiveness of our approach in reducing question irrelevance sentences by about 18% using the ROUGE scores and in significantly improving summary content and coherence with a p-value of 0.00281 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 [JL06]'s dataset of reviews for summary content 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 summary content and coherence, respectively. These results show that our approach can effectively reduce question irrelevance and discourse incoherence of automatic summaries.

Acknowledgments

I would like to thank Dr. Leila Kosseim, my academic adviser, for all her guidance, support, and enthusiasm. I particularly appreciated that despite her enormous workload she always made herself available to me and took the time to revise my thesis and papers for publications.

I would like to express my deep appreciation to the members of my doctoral committee: Professors Gregory Butler, Sabine Bergler, Nizar Bouguila, and Guy Lapalme for their valuable feedback on my thesis. Their comments, questions and suggestions have been very useful to improve my Ph.D. work.

I would like to thank my husband, Quamrul Islam Khan, for his endless love, support, and care during all these years.

I would like to thank my parents for their moral support and inspiration which have brought me here today.

Contents

List of Figures **x**

List of Tables **xiii**

1 Introduction **1**

1.1 Motivation 1

1.2 Question Irrelevance and Discourse Incoherence 4

1.2.1 Question Irrelevance 4

1.2.2 Discourse Incoherence 6

1.3 Overview of the Thesis 8

1.4 Intended Contributions 12

1.4.1 Theoretical Contributions 12

1.4.2 Practical Contributions 13

1.5 Thesis Organization 14

2 Background **15**

2.1 Automated Text Summarization 16

2.2 Query-based Multi-Document Extractive Summarization Approaches 19

2.2.1 Pre-processing 20

2.2.2 Relevance Calculation 21

2.2.3 Sentence Selection 23

2.2.4 Post-processing 27
2.2.5 Discussion 29
2.3 Work on Opinionated Texts 30
2.3.1 General Work on Blogs 31
2.3.2 Work on Query-based Blog Summarization 34
2.4 Discourse Relations and Schema-based Approaches 39
2.4.1 Discourse Relations 39
2.4.2 Schema-based Approaches 44
2.4.3 Discussion 46
2.5 Summary Evaluation 47
2.6 Conclusion 53

3 Blog-Specific Summarization Errors 55
3.1 News Summarization versus Blog Summarization 56
3.2 Related Work 59
3.3 Error Analysis 60
3.3.1 Summary-Level Errors 64
3.3.2 Sentence-Level Errors 67
3.3.3 Intra-Sentence-Level Errors 69
3.3.4 Discussion 72
3.4 Conclusion 72

4 A Schema-based Framework Utilizing Discourse Relations 74
4.1 Overview of Our Schema-based Approach 75
4.1.1 Candidate Sentence Selection 76
4.1.2 Content Filtering and Organization 77

4.2 An Example to Demonstrate the Usability of Our Schema-based Approach . . . 91
4.3 Conclusion . . . 94

5 Rhetorical Predicate Identification 95

5.1 Introduction . . . 95
5.2 Rhetorical Predicates . . . 98
5.3 Approaches to Rhetorical Predicate Identification . . . 102
5.3.1 Tagging Inter-Clausal Rhetorical Predicates . . . 104
5.3.2 Tagging Intra-Clausal Rhetorical Predicates . . . 106
5.4 Conclusion . . . 116

6 BlogSum: Our Prototype Blog Summarizer 117

6.1 Candidate Sentence Selection . . . 119
6.1.1 Blog Pre-processing . . . 119
6.1.2 Candidate Sentence Selection . . . 120
6.2 Content Filtering and Organization . . . 128
6.2.1 Question Categorization . . . 129
6.2.2 Predicate Identification . . . 132
6.2.3 Schema Selection from the Pre-designed Schemata . . . 137
6.2.4 Summary Generation . . . 138
6.3 Sample Summaries . . . 140
6.4 Conclusion . . . 141

7 Evaluation 142

7.1 Baseline . . . 143
7.2 Evaluation of Content . . . 144
7.2.1 Automatic Evaluation of Content . . . 144

7.2.2 Manual Evaluation of Content 149
7.3 Evaluation of Discourse Coherence 159
7.3.1 Automatic Evaluation of Discourse Coherence 159
7.3.2 Manual Evaluation of Discourse Coherence with the TAC 2008 Dataset 160
7.3.3 Manual Evaluation of Discourse Coherence with the Review Dataset 162
7.4 Evaluation of the Rhetorical Predicate Identification 165
7.4.1 Corpora and Experimental Design 165
7.4.2 Results 167
7.5 Effects of the Post-Schema Heuristics 173
7.5.1 Corpora and Experimental Design 173
7.5.2 Results 174
7.6 Conclusion 176

8 Discussion and Conclusion **178**
8.1 Main Findings and Contributions of the Thesis 179
8.1.1 Theoretical Contributions 179
8.1.2 Practical Contributions 182
8.2 Directions for Future Research 186
8.2.1 Extensions 186
8.2.2 Future Directions 189

Bibliography **190**

A Sample Summaries Generated by BlogSum **207**

B Sample Manual Evaluation for Content and Discourse Coherence **211**

List of Figures

1 Sample Summary Showing Question Irrelevance 5

2 Sample Summary Showing Discourse Incoherence 7

3 Sample Query-based Summary 19

4 Sample Blog Post from the BLOG06 Corpus 32

5 Definition of the *Evidence* Relation in RST (from [MT88]) 41

6 Sample RST Tree (from the RST corpus) 42

7 Identification Schema (from [McK85]) 45

8 Sample News Article from the AQUAINT-2 Collection 57

9 Sample Blog Post from the BLOG06 Corpus 58

10 Sample Summary from TAC 2008 Opinion Summarization Track . . . 62

11 Sample Summary from TAC 2008 Update Summarization Track . . . 63

12 Types of Errors in Blog vs. News Summaries 63

13 Partial Candidate List Used as Input 76

14 Architectural Design 77

15 Sample RST Tree 80

16 Candidate Sentences along with Rhetorical Predicates 81

17 The Reason Schema 83

18 Example of Constraint on Sentence Polarity 84

19 The Comparison Schema 85

20 The Suggestion Schema 86

21 Example of Topical Similarity from a News Article 88
22 Partial Candidate List Used as Input 91
23 Candidate Sentences along with Rhetorical Predicates 92
24 Summary Generated using the Reason Schema 93
25 Rhetorical Predicates that we Considered 102
26 Dependency Relations for the Sentence: *The movie was genuinely funny.* 108
27 Sample sentences from the Topic-opinion Dataset 109
28 Topic-opinion Dependency Relations Trees 109
29 Example of Topic-opinion Heuristic 1 110
30 Example of Topic-opinion Heuristic 3 111
31 Example of Topic-opinion Dependency Relations Tree 112
32 Sample Sentences from the Attributive Dataset 113
33 Attributive Dependency Relations Tree 113
34 Example of Heuristic 1 to Tag the Attributive Predicate 114
35 Example of Heuristic 2 to Tag the Attributive Predicate 115
36 Example of Heuristic 3 to Tag the Attributive Predicate 115
37 Detailed Architecture of BlogSum 118
38 Blog Pre-processing 120
39 Partial Candidate List Used as Input 127
40 Sample from the TAC 2008 opinion summarization Dataset 131
41 Sample Code for Reason Schema 138
42 Sample Summaries Generated by BlogSum 140
43 Sample DUC 2007 Dataset 147
44 Sample Summary Generated by BlogSum Used in the Manual Evaluation 150
45 Sample OList Summary Used in the Manual Evaluation 151
46 Sample TAC Best System Summary Used in the Manual Evaluation . 152

47 Results Comparison of the TAC and Review Dataset for Content Evaluation 158
48 Results Comparison of the TAC and Review Dataset for Discourse Coherence Evaluation 164
49 Effect of the Post-schema Heuristic Rules on the Summary Quality . 175
50 BlogSum-generated Sample Summary 1 207
51 BlogSum-generated Sample Summary 2 208
52 BlogSum-generated Sample Summary 3 208
53 BlogSum-generated Sample Summary 4 209
54 BlogSum-generated Sample Summary 5 209
55 BlogSum-generated Sample Summary 6 210
56 BlogSum-generated Sample Summary 7 210
57 Sample Manual Evaluation for Question Relevance and Discourse Coherence 211
58 Sample Summaries Distributed for Evaluation Generated by BlogSum and OList 212
59 Sample Dataset Distributed for Comparison Predicate Identification for the Manual Performance Evaluation 213
60 Sample Dataset Distributed for Attributive Predicate Identification for the Manual Performance Evaluation 214
61 Sample Dataset Distributed for Topic-Opinion Predicate Identification for the Manual Performance Evaluation 215

List of Tables

1 Our Work Situated Within the Summarization Dimensions 19

2 Linguistic Quality Scores of Automatic Summarizers at TAC 2008 . . 28

3 Examples of Word Polarity and Sentiment Degree in the MPQA Lexicon 35

4 Human and Automatic System Performance at Various TAC Competitions 51

5 Human and Automatic System Performance in ROUGE at TAC 2008 Update Summarization 52

6 TAC-2008 Summarization Results - Blogs vs. News 56

7 Summary-Level Errors - Blogs vs. News 66

8 Sentence-Level Errors - Blogs vs. News 68

9 Intra-Sentence-Level Errors - Blogs vs. News 71

10 TAC 2008 Question Distribution 79

11 Corpus Analyzed to Design Schemata 82

12 Discourse Markers 89

13 Sample Sentences 89

14 Frequency of Intra-Clausal Rhetorical Predicates 103

15 Rhetorical Predicates Considered and Identification Approaches Used 104

16 Sample Dependency Relations between Words (taken from [FHW06]) 108

17 Topic-opinion Heuristic Occurrence Distribution 111

18 Attributive Heuristic Occurrence Distribution 116

19 Comparison of Various Similarity Measures 122
20 Examples of Word Polarity and Subjectivity in the MPQA Lexicon . 123
21 Sentence Polarity Corresponding to Question Polarity 125
22 Accuracy of Our Polarity Identification Approach 126
23 Automatic Evaluation of MEAD based on Summary Content 128
24 Datasets used to Question Pattern Analysis 130
25 Lexico-syntactic Patterns for Question Categorization 132
26 Predicate Tagging Distribution based on Classifier Used 136
27 Predicate Tagging Distribution based on Number of Tags Occurs . . . 137
28 Evaluation Performed to Measure Summary Content and Coherence . 142
29 Comparison of Possible Baselines on TAC 2008 143
30 Automatic Evaluation of BlogSum based on Content 145
31 Automatic Evaluation of BlogSum based on Summary Content with the DUC 2007 Dataset 148
32 Automated vs. Manual Evaluation at TAC 2008 149
33 Comparison of the TAC Best System, OList, and BlogSum based on the Manual Evaluation of Summary Content with the TAC 2008 Dataset 153
34 Manual Evaluation of the TAC Best System, OList and BlogSum based on Summary Content with the TAC 2008 Dataset 154
35 Comparison of OList and BlogSum based on the Manual Evaluation of Summary Content with the Review Dataset 157
36 Manual Evaluation of OList and BlogSum based on Summary Content with the Review Dataset 157
37 Comparison of the TAC Best System, OList, and BlogSum based on the Manual Evaluation of Discourse Coherence with the TAC 2008 Dataset 161

38 Manual Evaluation of the TAC Best System, OList, and BlogSum based on Discourse Coherence with the TAC 2008 Dataset 161
39 Comparison of OList and BlogSum based on the Manual Evaluation of Discourse Coherence with the Review Dataset 163
40 Manual Evaluation of BlogSum and OList based on Discourse Coherence with the Review Dataset 163
41 Performance of Different Predicate Identification Approaches 168
42 Baseline and Human Performance of the Rhetorical Predicate Identification Approaches 171
43 Inter-Annotator Agreement on Predicate Tagging 173
44 Possible Configurations to Evaluate the Post-Schema Heuristics . . . 174
45 Configurations Used to Evaluate the Post-Schema Heuristics 174
46 Details of the Effect of the Post-schema Heuristic Rules on Summary Quality 175

Chapter 1

Introduction

1.1 Motivation

Because of the rapid growth of the Social Web, a large amount of informal opinionated texts are now easily available. These discuss a variety of topics ranging from politics, movies, music to newly launched products and are available in a variety of medias ranging from weblogs (or blogs), wikis, online-forums, review sites, twitter, and social networking web sites. Natural language tools for automatically analyzing these opinions become necessary to help individuals, organizations, and governments make 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. [McK85, YCWK03, RK08]) and summarization systems (e.g. [Mar97a, Bos04, BGM06, LLWH07]) 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. Dealing with opinionated texts (e.g. blogs or online reviews) is more challenging compared to fact-based texts because opinionated texts often contain subjective information such as writer's opinions, emotions, and speculations which are usually absent in fact-based texts. Natural language tools which deal with opinionated texts need to perform subjectivity or sentiment analysis accurately to be successful; however, subjectivity calculation is a difficult task on its own (see [PL08]). Moreover, opinionated texts are often informal in nature and written in casual and informal language. They may contain much and sometimes only unrelated information such as ads, photos, and other non-textual elements. They also contain spelling and grammatical errors, and punctuation and capitalization are often missing. Current state of the art natural language tools to process opinionated texts such as question answering systems (e.g. [LTHZ09]), summarizers (e.g. [MJCN08, BG08]) or opinion mining tools (e.g. [PL08]) have a much weaker performance than their manually-created counterparts. This lack of effective methods for dealing with opinionated texts motivated us to **develop a more effective query-based extractive multi-document opinion summarization approach for blogs**. Query-based opinion summarizers present what people think or feel on a given topic in a condensed manner to analyze others' opinions regarding a specific question (e.g. *"Why do people like Chrome better than Firefox?"*).

Over time, different extractive summarization techniques for factual texts (e.g. centroid-based [RABG$^+$04], graph-based [Bos04], machine learning-based [NFK02]) have been developed and evaluated. Although significant improvement continues to be made, the summaries generated automatically are by no means of the same quality as their manually-created counterparts. Opinion summarization approaches try

to use summarization techniques from fact-based summarization. However, evaluation results show that opinion summarizers are currently weaker than fact-based summarizers (see Section 3.1). Over the years, manual evaluation of the results of the Text Analysis Conference (TAC)[1] and the Document Understanding Conference (DUC)[2] also show that in both cases, factual summarization and opinionated summarization, system-generated summaries are significantly weaker than human-generated summaries [DO08, CD08]. For example, at TAC 2008, the average summary content evaluation scores in pyramid evaluation for opinion summarization of human-generated summaries and automated summaries were 0.446 and 0.102, respectively [Dan08] (these measures will be explained in Section 2.5).

In addition to a weaker summary content, previous studies (e.g. [DO08, CD08, GLYM09]) have also shown that the linguistic quality of system-generated summaries is significantly weaker than that of human-generated summaries and the area in which automatic summaries differ most from human-generated summaries is text coherence [ORL02, CD08, GLYM09]. Recently, [GLYM09] demonstrated that the performance of automatic summarizers in terms of linguistic quality is significantly weaker compared to that of a baseline consisting of sentences extracted from the source documents by 5 human extractors and added to the summary without any modification. This result indicates that there is still much space to improve the linguistic quality of summaries even for pure extractive summaries. All these studies indicate that extractive summaries need to be improved for both content and linguistic quality, especially coherence.

In query-based extractive summarization, one of the main causes of poor contents is *question irrelevance* [KLC06, MK09]; that is, the sentences extracted from the original documents as candidates to be included in the summary, are not relevant to the

[1]TAC: http://www.nist.gov/tac
[2]DUC: http://duc.nist.gov

query. On the other hand, coherence problems can be the result of different phenomena: discourse incoherence, redundancy, temporal incoherence, grammatical mistakes or many other linguistic problems [ORL02, MK09]. In a manual analysis of 15 summaries, [ORL02] showed that coherence problems are caused mostly by *discourse incoherence* (34%). Our work also shows that question irrelevance and discourse incoherence are the most frequently occurring errors in blog summaries [MK09] (see Section 3.3). As a result, an effective query-based extractive summarization approach needs to deal with question irrelevance and discourse incoherence.

1.2 Question Irrelevance and Discourse Incoherence

In this thesis, we are addressing two important problems: *question irrelevance* and *discourse incoherence*. Let us first define these two problems.

1.2.1 Question Irrelevance

A query-based summary is produced from a document or a set of documents to satisfy a request for information expressed by a question. If the sentences in a query-based summary are not relevant to the question, then the summary exhibits a question irrelevance. A question irrelevant summary does not fulfil the user's information need because it does not relate to his/her question.

Figure 1 shows a sample summary taken from the TAC 2008 opinion summarization track. Summary 1 contains question irrelevance 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

Figure 1: Sample Summary Showing Question Irrelevance

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. [...]

sentences are extracted from the source documents and then these sentences are ordered to create a summary. An inadequate content selection (sentence extraction) can result in question irrelevance.

To select sentences, current query-based summarization approaches typically use the similarity between the question and candidate sentences [MJCN08, JHZ09]. To calculate the similarity, both linguistic and statistical methods are used. Most of these approaches represent the question and sentences as bag-of-words to find the similarity between question and document sentences. As a result, sentences which share common words with the question will very likely be added in the summary. However, a sentence containing similar words to the question might be irrelevant to answer the question (e.g. see the second sentence of Summary 1 in Figure 1). Since current approaches do not consider how words are related to the topic of the sentence or do not consider the deeper semantic interpretation of a sentence, they often add question irrelevant sentences to the summary. Current approaches (e.g. [RABG$^+$04, MJCN08]) often utilize predefined features such as sentence position, sentence length or word frequency in the document to calculate the similarity between the question and the sentence. However, many of these features are not as useful for unstructured genres like blogs [BG08] because these texts do not have predictable discourse structures. In addition, the semantic category of the question (e.g. a comparison versus a reason), which is typically used by human writers to answer

a specific type of question, is ignored in current approaches. As a result, question irrelevance still remains an open issue for query-based blog summarization.

1.2.2 Discourse Incoherence

Computational theories on discourse coherence were introduced by [Hob85, MT88]. According to [MT88], a discourse is coherent if the hearer knows the communicative role of each of its portions; that is, if the hearer knows how the speaker intends each clause to relate to each other. For example, the sentence "*I'll have to cancel dinner tonight because I lost my car keys.*" is coherent because both clauses are related; clause *b* provides reasons to support clause *a*.

a. *[I'll have to cancel dinner tonight]*

b. *[because I lost my car keys.]*

A summary will exhibit discourse incoherence if the reader cannot identify the communicative intentions of the writer from the clauses or if the clauses do not seem to be interrelated. A summary with poor coherence confuses the readers and degrades the quality and readability of the summary. [Lap03] experimentally showed that the time to read a summary strongly correlates with the arrangement of sentences. In addition, [BEM02] has shown empirically that proper order significantly improves the readability of summaries.

Summary 2 in Figure 2 shows a sample summary that contains discourse incoherence. 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 (shown as Summary 3 and Summary 4 in Figure 2).

In extractive summarization, sentences may be selected from multiple documents or without consideration to their interdependency with other sentences. Moreover, in

Figure 2: Sample Summary Showing Discourse Incoherence

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

Summary 2:
(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.

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

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

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.

Two major types of sentence reordering approaches are used to address discourse incoherence: making use of chronological information [BEM02, MKH+02], and learning the natural order of sentences from large corpora [BL04, Lap03]. 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 the models will perform poorly. In other work, (e.g. [Bos04, BGM06, Mar97a, ZF11]) 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. Most of this work is developed for a particular domain or genre (e.g. news articles, scientific research papers). Some schema [McK85] and template-based approaches have

been used successfully in achieving coherence (e.g. [SB09, JKN10]); however, they are either domain dependent or applied to a very structured domain (e.g. Wikipedia pages).

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.

Since question irrelevance and discourse incoherence are the outcomes of an inadequate content selection and content organization of the extractive summarization approach, we believe that if the summary contents are selected properly and the selected contents are organized properly then question irrelevance and discourse incoherence could be significantly reduced.

1.3 Overview of the Thesis

The goal of this thesis is to design an effective extractive query-based summarization approach for blogs and evaluate it experimentally using current benchmarks. Since blog summarization is a more recent endeavor compared to news summarization and the current state of the art systems are typically weaker than news summarizers (discussed in Chapter 3), we first tried to identify and categorize problems which typically occur in opinion summarization through an error analysis of the state of the art blog summarizers. In this error analysis, we also compared blog summaries with news summaries to assess whether there is any information processing difference needed for blogs. For this error analysis, we used summaries from participating systems of the TAC 2008 opinion summarization track and the TAC 2008 update summarization track. Our study (detailed in Chapter 3) shows that question irrelevance, topic irrelevance, and discourse incoherence are the most frequently occurring problems

for blog summarization and these problems occur more frequently in blog summaries compared to news summaries.

Evaluation results of various studies (e.g. [CD08, GLYM09]) as well as ours [MK09] show that question irrelevance and discourse incoherence are two major areas where automatic summaries need to be improved. Our next goal was, therefore, to develop an effective blog summarization approach that addresses these most frequently occurring problems. The heart of our approach is based on discourse relations and text schemata.

According to [Tab06], "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 "*Where some operations in iPhoto take a few clicks in unexpected places, in Picasa they are almost always conveniently close to where you are currently working.*" a *contrast* relation is expressed. Discourse relations have been found useful in natural language generation [McK85] and in news summarization (e.g. [BGM06, Bos04]) to improve coherence and better simulate human writing. However, to the best of our knowledge, they have never been used for blog summarization.

Text schemata (described in Chapter 4) are patterns of discourse organization used to achieve different communicative goals. Text schemata were first introduced by McKeown [McK85] based on the observation that specific types of schemata are more effective to achieve a particular communicative goal. Schema-based approaches were also used by other researchers (e.g. [Par85, CN94]) in the context of question answering and text generation to generate relevant and coherent text. However, schema-based approaches are usually domain-dependent where the domain knowledge is pre-compiled and explicitly represented in knowledge bases or is used for structured documents (e.g. Wikipedia articles). In our research, we have tried to investigate:

1. How discourse relations and text schemata can be utilized to reduce question irrelevance and discourse incoherence. Specifically we tried to find a suitable schema-based model to make use of discourse relations for blog summarization.

2. How different types of discourse relations can be identified automatically for any given domain.

In this thesis, we propose a domain-independent query-based blog summarization approach using intra-sentential discourse relations within the framework of schemata.

In our approach, candidate sentences are first ranked using the topic and question similarity to give priority to topic and question relevant sentences. Since we are working with blogs, which are opinionated in nature (see Section 2.3), to rank a sentence we have also considered its semantic orientation or polarity (e.g. positive, negative or neutral) calculated using a subjectivity score (described in Section 6.1.2). The subjectivity score of a sentence is also used to calculate its relevance to the question. In the second step, questions are categorized based on their communicative goals to answer different types of questions differently and schema are designed for each question type. In the next step, sentences are categorized based on the discourse relations that they convey; we called this step "predicate identification". This step is critical because the automatic identification of discourse relations renders our approach independent of the domain. This step also plays a key role in the reduction of question irrelevance and discourse incoherence as schemata are designed using these relations. For predicate identification, first we compiled a list of rhetorical predicates (the basic unit of schema that characterize the predicating acts a writer may use and describe discourse relations between clauses) which we wanted to utilize (described in Chapter 5). Then we used four approaches to identify these predicates: a) the Rhetorical Structure Theory (RST) [MT88] based discourse parser SPADE [SM03], which can automatically identify discourse relations within a sentence; b) a comparison relations

classifier adapted from [JL06]; c) our topic-opinion discourse relation tagger based on the dependency relations of words defined by the dependency grammar [dMM08], and d) our own attributive tagger (described in Section 5.3.2). In the final step, a 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. As multiple sentences can be candidates to fill the same position in a schema, we devised further selection heuristics based on a corpus analysis.

To validate the approach described above, we have implemented a prototype system named BlogSum and evaluated its performance for question relevance and coherence using a subset of the BLOG06 corpus[3] (described in Section 2.3.2). The results show the effectiveness of our approach in reducing question irrelevance sentences by about 18% using ROUGE scores and in significantly improving question relevance and summary coherence with a p-value of 0.00281 and 0.0223 in a t-test using a manual likert scale of 1 to 5 compared to the original candidate list. We have also conducted another manual experiment to evaluate BlogSum-generated summary content and coherence using the OpinRank dataset[4] and [JL06]'s dataset of reviews. 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 summary content and coherence, respectively.

We have conducted another automatic experiment to evaluate BlogSum-generated summary content using the ROUGE metric with the DUC 2007 dataset on news articles. Evaluation results show that even though BlogSum was designed for opinionated texts, it performed quite satisfactorily with news articles; very close to the average performance of the DUC 2007 participants.

[3]BLOG06: http://ir.dcs.gla.ac.uk/test_collections/blog06info.html
[4]OpinRank Dataset: http://kavita-ganesan.com/entity-ranking-data

1.4 Intended Contributions

In this thesis, we show that discourse relations can be successfully used in a schema-based framework to reduce question irrelevance and discourse incoherence in blog summarization. The theoretical and practical development proposed in this thesis contributes to research in Natural Language Processing in the following ways:

1.4.1 Theoretical Contributions

Analysis of Summary-Specific Errors

A systematic manual analysis and comparison of the current state of the art blog summaries and news summaries has been performed with the goal of identifying frequently occurring errors in blog summaries and quantifying the information processing difference between the two genres. This was published in [MK09] and described in Chapter 3.

Analysis of Performance Issues of Automated Summaries

A study of current extractive summaries based on a literature survey and a summary analysis was conducted to reveal the main performance issues of query-based extractive summaries. This study also helped to identify why current approaches are not capable to address these issues (see Section 2.2).

Development of a Schema-based Summarization Approach

The development of a schema-based approach to use discourse relations for query-based blog summarization was performed and improved the current state of the art. To do so, we designed schemata and question categorization patterns. This was published in [MK10, MK11b] and described in Chapter 4.

Analysis of Current Predicate Tagging Approaches

We methodically analyzed and compared currently available discourse relations tagging approaches to evaluate the current state of the art. This led to the publication [MK11a] (see Section 5.3 and Section 7.4).

Identification and Development of a Predicate Tagging Approach

We have introduced a predicate tagging approach for the attributive predicates. See Section 5.3.2.

1.4.2 Practical Contributions

Design of a Prototype Summarizer

We have designed a prototype to show how our proposed approach can be used in a summarizer (called BlogSum). This is described in Chapter 6.

Evaluation of Performance

The evaluation of BlogSum with standard benchmarks empirically supports our theoretical developments (see Sections 7.2 and 7.3).

Identification of Summary Evaluation Issue

We empirically pointed out the need for a better automated summary evaluation metric rather than the standard ROUGE metric[5] [Lin04] (accepted in [MKP12]). See Section 7.2.

We believe that our proposed approach can also be used in other applications such as natural language generation and question answering in order to produce

[5]http://berouge.com/default.aspx

more coherent texts. Our work also provides a guidance to continue future research in summarization and other related domains.

1.5 Thesis Organization

The thesis is organized as follows: the current state of the art summarization approaches are reviewed in Chapter 2. This chapter also clarifies the terminology used in summarization research and provides a description of summary evaluation metrics. Chapter 3 describes specific considerations when dealing with blog summaries. This chapter specifically discusses our methodological study of blog summaries to identify blog specific errors and our attempt to quantify the information processing difference between blog and news summaries. Chapter 4 describes the heart of our approach: we show how discourse relations can be utilized in a schema-based framework in a domain-independent way. This chapter also demonstrates how our schema-based approach is able to reduce question irrelevance and discourse incoherence of blog summaries. Current discourse relations identifications approaches are described in Chapter 5. This chapter also gives a description of our attributive discourse relation tagger. Chapter 6 provides a complete description of how our schema-based summarization approach has been implemented. This chapter also describes our summarizer named BlogSum which we developed to validate our approach. Chapter 7 contains evaluation results of summary content and coherence. This chapter also presents evaluation results of the predicate identification approaches and the effects of the various heuristics we defined for sentence ordering. Finally, Chapter 8 presents the conclusion and a summary of possible future work.

Chapter 2

Background

In the last decade, research in the area of automated text summarization has gained intensive attention by researchers within the Natural Language Processing (NLP) community. This is in part due to the availability of more stable basic NLP tools (such as taggers, parsers, named entity taggers ...) from which second generation tools can be built. In addition, because huge amounts of online information is available today on the Internet, the need for automatic text analysis systems has reached a critical point. Hence, text summarization has become an important and timely tool for assisting and interpreting text information in today's fast-growing information age.

This chapter provides background information on text summarization which will be helpful to better appreciate the rest of the thesis. This chapter also discusses related work on top of which we built our own work and how our work differs from others.

2.1 Automated Text Summarization

[RHM02] define a summary as "*a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that*". This definition shows three main aspects of summarization: summaries can be produced from single or multiple documents; summaries need to contain important information from the source documents; and summaries need to be short. Automatic text summarization was first attempted in the 1950s, in the form of Luhn's [Luh58] auto-extracts; but for a long time after, little progress has been made in the field. However, since the 1990s, the increasing amount of online text and advances in Natural Language Processing (NLP) technologies have made summarization an active research area and today, it is one of the most active research area in NLP. Since the 1990s, various summarization approaches have been developed with the goal of producing summaries which contain topics from the original documents while keeping the redundancy to a minimum and presenting the information in a shorter and coherent form. In this section, some of the dimensions used to classify current summarization approaches are presented.

Summarization approaches are typically characterized by the following dimensions:

Indicative Summarization vs. Informative Summarization

An automatic summary is said to be *indicative* if it provides pointers to some parts of the original documents. Indicative summarization produces short summaries (two to three lines) that suggest which contents of the original documents are closely related to a user query. This type of summary facilitates a quick scanning of the original documents. On the other hand, an *informative* summary covers all relevant information

from the original documents [GL10]. An informative summary is meant to represent the original document. It provides a brief description of the original document, providing an idea of what the whole content of the document is about.

Extractive Summarization vs. Abstractive Summarization

Text summarization methods can be classified as *extractive* or *abstractive*. An extractive method selects and directly inserts important textual elements from the original texts into the summaries. Textual elements can be phrases, sentences or entire paragraphs [AKS05]. The importance of these textual elements can be decided based on statistical features (e.g. term frequency, sentence position), linguistic features (e.g. parts of speech, noun phrases) or both. Most of the current summarization approaches (e.g. [Bos04, BGM06, MJCN08]) employ an extractive approach and most of these work have been carried out on news articles where the main concern is what the summary *content* should be. Hence, extractive summarization approaches often suffer from linguistic problems such as incoherence.

On the other hand, abstractive summarization approaches (e.g. [RM98, XEN08]) first identify the most salient concepts from the original documents; then combine, reformulate, and appropriately present these concepts usually through Natural Language Generation (NLG) techniques [AKS05]. Abstractive approaches try to present important concepts in a new way by using NLG techniques such as *fusion* and *compression* (e.g. [RHM02]), where fusion combines extracted information coherently and compression removes unimportant information from texts.

Purely extractive summarization approaches often perform better than abstractive summarization approaches in shared evaluation tasks (e.g. TAC 2008, see Section 2.3.2) because abstractive approaches often require inference and natural language

generation which are relatively harder to perform compared to a data-driven approach such as sentence extraction [Rad04].

Single-Document Summarization vs. Multi-Document Summarization

Summarization approaches which produce a summary from a *single* document or *multiple* documents are known as single-document summarization (e.g. [Mar97a]) and multi-document summarization (e.g. [BGM06, MJCN08]), respectively. Research on single-document summarization started in the 1950s while multi-document summarization has gained interest in the mid 1990s. Most of the single-document and multi-document summarization applications have been developed in the domain of news articles. Multi-document summarization is more challenging compared to single-document summarization because of redundancy, temporal dimension, compression ratio, and incoherence problems [GMCK00].

Generic Summarization vs. User-oriented Summarization

Generic summarization approaches (e.g. [Mar97a]) try to incorporate as much information as possible from the original documents by maintaining the topical organization of the original documents (described in [Rad04]). On the other hand, *user-oriented* (or query-based) summarization approaches (e.g. [BGM06, MJCN08]) add contents based on users' preference or information needs that are expressed in terms of a query.

In our research, we developed a *query-based multi-document extractive* summarization approach to produce *informative* summaries. Table 1 situates our work within the dimensions presented above. The rest of the chapter will therefore focus on these types of summarization approaches.

Table 1: Our Work Situated Within the Summarization Dimensions

Dimension1		**Dimension2**	
Indicative	Informative	Extractive	Abstractive
	✓	✓	
Dimension3		**Dimension4**	
Single-Document	Multi-Document	Generic	User-Oriented
	✓		✓

2.2 Query-based Multi-Document Extractive Summarization Approaches

A query-based multi-document summarization approach includes contents based on users' preference or information needs that are expressed in terms of a query (or question). Figure 3 shows a query-based multi-document summary.

Figure 3: Sample Query-based Summary

Topic: Jiffy Lube

Question: What reasons are given for liking the services provided by Jiffy Lube?

Summary

Fortunately, I was right in front of a Jiffy Lube. I know it's fine cause Jiffy Lube sent me a little card in the mail and I have about a month before I need an oil change. When we first got here my bf took his car to jiffy lube cuz we didn't have a jack and its too low to crawl under. When I worked for my last auto shop, our best oil change was 49.99 to the same exact one for jiffy lube was damn near 100 bucks. Well, I suppose it is a little bit of a PITA to figure out what to do with the spent oil, but after some digging, I found out that every Jiffy Lube will take used oil for free! My bet is that Dallas will probably just go back to Jiffy Lube, since I'm not capable of a simple oil change. I had the biotches from Jiffy Lube strip out my oil drain plug! Mom got stuck in the snow, and the jiffy lube people pushed her out.

Most of the summarization approaches have been developed for generic summarization; query-based (or user-oriented) summarization is a relatively younger research area inspired from information retrieval. However, query-based approaches share many common characteristics with generic summarization approaches as both need

to incorporate important information into the summary and output summaries need to be coherent. As a result, many of the query-based summarization approaches have been developed using concepts developed for generic summarization or by modifying the generic approach by adding the user-given query as a dimension.

Query-based multi-document extractive summarization approaches rank input document sentences according to their importance. The importance of a sentence is calculated based on its relevance or similarity with the user given query or informativeness with respect to the whole input document set. Then top ranked sentences are selected up to a length limit to produce the final summaries. Current approaches often use post-processing to improve the readability of the summaries.

Current query-based extractive summarization approaches (e.g. [FR08, GLNW09, NFP08]) mainly perform the following tasks:

1. Pre-processing
2. Relevance Calculation
3. Sentence Selection
4. Post-processing

Common approaches which are used to perform these tasks are discussed below.

2.2.1 Pre-processing

In query-based summarization, current approaches often perform query expansion as part of the pre-processing where related words or phrases are added to the query in order to increase the possibility of finding matching sentences in the document set. To expand the query, WordNet[1], Wikipedia, word co-occurrences collected from a corpus are the most commonly used techniques [DO08].

[1]WordNet: http://wordnet.princeton.edu/

2.2.2 Relevance Calculation

Most approaches determine the relevance of a sentence based on its similarity or relevance with the user given query. This can be calculated by linguistic features, purely statistical features or a combination. Since in human writing, sentences may be included in a summary even if they are not relevant to the query in order to improve its informativeness [SB01], query independent features such as position in the text, overall frequency of the words they contain, or key phrases indicating the importance of the sentences are also used to calculate the relevance in the process of assigning scores to a sentence. Commonly used measures to calculate the relevance are listed below:

Linguistic Approaches

- **Linguistic Features** To improve the search for relevant sentences, some approaches (e.g. [GLNW09, CHJ08]) perform lemmatization, part-of-speech (POS) tagging, named entity (NE) recognition on both the query and the documents. These features are used to calculate the overlap between the query and the sentence and also to find the central concepts in the document sets in order to consider the informativeness of the sentences.

- **Deeper Linguistic Processing** Some approaches employ deeper linguistic processing. [BBW08], for example, performs similarity measurement using clustered noun phrases. [MBB98] compares clauses instead of sentences. [NFP08] uses a graph representation of named entities (NE) of the document sets which are connected by dependency relations.

Statistical Approaches

- **Cosine Similarity** To calculate the similarity between the query and a sentence, a commonly used measure is the *cosine similarity* (e.g. [Bos09, MJCN08]). Here both sentences and queries are represented as a weighted word vector often based on *tf.idf* (for sentences) and *tf* (for queries). The *idf* of a sentence is the inverse document frequency, which is defined by:

  $idf = \log \frac{|N|}{|n_i|}$

  where N is the total number of the documents in a collection, and n_i is the number of documents in which word i occurs. Term frequency *tf* is simply the frequency of a term in that sentence. The cosine similarity overlap of a sentence with a query is measured by computing the angle between the sentence vector and the query vector (similar sentences will have a small angle value) as follows:

  $\theta = \cos^{-1} \frac{\overrightarrow{q} . \overrightarrow{s}}{||\overrightarrow{q}||X||\overrightarrow{s}||}$

  where, $\overrightarrow{q}$ and $\overrightarrow{s}$ represent the query and the sentence vector, respectively.

- **Heuristic Features** Sentence position in the document and sentence length are also two widely used features to calculate relevance [DO08]. Sentence position is based on the assumption that early sentences in a document are more likely to contain focused, and important information. On the other hand, sentence length is based on the hypothesis that very short and very long sentences are unlikely to be useful.

Semantic Approaches

- **Semantic Resources** To determine each sentence's relevance to the query, many systems (e.g. [CKK+08]) calculate the degree of overlap between the query and the sentence using external semantic resources such as WordNet.

2.2.3 Sentence Selection

Sentences in the documents are ranked based on their relevance to the query or informativeness; then the top ranked sentences are selected to produce the summary. The main approach used to select sentences is ranking. To rank sentences, a combination of various features, language models, and graph-based approaches are commonly used. To select sentences, various clustering algorithms or machine learning approaches are also used.

Combination of Features

In the process of ranking, a score is assigned to each sentence to indicate its priority to be included in the final summary. The final score of a sentence is often calculated using a weighted combination of individual features such as query relevance, sentence position, sentence length values... The weight of each feature is usually calculated experimentally. In some work, only query relevant features are used; in others, both query relevant and query independent types of features are used.

Language Models

In some work (e.g. [Jag06, YYL+07]), language models are used to rank sentences. These approaches are based on the assumption that a document or a sentence is relevant to an information need, if the query can be treated as a representative sample of the document or sentence. This idea is akin to using language models for information retrieval. If a query is a better representative sample of a document d than $\acute{d}$, then d is assumed to be more relevant to the query. In these work, n-gram language models are used to predict the probability of natural word sequences; in other words, to assign a high probability to word sequences that occur frequently (and a low probability to word sequences that rarely or never occur).

Graph-based Approaches

Ranking can also be the result of a ranking algorithm applied to a sentence graph. In graph-based approaches, a graph is created using the input document. Sentences are vertices of the graph and edges are relations between sentences where, the relations are generated following different heuristic rules. In most graph-based approaches (e.g. [MR06, Bos04]), a centric graph is produced from all source documents and guides the summarizer to search for candidate sentences to be added to the output summary. A centric graph is a graph which has the highest number of bushy nodes; i.e. a node which is connected to many other nodes. The idea is that bushy nodes are the most important nodes in the graph because they are highly connected by/to other nodes [MR06]. This indicates that they contain the core concepts/entities about which the document is focusing. This concept, introduced in information retrieval is akin to the PageRank algorithm [PBMW99].

[Bos04] used a graph-based approach. In this work, the highest ranked sentence is calculated based on the query relevance and is added to the graph as the starting point. Later, a centric graph for the document is created by adding sentences based on their relation with the highest ranked sentence. RST (Rhetorical Structure Theory) [MT88] is used to create the graph representation of the document. In [Bos04]'s work, vertices of the graph are document sentences and edges are discourse relations between sentences where the relation strength is used as the weight of the edge. The approach works in two steps. First, the relations between sentences are defined in a discourse graph. Then, a graph search algorithm is used to extract the most relevant sentences from the graph for the summary. The sentences with the minimum path from the entry point (the highest ranked sentence) are selected.

Later, [Bos09] used another graph-based approach where sentences relevant to the query are added first, then sentences which are relevant to the already added

candidate sentences are added. This work adds contextual sentences in addition to the query relevant sentences to improve contextual description. In this work, first a query relevant graph is created where the edges of the graph show the cosine score using *tf.idf* between the query and the candidate sentences. Then a centric graph is created (called the contextual graph) where vertices are sentences from the same documents and edges show the cosine score between two sentences instead of a sentence and the query. In this graph representation, to calculate the relation strength between two vertices, the cosine similarity score is used. This means that if sentences share common words, then there will be edges between them. Finally, the relevance of each sentence is calculated using both graphs and the highest ranked sentences are included with the goal of improving the context.

Clustering

To select mostly query independent sentences which are very informative to represent the topic of the documents, clustering is also used (e.g. [DO08]). In this approach, sentences are clustered according to their similarity and then central sentences from each cluster are chosen for the summary. In the process of clustering sentences, similarity is mostly calculated based on *tf.idf* or Latent Semantic Analysis. Later, to choose central sentences from a cluster, two factors are considered: a) how relevant the sentence is to the general topic of the entire cluster using *tf.idf* and b) redundancy among sentences within a cluster.

Machine Learning Approaches

In query-based summarization, machine learning-based approaches are also used to improve the output by combining various features. [CSS05] designed a summarizer based on a Hidden Markov Model (HMM) where linguistic features, patterns with

lexical cues for sentence and phrase elimination and query terms are used as features. To reduce the search space for candidate sentences, [CSS05] remove sentences or phrases from the candidate sentence list using heuristic patterns based on lexical information such as gerund clauses, lead adverbs, etc. However, they found that full sentence elimination was not very useful. In this work, to identify query terms, part-of-speech (POS) tagging is done and the POS information is utilized as features. [CSS05] also used named entities to give importance to proper nouns, location, etc. At the end, the HMM model utilizes all these features to score the individual sentences classifying them as summary and non-summary sentences. From the evaluation, they identified that query terms was a very effective feature to find the best sentences for the summary. On the other hand, named entities were not a very useful feature due to the named entity identifier's mistakes and its coarse-grained classification. For example, LOCATION included cities, states/provinces/etc., countries, geographic features, etc. In their DUC 2005 participation, their method scored within the top group of systems for both ROUGE and pyramid evaluation (see Section 2.5 for a description of these measures).

[SK08] developed a machine learning approach to rank query relevant sentences. In this approach, a Support Vector Machine (SVM) classifier was trained using various features based mainly on word frequencies of words in the clusters, documents and topics. In their training set (DUC 2005), a cluster contained 25 documents and was associated with a particular topic. The topic contained a topic title and the topic descriptions. The topic title was a list of key words or phrases describing the topic. The topic description contained the actual query or queries (e.g., *Describe steps taken and worldwide reaction prior to introduction of the Euro on January 1, 1999.*). They used 8 features including: topic title frequency, topic description frequency, document frequency, sentence length, sentence position and so on. For evaluation, they used

the DUC 2006 and 2007 datasets and their classifier's scores corresponded to rank 6 for DUC 2007 (out of 32 systems) and rank 2 for DUC 2006 (out of 35 systems). From the evaluation, they found that the document frequency feature was the most important feature for sentence ranking.

MEAD [RABG+04], a publicly available and a widely used summarizer provides support for trainable summarization using decision trees, Support Vector Machines or Maximum Entropy. MEAD uses different features such as centroid, sentence length, query overlap, and so on. In the MEAD system, users can also define summary compression rates, for example, 10% of the original document sets. The MEAD summarizer consists of three components: a feature extractor, a sentence scorer, and a sentence reranker. The feature extractor first computes the weight of user-defined features such as position, centroid or length of each sentence. Once the feature extractor has assigned a weight for each feature then the sentence scorer computes the score of each sentence based on a linear combination of its features. At the end, the reranker arranges the summary sentences based on their scores beginning with the highest ranked sentence.

2.2.4 Post-processing

Most of the sentence selection approaches described above generate summaries that may be incoherent, redundant, and exhibit other linguistic problems. Most approaches try to reduce redundancy as part of their post-processing. Redundancy removal is typically done using the cosine similarity score (e.g. [MJCN08]) or Maximal Marginal Relevance (MMR) (e.g. [LOHW08]). In these approaches, first the highest ranked sentence is included in the final summary. Then the next candidate sentence is compared with this sentence for similarity. If the similarity score of the candidate sentence is above a threshold value then that sentence will not be added

to the summary. Otherwise, it is added, and the next candidates are then compared with each sentence already in the summary. This process continues for all candidate sentences until the summary reaches its length limit. Redundancy reduction can also be done using clustering (e.g. [VPK+08]). In this approach, sentences are clustered based on their similarity and from each cluster, then the top n ranked sentences are selected to produce the final summary. From a cluster, first the highest ranked sentence is added to the summary; then candidate sentences from the cluster are compared for similarity. If the similarity of the candidate sentence with the already added sentences is above a given threshold, the sentence is not added to the summary. In this approach, sentences are selected from different clusters to achieve high information coverage.

Table 2: Linguistic Quality Scores of Automatic Summarizers at TAC 2008

Non-redundancy	**Structure and Coherence**	**Fluency**
2.98	1.39	1.87

Even though most of the approaches try to address redundancy, current approaches pay little attention to improve other linguistic qualities such as coherence in their post-processing phase. Summary evaluation results also reflect this. Indeed, Table 2 shows the average scores for linguistic quality of the TAC 2008 opinion summarization track [Dan08]. In this evaluation, 3 linguistic criteria (non-redundancy, structure and coherence, and fluency) were evaluated manually by human assessors from the National Institute of Standards and Technology (NIST) on a scale of 1 to 5 where 1 meant "very poor" and 5 meant "very good". Table 2 shows that participant systems perform better in redundancy reduction compared to other linguistic qualities, but that much work still needs to be done to improve the linguistic quality of automatic summarizers.

2.2.5 Discussion

As discussed in the previous section, most of the query-based summarization approaches have been developed for news articles, and query and sentence overlap is computed by considering them as a bag-of-words without analyzing sentences semantically or at the discourse level. As a result, a sentence having a high similarity score with the query is very likely to be added in the final summary even if it is query irrelevant. As a result, current approaches often suffer from query irrelevance.

Summarization is seen as an extraction task in most of the current query-based summarization approaches. The main concern of current approaches is to improve summary content but very little attention is paid to improving summary organization. To do sentence organization, most of these approaches mainly use sentence scores to rank sentences in descending order which cannot ensure coherence. In some work, the original document sentence order (e.g. [KC08]) or the temporal order of the events (e.g. [BEM02]) are used to organize summaries. But approaches which use the sentence order of the original document are not useful for unstructured texts such as blogs and approaches which use temporal order are also not useful if the test domain is not event based. Current approaches try to improve redundancy rather than coherence; and the resulting summary often suffers from incoherence.

In statistical-based approaches and machine learning-based approaches various features such as sentence location are used to find query relevant sentences. However, these features are not useful for unstructured genres like blogs because these do not have predictable discourse structures [BG08]. Moreover, machine learning-based approaches require annotated data for training and are not as effective when applied to other domains. In graph-based summarization, the interrelatedness between sentences are considered which helps to improve context. But again in most of the graph-based approaches, sentences are viewed as bag-of-words. To improve query relevance and

coherence, sentences need to be interpreted at a deeper level of semantics or at the discourse level. There are some graph-based approaches (e.g. [Bos04, BGM06]) which use discourse relations. However, they are either domain-dependent or use only few discourse relations (see Section 2.4.1).

As we will see in Section 6.1.2, our summarization approach uses the standard technique of cosine similarity to calculate query relevance and uses a combination of features using attributes such as topic relevance, query relevance, and subjectivity scores to select candidate sentences. Our approach also performs post-processing to reduce redundancy 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 of output summaries (see Chapter 4).

2.3 Work on Opinionated Texts

Summarization from opinionated texts, or opinion summarization, is a fairly recent field. Query-based opinion summarization uses opinionated documents such as blogs, reviews, newspaper editorials or letters to the editor to answer opinionated questions as opposed to summarization of traditional news texts which uses fact-based information such as formal and event-based texts. Research on opinion summarization has been applied to diverse genres of texts such as customer reviews (e.g. [HL04]), conversations (e.g. [WL11]), and blogs (e.g. [MJCN08]).

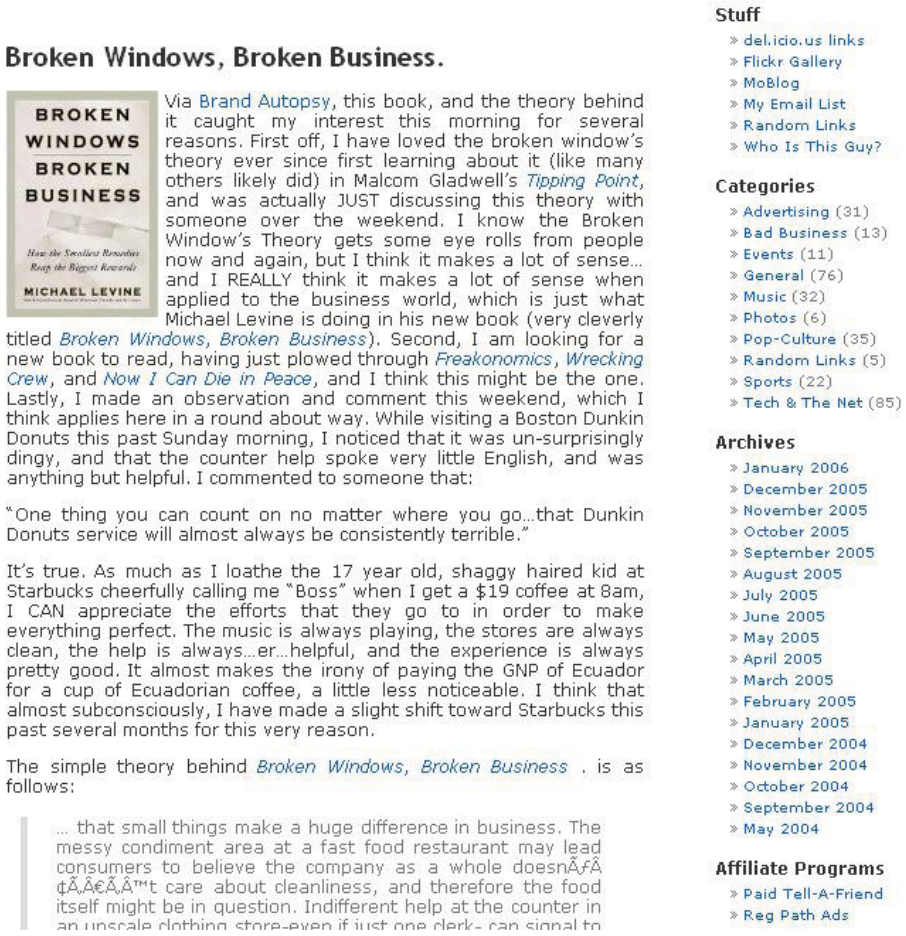
2.3.1 General Work on Blogs

Many blogs (or weblogs) are online diaries that appear in chronological order. Blogs reflect personal thinking and feelings on all kinds of topics including day to day activities of bloggers. Hence an essential feature of blogs is their subjectivity. Some blogs focus on a specific topic while others cover several topics; some describe personal daily lives of bloggers while others describe common artifacts or news. Many different sub-genres of blogs exist. The two most common are personal journals and notebooks [ABU07]. Personal journals discuss internal experiences and personal lives of bloggers and tend to be short. They are usually informal in nature and written in casual and informal language. They may contain much and sometimes only unrelated information such as ads, photos, and other non-textual elements. Personal journals also contain spelling and grammatical errors, and punctuation and capitalization are often missing. Figure 4 shows an example of a personal journal. On the other hand, notebooks contain comments on internal or external events. Similarly to newspaper articles, they are usually long and written in a more formal style [ABU07]. Most NLP work on blogs has tended to study personal journals as opposed to notebooks. For example, the BLOG06 corpus [MOS07], used at Text REtrieval Conference (TREC)[2] and at the Text Analysis Conference (TAC), contains mostly personal journals.

Blog is a useful media to understand peoples' responses to events, gather opinions on products and services. Even though natural language processing on blogs is a fairly new trend, its popularity is growing rapidly. Many conferences and workshops (e.g. Document Understanding Conference (DUC), Text Analysis Conference (TAC), Text REtrieval Conference (TREC)) are taking place to address different aspects of the analysis of blog entries. Over time, various studies have been conducted on blogs including subjectivity and sentiment analysis of blogs, blog post tagging, spam blog

[2]Text REtrieval Conference: http://trec.nist.gov

Figure 4: Sample Blog Post from the BLOG06 Corpus



Broken Windows, Broken Business.

Via Brand Autopsy, this book, and the theory behind it caught my interest this morning for several reasons. First off, I have loved the broken window's theory ever since first learning about it (like many others likely did) in Malcom Gladwell's *Tipping Point*, and was actually JUST discussing this theory with someone over the weekend. I know the Broken Window's Theory gets some eye rolls from people now and again, but I think it makes a lot of sense... and I REALLY think it makes a lot of sense when applied to the business world, which is just what Michael Levine is doing in his new book (very cleverly titled *Broken Windows, Broken Business*). Second, I am looking for a new book to read, having just plowed through *Freakonomics*, *Wrecking Crew*, and *Now I Can Die in Peace*, and I think this might be the one. Lastly, I made an observation and comment this weekend, which I think applies here in a round about way. While visiting a Boston Dunkin Donuts this past Sunday morning, I noticed that it was un-surprisingly dingy, and that the counter help spoke very little English, and was anything but helpful. I commented to someone that:

"One thing you can count on no matter where you go...that Dunkin Donuts service will almost always be consistently terrible."

It's true. As much as I loathe the 17 year old, shaggy haired kid at Starbucks cheerfully calling me "Boss" when I get a $19 coffee at 8am, I CAN appreciate the efforts that they go to in order to make everything perfect. The music is always playing, the stores are always clean, the help is always...er...helpful, and the experience is always pretty good. It almost makes the irony of paying the GNP of Ecuador for a cup of Ecuadorian coffee, a little less noticeable. I think that almost subconsciously, I have made a slight shift toward Starbucks this past several months for this very reason.

The simple theory behind *Broken Windows, Broken Business* . is as follows:

> ... that small things make a huge difference in business. The messy condiment area at a fast food restaurant may lead consumers to believe the company as a whole doesnÃƒÂ¢Ã‚Ã‚Â™t care about cleanliness, and therefore the food itself might be in question. Indifferent help at the counter in an upscale clothing store-even if just one clerk- can signal to

Stuff
- del.icio.us links
- Flickr Gallery
- MoBlog
- My Email List
- Random Links
- Who Is This Guy?

Categories
- Advertising (31)
- Bad Business (13)
- Events (11)
- General (76)
- Music (32)
- Photos (6)
- Pop-Culture (35)
- Random Links (5)
- Sports (22)
- Tech & The Net (85)

Archives
- January 2006
- December 2005
- November 2005
- October 2005
- September 2005
- August 2005
- July 2005
- June 2005
- May 2005
- April 2005
- March 2005
- February 2005
- January 2005
- December 2004
- November 2004
- October 2004
- September 2004
- May 2004

Affiliate Programs
- Paid Tell-A-Friend
- Reg Path Ads

post detection, opinion mining where blogs are mined for useful information (e.g. popular culture trend), opinionated question answering, and blog summarization.

Subjectivity and sentiment analysis include classifying sentiments of reviews (e.g. [And09, PLV02]) and analyzing bloggers' mood and sentiment on various events (e.g. [MG06]). Sentiment classification of reviews on different events is often done on movie or product reviews. Rating indicators of reviews are used to identify the polarity of the blogs, namely positive, negative or neutral. To analyze bloggers' mood and sentiment, systems make use of information regarding bloggers' mood varying over time. To record bloggers' varying mood, the polarity information of the blog post is often used.

[LKS06] developed the Lydia system to analyze blogs. They analyzed the temporal relationship between blogs and news articles. In particular, they analyzed how often bloggers report a story before newspapers and how often bloggers react to news that have already been reported. Later on, [GSS07] developed a large-scale sentiment analysis system on top of the Lydia text analysis system for news articles and blog entities. They determined the public sentiment on various entities and identified how this sentiment varies with time. They found that the same entities (persons) except certain controversial political figures received comparable opinions (favorable or adverse) in blogs and news texts. Controversial political figures received different opinions in blogs compared to news articles because of the political biases among bloggers, and perhaps the mainstream press.

Question answering (QA) on blogs and on linked data are relatively new fields. Since today, huge amounts of texts are available due to the popularity of the social media, there is an urge for a system or an architecture that can make connections between related pieces of information. To fulfill these needs, workshops are taking place on linked data[3] and approaches are being developed to address these (e.g. [Dub11]). The most notable QA work on blogs was conducted at TREC 2007 [MOS07] and TAC 2008 (see Section 2.3.2). To answer queries on an event or entity, TREC provided a blog corpus in addition to the AQUAINT newspaper corpus. The TREC blog track [MOS07] provided an opportunity to build new techniques of sentiment tagging on blog posts. The task was to identify and rank blog posts on a given topic from a corpus of blogs. [SWWS07] developed an opinion question answering approach for blogs and news articles. They exploited attitude information, namely sentiment and argument types, to answer opinion questions. They achieved comparable result with both text types.

[3]http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/index.php?x=home&q=1

2.3.2 Work on Query-based Blog Summarization

The availability of huge amounts of social media documents (e.g. blogs) has recently drawn researchers' interest; hence query-based blog summarization is a relatively new but very active field.

Blog Summarization Approaches

Similar to news article-based approaches (see Section 2.2), blog-based approaches also use extractive summarization and rank sentences based on their importance, where the importance is calculated using query relevance and informativeness of the sentence using query independent features (e.g. [HCGL08, CS08, VPK+08]). To calculate the relevance, blog summarization approaches also commonly utilize word overlap between the query and sentences using the cosine similarity score based on *tf.idf* information (e.g. [MJCN08, BG08]). Similarly to news summarization, linguistic features such as lemmatization, POS tagging, named entity (NE) as well as heuristic features of sentence position [BG08], sentence length [BG08] are also used (e.g. [MJCN08]). However, in addition, blog-based summarization also uses the polarity information (e.g. positive, negative, neutral) and sentiment degree (subjectivity scores) to rank sentences.

A major difference with query-based blog summarization is the use of the polarity information and sentiment degree. Opinion dictionaries such as the General Inquirer[4], the MPQA Opinion Corpus[5] (used in [MJCN08, HCGL08]) and different machine learning techniques based on polarity-annotated corpora (e.g. [VPK+08, BG08]) are used to identify the polarity of the question and sentences and to assign a subjectivity score. Table 3 shows the polarity and sentiment degree of a few words from the MPQA lexicon (more details will be provided in Section 6.1.2).

[4]General Inquirer: http://www.wjh.harvard.edu/∼ inquirer
[5]MPQA: http://www.cs.pitt.edu/mpqa/mpqa_corpus.html

Table 3: Examples of Word Polarity and Sentiment Degree in the MPQA Lexicon

Word	**Sentiment Degree**	**Polarity**
Congratulation	Strong	Positive
Ability	Weak	Positive
Hate	Strong	Negative
Complication	Weak	Negative
Eat	Not Applicable	Neutral

To select sentences, sentence ranks are mostly used where the ranks are calculated based on a weighted combination of question relevance scores, query-independent scores, and subjectivity scores (e.g. [BG08, MJCN08]). At the end, the top ranked sentences are selected to produce summaries. Similarly to news summarization (see Section 2.2.4), blog summarization approaches (e.g. [HCGL08]) also perform redundancy removal to improve readability. To select sentences, some work also use language models (e.g. [HCGL08]).

[KLC06] developed a language independent opinion summarization approach. For summarization, they retrieved all sentences which are relevant to the main topic of the document set and determined the polarity and subjectivity scores of these relevant sentences. They also found that the identification of correlated events on a time interval is also important for opinion summarization. They tested their approach for blogs and news articles for English and Chinese languages. From their evaluation, they found that blogs contain more question irrelevant information compared to news articles. Their results confirm our own results (see Section 3.3). [KLC06] also found that news articles use a larger vocabulary compared to blogs which makes the filtering of non-relevant sentences harder for news articles. On the other hand, this larger vocabulary helps to determine the polarity. Due to the limited vocabulary the judgment of polarity of blogs was difficult.

Recently, [PZG10] developed a blog summarization approach to highlight the contrast between multiple viewpoints expressed towards a topic by developing a model to jointly represent topic and viewpoints in the text. Other researchers also attempted to add new dimensions in blog summarization such as usage of comment of blog posts [PB10] rather than addressing question irrelevance and discourse incoherence.

Text Analysis Conference 2008

The most notable resource in query-based blog summarization is TAC 2008. In 2008, the Text Analysis Conference (TAC) introduced a query-based opinion summarization track[6]. At this track, participants were given a set of target topics on various events or entities collected from the BLOG06 corpus[7]. BLOG06 is a TREC test collection, created and distributed by the University of Glasgow to support research on information retrieval and related technologies. BLOG06 consists of 100,649 blogs which were collected over an 11 week period (a total of 77 days) from late 2005 and early 2006. The total size of the collection is 25 gigabytes. In this corpus, blogs vary significantly in size, ranging from 44 words to 3000 words.

At the TAC 2008 opinion summarization track, for each topic, a set of questions and a set of relevant blog entries (mostly personal journals) were provided. For example, for the topic "*UN Commission on Human Rights*", two questions were asked:

1. "*What reasons are given as examples of their ineffectiveness?*"

2. "*What steps are being suggested to correct this problem?*"

and a set of IDs of related blogs in the BLOG06 corpus were provided. Participating systems needed to extract answers to questions from these specified sets of

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

[7]BLOG06 Corpus: http://ir.dcs.gla.ac.uk/test_collections/blog06info.html

blogs and summarize them. Optionally, additional input were provided in the form of answer-containing text snippets found by question answering systems 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. Here is one sample snippet for the topic *UN Commission on Human Rights*:

1. "*Issues regular resolutions condemning Israel while overlooking real offenders.*"

In the TAC 2008 opinion summarization track, 50 questions on 28 topics were distributed. For each question, from 9 to 39 relevant blogs, which are subset of the BLOG06 corpus, were distributed. In total, 600 blogs were provided. The National Institute of Standards and Technology (NIST) received 45 runs from 19 teams for the opinion summarization task. Each team submitted up to three runs, ranked by priority. Due to assessing constraints, NIST manually evaluated only runs with priority 1 and 2. At the end, they evaluated a total of 36 runs. NIST provided a standard evaluation forum based on the automatic metrics discussed in Section 2.5. In addition, NIST assessors manually evaluated summary content using pyramid scores and linguistic quality of the summary on a likert chart (discussed in Section 2.5).

To this day, TAC 2008 remains the reference on blog summarization because there has been no other main "bake-off" style conference on blog summarization.

Most of the summarization approaches designed at TAC 2008 (e.g. [KPVZ08, MJCN08]) 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. Some systems (e.g. [HB08]) 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 which were provided with

the TAC 2008 dataset.

Most of the approaches at TAC 2008 used sentence scores to order final summaries. The highest ranked system for summary coherence at TAC 2008 [CS08] 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 [CS08]. The second best ranked system at TAC 2008 for summary coherence [BG08] 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 appeared in the summary were positive sentences. However, none of the top ranking systems at TAC 2008 used discourse relations to address summary coherence.

The TAC 2008 conference also provided an update summarization track. In this track, participants needed to generate short (about 100 words) fluent multi-document summaries of news articles under the assumption that the user had already read a set of earlier articles. The purpose of each update summary was to inform the reader about new information about a particular topic. In this track, the test dataset was composed of 48 topics. On each topic, a topic statement and 20 relevant documents divided into 2 sets (Document Set A and Document Set B) were distributed. Each document set contained 10 documents, where all the documents in Set A chronologically preceded the documents in Set B. The documents were collected from the AQUAINT-2 collection of news articles[8]. For example, on the topic "*Airbus A380*", the topic statement was:

[8]AQUAINT-2 collection: http://www.nist.gov/tac/data/data_desc.html#AQUAINT-2
The AQUAINT-2 collection is a subset of the Linguistic Data Consortium (LDC) English Gigaword Third Edition. The AQUAINT-2 collection comprises approximately 2.5 GB of text (about 907K documents). These documents were collected over the time period of October 2004 to March 2006. Articles of the AQUAINT-2 collection are in English and come from a variety of sources including Agence France Presse, Central News Agency (Taiwan), Xinhua News Agency, Los Angeles Times; Washington Post News Service, New York Times, and the Associated Press.

1. Describe developments in the production and launch of the Airbus A380.

As will be described in Chapter 3, we have compared summaries generated by participants of the update summarization track and the opinion summarization track to find blog-specific errors in summaries.

Discussion

Since most of the blog summarization approaches consider sentences and queries as bag-of-words without applying deep natural language analysis, they often produced query irrelevant summaries. Most of these summarization approaches (e.g. [MJCN08, BG08]) use sentence scores for summary organization. Some of these approaches (e.g. [KC08]) use the sentence order of the original documents to specify the sentence order of the summary. Recent work (e.g. [PZG10]) on blog summarization also mostly uses sentence scores for summary generation. However, these approaches can hardly be effective in coherence improvement. To improve the state of the art, in our work, we have tried to go beyond the bag-of-words approach and have attempted to use discourse relations to address query irrelevance and discourse incoherence.

Since we utilized discourse relations in a schema-based framework to address question irrelevance and discourse incoherence, the next section will give an overview of discourse relation-based and schema-based approaches.

2.4 Discourse Relations and Schema-based Approaches

2.4.1 Discourse Relations

It is widely accepted that in a text, sentences and clauses should not be understood in isolation but in relation to each other through discourse relations that may or may not

be explicitly marked. A text is not a linear combination of clauses but a hierarchial organized group of clauses placed together based on informational and interactional relations to one another. 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; they become more meaningful when we realize that they are related through the discourse relation *condition*.

In a discourse, different kinds of relations such as *contrast*, *causality*, *elaboration* may be expressed. Discourse relations can occur within a sentence or across sentences. For example, "*Although they represent only 2% of the population, they control nearly one-third of the discretionary income.*" is an example of intra-sentential relation (concession relation). On the other hand, "*The projects are big. They can cost $1 billion plus.*" is an example of inter-sentence relation (elaboration relation).

To describe discourse relations, different theories have been developed such as Rhetoric [Ari54], Rhetorical Predicates [Gri75, Hob85], Discourse Representation Theory [Kam81, Ash93], Rhetorical Structure Theory [MT88] and other theories by [Gro85, GL86, KD94, Hov93, HM93]. Some theories are inclusive compared to others with respect to discourse structure definition and applicability. For example, Rhetorical Structure Theory (RST) [MT88] is comprehensive compared to its predecessors because it provides extensive definitions of various discourse relations and showed that a plan based approach can be used to apply these relations computationally [Mit93].

Rhetorical Structure Theory (RST) is one of the most widely used discourse theory for computational work (e.g. [SM03, CM01]). RST is a theory of text organization created in the 1980s as a result of exhaustive analysis of texts. Mann and Thompson [MT88] produced a list of approximately 25 discourse relations (e.g. *elaboration*,

contrast) and claimed that these relations are sufficient to describe the discourse structure of all coherent English texts. According to Mann and Thompson's RST, a relation typically holds between two non-overlapping text spans called a nucleus (the central segment of the relation) and a satellite (the supporting information). Sometimes two related text spans are equally important; in that case, a multinuclear

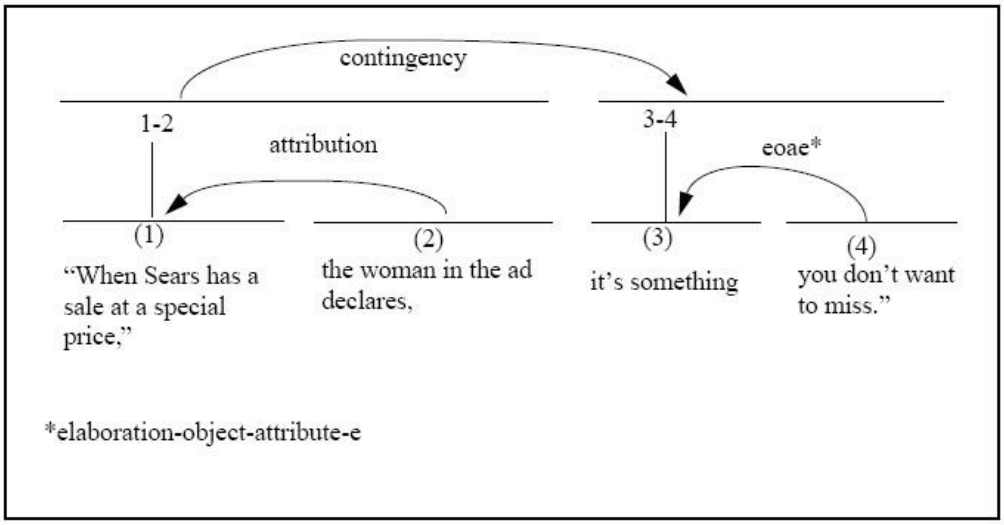
Figure 5: Definition of the *Evidence* Relation in RST (from [MT88])

Relation name	*EVIDENCE*
Constraints on N :	*The reader R might not believe the information that is conveyed by the nucleus N to a degree satisfactory to the writer W.*
Constraints on S :	*The reader believes the information that is conveyed by the satellite S or will find it credible.*
Constraints on N + S combination :	*R's comprehending S increases R's belief of N.*
The effect :	*R's belief of N is increased.*
Locus of the effect :	*N.*
Example :	*[The truth is that the pressure to smock in junior high is greater than it will be any other time of one's life :] [we know that* 3,000 *teens start smocking each day.]*

relation holds [Mar97b, Hov93]. RST shows how texts can be decomposed recursively into smaller segments (down to the clause level) where these segments are related to each other by discourse relations. Each relation is defined in terms of a distinctive set of constraints on the information presented on the segments, on the speaker/hearer belief state, and on the effect that the speaker wants to achieve through this relation. In this process, the constraints hold on the nucleus, satellite, and the combination of nucleus and satellite. Figure 5 shows the description of the *evidence* relation in RST. Here, an *evidence* relation holds between the nucleus N (*The truth is that ...*) and the satellite S (*we know that ...*). The writer believes that the information expressed in the nucleus N is insufficient to achieve the reader's acceptance. The information

expressed in the satellite S is assumed to be believed by the reader. The combination of nucleus and satellite is assumed to help increase the reader's belief in the nucleus. The effect of this relation is to increase the belief on the information that is presented in the nucleus.

Figure 6: Sample RST Tree (from the RST corpus)



The analysis of a discourse using RST builds a tree-like structure linking nuclei and satellites with the rhetorical relations. Figure 6 shows a sample RST tree for the sentence: *"When Sears has a sale at a special price," the woman at the ad declares, "it's something you don't want to miss."* In the tree, the arrows link the satellite to the nucleus of a discourse relation. Arrows are labeled with the name of the discourse relation that holds between the linked units. Horizontal lines correspond to text spans, and vertical lines identify text spans which are nuclei.

The use of discourse structures have been found useful in many applications such as document summarization and question answering (e.g. [McK85, Mar97a]). For example, [McK85] showed that discourse relations can be used to select the content and generate coherent text in question answering with the help of schemata (see Section 2.4.2). Discourse relations have also been found useful for anaphora resolution

(e.g. [McK85]) and machine translation (e.g. [Mit93]).

Discourse relations have been used for text summarization. Most notably, [Mar97a] used discourse relations for single document summarization and proposed a discourse relation identification parsing algorithm. [MBB98, ORL02] experimentally showed that discourse relations can improve the coherence of multi-document summaries. In some work (e.g. [Bos04, BGM06]), discourse relations are exploited successfully for multi-document summarization. In this work, only discourse relations across sentences are utilized. The great difficulty in using discourse relations computationally for text analysis is the lack of availability of systems to identify discourse relations across sentences automatically. Currently, manually annotated discourse relations corpora such as the Penn Discourse Treebank [PMD+08] and the RST Discourse Treebank [CM01] are available. These corpora facilitate the computational use of discourse relations. [Bos04] shows the effectiveness of discourse relations to incorporate additional contextual information for a given question in a query-based summarization. In this work, the evaluation was done on selected domains for which annotated discourse relations were available. [BGM06] used discourse relations for content selection and organization of automatic summaries and achieved improvement in both cases. They considered only two discourse relations for their work and specified criteria to identify these relations based on text analysis. They adopted an unsupervised approach to recognize discourse relations from [ME02] to identify these relations automatically. These discourse relations are used as content selection features and for content organization. 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*.

In our work, we have considered intra-sentential discourse relations only; because in extractive summarization, question answering, information retrieval and in many

other applications, individual sentences (candidate sentences) are extracted from different documents or from different positions of a document to build a candidate sentence list. As a result, it is 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 [McK85, BG07] (see Section 2.4.2). 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 [BG07] who showed that 95% of the time, causality occurred within sentences in the T corpus[9]. In addition, [SM03] notes that 95% of the sentences in the RST Discourse Treebank corpus[10] contain intra-sentential relations. This is why in our research, we have exploited intra-sentential discourse relations.

2.4.2 Schema-based Approaches

In previous work, schema (or template-based) approaches have been used successfully to achieve text coherence (e.g. [SB09, JKN10]). In [McK85], 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. In McKeown's schema-based approach, clauses are classified into a predefined set of rhetorical predicates which correspond to organizing relations which are used in discourse [Gri75, McK85] (see Chapter 5). After a corpus analysis, McKeown observed that some combinations of rhetorical predicates are more likely to occur than others and some combinations are more appropriate

[9] A gigaword newswire corpus of 4.7 million newswire documents.
[10] http://www.isi.edu/∼marcu/discourse/Corpora.html

for a particular communicative goal. For example, the definition of an object is often provided by a particular combination of predicates, whereas a comparison of two objects uses a different combination. These standard combinations of predicates to achieve a particular communicative goal (e.g. compare two objects) are defined by schemata. Schemata provide partially ordered flexible text structures.

To illustrate the text schema-based approach, the identification schema, designed by McKeown, which shows a strategy to provide definitions, is shown in Figure 7[11]. This schema is suitable to answer questions for a definition; for example, "*What is a Hobie Cat?*". It uses predicates such as *identification*, *analogy*, *constituency*, ... shown in the Identification schema (Figure 7). The Identification schema stipulates that a definition of an item should first provide information on its generic class (using the *identification* predicate), then use sentences that provide *constituency* or *attributes* (using the *constituency* or *attributive* predicates), followed by sentences that provide examples (by using the *particular-illustration* or *evidence* predicates), followed by optional sentences that contains analogies or examples.

Figure 7: Identification Schema (from [McK85])

1. $Identification\ (class\ \&\ attribute/function)$
2. $\{Analogy/Constituency/Attributive/Renaming/Amplification\}^*$
3. $Particular-illustration/Evidence^+$
4. $\{Amplification/Analogy/Attributive\}$
5. $\{Particular-illustration/Evidence\}$

Schemata may be nested within others; as a result, a portion of text can be omitted or repeated based on the information need. A focusing technique was used to fully order texts when a schema does not completely constraint the choices. This mechanism was developed based on prioritizing constraints on how focus of attention

[11]The symbol / indicates an alternative, { } indicates optionality, * indicates that the item may appear 0 to n times, + indicates that the item may appear 1 to n times.

can shift from one sentence to the next.

McKeown also demonstrated the usability of her schema-based approach for a domain-dependent question answering application. In this application, McKeown designed various schemata that incorporate discourse relations which are typically used in human writing for a specific question type (e.g. identification).

Text schemata were later used by other researchers (e.g. [Par85, Tat91, CN94]) where specific schemata were designed according to the specific applications and knowledge of the user. In more recent work, [SB09, JKN10] also tried to utilize discourse structures learned from domain relevant articles (e.g. scientific research paper) to design schemata for summarization where they applied schema-based approach for a well structured documents.

2.4.3 Discussion

Discourse relations have been used in diverse domains to generate coherent text as well as for text summarization. Most of these summarization approaches are developed for a single document and for a generic summary generation instead of a query-focused multi-document summary generation. Only a few query-based summarization approaches (e.g. [Bos04, BGM06]) have used discourse relations. Even though discourse relations across sentences are found useful for news summarization, available approaches are either domain-dependent or use only few discourse relations because of the unavailability of reliable automatic identification of across sentence relations. However, to the best of our knowledge, discourse relations have never been used for blog summarization. In our work, we used intra-sentential discourse relations that are genre and domain independent in a schema-based framework. We have also developed and analyzed automatic approaches to identify many intra-sentential discourse relations (see Section 5.3). Available schema-based approaches are typically

domain-dependent and the domain knowledge is explicitly represented in knowledge bases and later used to identify discourse structures. As opposed to targeting only a specific domain and tagging discourse relations in advance in a knowledge base, we have used a text schema-based approach applicable to any domain by identifying discourse relations automatically. In previous work, schema-based approaches have been used for well structured documents (e.g. Wikipedia pages); in contrast, in our work, we use schema for very unstructured documents: blogs.

2.5 Summary Evaluation

Nowadays, any NLP endeavor must be accompanied by a well-accepted evaluation scheme. Summary evaluation is a critical issue in text summarization research. During the last 15 years, to evaluate automated summarization systems, sets of evaluation data (corpora, topics, ...) and baselines have been established in text summarization competitions such as TREC[12], DUC[13], and TAC[14] (see Section 2.3.2). Although evaluation is essential to verify the quality of a summary or to compare different summarization approaches, the evaluation criteria used are by no means accepted unanimously.

The available evaluation techniques are divided into two categories: manual and automatic. To do a manual evaluation, human experts assess different qualities of the system generated summary. On the other hand, for an automatic evaluation, tools are used to compare the system generated summary with a human generated gold standard summary or reference summary. Although they are faster to perform and result in consistent evaluations, automatic evaluations can only address superficial concepts such as n-gram matching, because many required qualities such as coherence

[12]Text REtrieval Conference: http://trec.nist.gov
[13]Document Understanding Conference: http://duc.nist.gov
[14]Text Analysis Conference: http://www.nist.gov/tac

and grammaticality cannot be measured automatically. As a result, human judges are often called for to evaluate or cross check the quality of the summaries, but in many cases human judges have different opinions. Hence inter-annotator agreement is often computed as well.

The quality of a summary is assessed mostly on its content and linguistic quality [LN08]. Content evaluation of a query-based summary is performed based on the relevance with the topic and the question and the inclusion of important contents from the input documents. The linguistic quality of a summary is evaluated manually based on how it structures and presents the contents. Mainly, subjective evaluation is done to assess the linguistic quality of an automatically generated summary. Grammaticality, non-redundancy, referential clarity, focus, structure and coherence are the commonly used factors considered to evaluate the linguistic quality.

Available manual and automatic evaluation techniques can be further divided into two classes: intrinsic methods and extrinsic methods. An intrinsic evaluation compares the results of the system to a gold standard to evaluate the system in isolation. This method measures the quality of the summary such as integrity of sentences and readability. This approach gives a quantitative measure but it is hard to get a well agreed gold standard summary [AKS05, GKV06].

An extrinsic evaluation measures the quality of a system in term of its utility to solve a particular task [GKV06]. For example, in a query-based summary the end goal is to answer the user query. In this case, human judges evaluate the summary output in terms of how well it answers the users query and not necessarily how it compares with a gold standard.

There exist different measures to evaluate an output summary. The most commonly used metrics are *recall*, *precision*, *F-measure*, *Pyramid score*, and *ROUGE/BE*.

Recall, Precision, and F-measure

Recall, precision, and F-measure scores (described in [JM00]) are often used in both automatic and manual summary evaluation. These are standard measures in many NLP applications such as summarization (e.g. [YYL+07, MJCN08]), question answering (e.g. [CHJ08]), and information retrieval.

Recall is a measure of how much relevant information the system has extracted from the document. It is defined as:

$$Recall = \frac{\#\ of\ correct\ sentences\ extracted\ by\ the\ system}{total\ of\ possible\ correct\ sentences\ in\ the\ gold-standard\ summary}$$

Precision is a measure of how much of the information that the system has extracted is actually correct. It is defined as:

$$Precision = \frac{\#\ of\ correct\ sentences\ extracted\ by\ the\ system}{\#\ of\ sentences\ extracted\ by\ the\ system}$$

F-measure is a weighted harmonic mean of precision and recall; it is defined as:

$$F - measure = \frac{(\beta^2+1)PR}{\beta^2 P+R}$$

where, β is a parameter to balance recall and precision. When β is 1, precision and recall are considered equally important. When β is greater than 1, precision gets more weight, and when β is less than 1 recall gets more weight. P and R stand for precision and recall, respectively.

Pyramid Evaluation Method

The pyramid evaluation (described in [PNMS05]) is a manual evaluation metric that creates a map between the summary sentences and source documents by identifying summarization content unit (SCUs). SCUs are minimal unit of informative ability. Weights are assigned to SCUs based on the number of human judges who agreed on

the source of SCUs. These SCUs build different layers of a pyramid based on their score. SCUs in higher level of the pyramid are assumed to be the more salient information from the original text. An output summary is evaluated based on the number of SCUs present and summing up their weight [GKV06]. The pyramid metric was used in the DUC conference in 2006 and 2007 and currently, the TAC conference also employs this metric to evaluate participants' summary content.

ROUGE/BE Evaluation Method

The ROUGE metric has become a standard of automatic evaluation of summary content. This metric was used in DUC from 2004 to 2007 and is use in TAC for participants' summary evaluation. In the ROUGE/BE [Lin04], basic elements (BE) are matched between the source document and the output summary. BEs can be matched using simple string matching, n-gram overlap or more complex matching methods. The main idea of this approach is to identify minimal units of information from the original text to the summary by locating similar meaning. The BE approach actually defines a family of measures depending on which basic element is used. ROUGE is a specific instance of the BE approach where BEs are word unigrams or n-grams of a higher order. The ROUGE evaluation tool is often used to compute the ROUGE-2 and ROUGE-SU4 score. The ROUGE-2 score is based on the overlap of bi-grams (using words as tokens) between the automatically generated summaries and human generated gold standard summaries (or reference summaries) [DOCS07]. 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-bigram), and includes uni-gram co-occurrence statistics as well [DOCS07].

The above mentioned evaluation metrics are used to evaluate both opinionated and news article based summarization approaches. Shared evaluation tasks such as

DUC and TAC competitions also use these methods to evaluate participants' summary. Table 4 shows the evaluation results of automatic systems' average performance at the TAC 2008 to 2010 conferences using the pyramid score, linguistic quality (Ling. Q.), and responsiveness (Resp.). In this evaluation, the pyramid score was used to calculate the content relevance; linguistic quality was used to evaluate grammaticality, coherence, non-redundancy, readability, and so on; and the responsiveness of a summary was used to judge the overall quality or usefulness of the summary, consid-

Table 4: Human and Automatic System Performance at Various TAC Competitions

	Model (Human)			**Automatic**		
	Pyramid	**Ling. Q.**	**Resp.**	**Pyramid**	**Ling. Q.**	**Resp.**
2010 Update	0.785	4.908	4.761	0.302	2.837	2.565
2009 Update	0.683	8.915	8.830	0.260	4.859	4.149
2008 Update	0.663	4.786	4.619	0.260	2.346	2.323
2008 Opinion	0.446	Unknown	Unknown	0.102	2.13	1.31

ering both the information content and readability. All three criteria were evaluated manually. The pyramid score was calculated out of 1 and the other two measures were calculated on a scale of 1 to 5 (1, being the worst). However, in 2009, linguistic quality and responsiveness were calculated on a scale of 1 to 10. Table 4 also shows a comparison between automatic systems and human assessors (model). In Table 4, the first 3 rows show the evaluation results of the TAC Update Summarization initial summary generation task (which were generated for news articles) and the last row shows the evaluation results of the TAC 2008 Opinion (blog) Summarization track (see Section 2.3.2). From Table 4, we can see that in all three criteria, automatic systems are weaker than humans.

Table 5 shows the average performance scores of human and participant systems at the TAC 2008 Update Summarization track using ROUGE-2 and ROUGE-SU4.

Table 5: Human and Automatic System Performance in ROUGE at TAC 2008 Update Summarization

	ROUGE-2	**ROUGE-SU4**
Model (Human)	0.12	0.14
Automatic	0.08	0.12

Interestingly, in this evaluation, we can see that there was no significant performance difference between human and automatic systems; they achieved similar ROUGE scores. [DO08] explains this the following ways: "automatic metrics, based on string matching, are unable to appreciate a summary that uses different phrases than the reference text, even if such a summary is perfectly fine by human standards". On the other hand, the TAC 2008 update summarization task showed that there exists a significant gap between automatic summarizers and human summarizers based on manual evaluation of summaries [DO08]. This indicates that ROUGE may not the most effective tool to evaluate summaries. Indeed the same phenomenon will be encountered in our summary content evaluation (see Section 7.2).

According to [DM07], a universal strategy to evaluate summarization systems is still absent. Summary evaluation is a difficult task because no ideal summary is available for a set of documents. It is also difficult to compare different summaries and establish a baseline because of the absence of standard human or automatic summary evaluation metrics. On the other hand, manual evaluation is very expensive. According to [Lin04], large scale manual evaluations of all participants' summaries as in the DUC 2003 conference would require over 3000 hours of human efforts to evaluate summary content and linguistic qualities. A study by [DM07] showed that evaluating the content of a summary is more difficult compared to evaluating its linguistic quality. Despite the inadequacy of summary evaluation standards, the evaluation metrics commonly used are discussed in this section.

As discussed in Chapter 1, to evaluate the content and coherence of summaries generated by our approach, we have used the standard measures of precision, recall, F-measure, and ROUGE-2 and ROUGE-SU4 scores to evaluate their content. Moreover, we have also conducted manual evaluations by human evaluators in order to evaluate their content and linguistic quality as well. This will be discussed in Chapter 7.

2.6 Conclusion

In this chapter, we have reviewed current query-based summarization approaches and discussed how they address question irrelevance and discourse incoherence issues. We have also discussed challenges involved in summary evaluation and described the current summary evaluation metrics.

Current query-based approaches are mostly developed for news articles and focus on content extraction. Generally, they pay little attention to summary organization which is a crucial issue for automated text summarization that still needs to be addressed. Available approaches mostly try to rank sentences by assigning them scores based on their similarity to the user given question. To calculate similarity, statistical approaches, machine learning approaches, and graph-based approaches may be used based on different features such as term frequency and sentence position. Currently, sentences are typically viewed as bag-of-words without considering the semantics of words, phrases, and larger units. As a result, these approaches often suffer from question irrelevance. To organize summary sentences, current approaches mostly use sentence scores which cannot ensure discourse coherence. As a result, they often produce incoherent summaries. For example, in the TAC 2008 opinion summarization track, participants' average scores for summary coherence was 1.39 out of 5.

Blog summarization is a relatively new endeavor which uses similar techniques as

query-based news summarization for content selection and organization. In addition to these, for content selection, blog summarization use polarity information (e.g. positive, negative, neutral) and subjectivity scores to rank sentences. The polarity and subjectivity scores are calculated using dictionaries or machine learning approaches. Current blog summarization approaches also often suffer from question irrelevance and discourse incoherence.

Schema-based and discourse relation-based approaches have been used to improve coherence and question relevance of automated summaries. However, these approaches are applied to very structured domains or genres and use only a few discourse relations. In our work, we will show how a wide range of discourse relations could be utilized domain independently for an unstructured genre like blogs.

In the next chapter, we will discuss our methodological study of blog summaries to identify blog specific errors and our attempt to quantify the information processing difference between blog and news summaries.

Chapter 3

Blog-Specific Summarization Errors

As discussed in Section 1.1, most summarization approaches have been developed to process factual information from traditional news articles. Blogs are different in style and structure compared to news articles. As a result, successful natural language approaches that deal with news articles might not be as successful for processing blogs; thus the adaptation of existing successful Natural Language Processing (NLP) approaches for news articles to process blogs is an interesting and challenging question. The first step towards this adaptation is to identify the differences between these two textual genres in order to develop approaches to handle this new genre of texts (blogs) with greater accuracy.

3.1 News Summarization versus Blog Summarization

As most previous work has been performed on news summarization, it is not surprising that the performance of such systems are generally higher than blog summarizers. Table 6 shows the summary evaluation using the pyramid score, linguistic quality, and responsiveness (described in Section 2.5) of the systems participating at the TAC 2008 conference. The pyramid scores were calculated manually on a scale of 0 to 1 and the last two criteria were evaluated by human assessors on a scale of 1 to 5 (1, being the worst). In this evaluation, the pyramid score was used to calculate content relevance; the linguistic quality score was used to measure linguistic quality such as readability, coherence; and the responsiveness of a summary was used to judge the overall quality and usefulness of the summary, considering both the information content and readability. As shown in Table 6, the average scores for news summaries (the update summarization track) are higher than for blog summaries (the opinion summarization track) using all 3 evaluation criteria. The best system for the news also performs better than the best blog summarization system.

Table 6: TAC-2008 Summarization Results - Blogs vs. News

Genre	Pyramid Score	Linguistic Quality	Responsiveness Score
Blogs (Average)	0.10	2.13	1.31
News (Average)	0.26	2.35	2.32
Blogs (Best)	0.25	2.18	1.95
News (Best)	0.36	3.25	2.79

The difference in performance between blogs and news summarization can be attributed to several factors, most notably the fact that news have been studied more than blogs, the availability of better and more training data and the differences in the two textual genres. Indeed, one of the essential characteristics of blogs as

opposed to news, is their subjectivity (or opinion) [ABU07]. Unlike traditional news summarization, sentiment (subjectivity) plays a key role in blog summarization where sentiment degree is often used to rank sentences and sentiment analysis is a difficult task on its own. In addition, as opposed to traditional news, blogs are usually written in casual language and may contain unrelated information such as ads, photos, music, videos... A sample news article from the AQUAINT-2 collection (described in Section 2.3.2) and a sample blog post from the BLOG06 corpus (described in Section 2.3.2) are shown in Figure 8 and Figure 9, respectively. The news article of Figure 8 was provided as an input document on the topic "Airbus A380". In this article, most sentences are relevant to the topic. On the other hand, the sample blog shown in Figure 9 was distributed as an input document on the topic "Starbucks coffee shops". This sample blog contains many topic irrelevant sentences as well as an image, ads, and links.

Figure 8: Sample News Article from the AQUAINT-2 Collection

Paris airport neighbors complain about noise from giant Airbus A380 TOULOUSE, France, April 27

An association of residents living near Paris's Charles-de-Gaulle airport on Wednesday denounced the noise pollution generated by the giant Airbus A380, after the new airliner's maiden flight.

French acoustics expert Joel Ravenel, a member of the Advocnar group representing those who live near Charles de Gaulle, told AFP he had recorded a maximum sound level of 88 decibels just after the aircraft took off from near the southwestern city of Toulouse.

The figure makes the world's largest commercial jet "one of the loudest planes that will for decades fly over the heads of the four million people living in the area" outside Paris, Advocnar said in a statement.

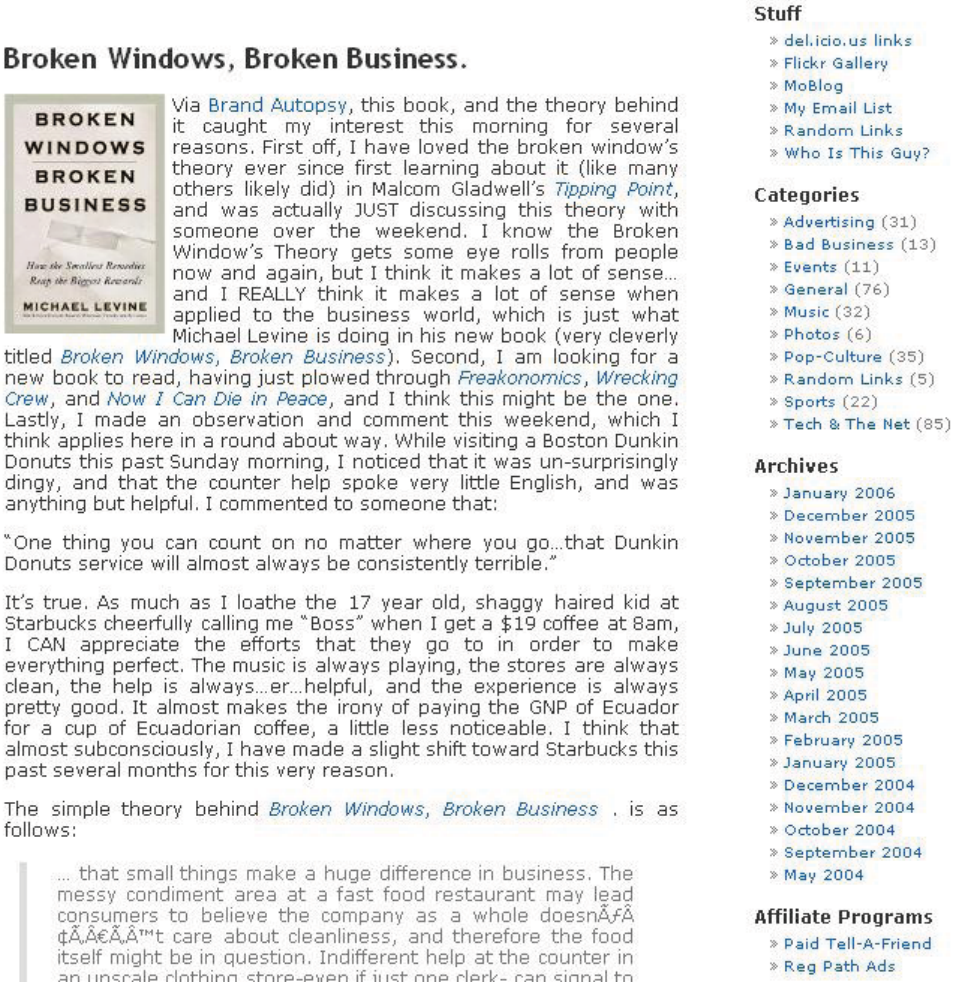
Ravenel said sound levels near Charles de Gaulle airport normally reached about 40 decibels.

Journalists watching the Airbus A380's first flight at Toulouse airport in southwestern France, however, noted how quiet the take-off and landing had seemed.

Tens of thousands of spectators cheered as the A380 touched down at the airport near Toulouse, home of the European aircraft maker Airbus Industrie, after a test flight of three hours and 54 minutes.

In general, for blogs, it is often difficult to find sentence boundaries because punctuation and capitalization are unreliable. As a result, for blog summarization, systems need to put additional efforts to pre-process the input texts (blogs) compared

Figure 9: Sample Blog Post from the BLOG06 Corpus



Broken Windows, Broken Business.

Via Brand Autopsy, this book, and the theory behind it caught my interest this morning for several reasons. First off, I have loved the broken window's theory ever since first learning about it (like many others likely did) in Malcom Gladwell's *Tipping Point*, and was actually JUST discussing this theory with someone over the weekend. I know the Broken Window's Theory gets some eye rolls from people now and again, but I think it makes a lot of sense... and I REALLY think it makes a lot of sense when applied to the business world, which is just what Michael Levine is doing in his new book (very cleverly titled *Broken Windows, Broken Business*). Second, I am looking for a new book to read, having just plowed through *Freakonomics*, *Wrecking Crew*, and *Now I Can Die in Peace*, and I think this might be the one. Lastly, I made an observation and comment this weekend, which I think applies here in a round about way. While visiting a Boston Dunkin Donuts this past Sunday morning, I noticed that it was un-surprisingly dingy, and that the counter help spoke very little English, and was anything but helpful. I commented to someone that:

"One thing you can count on no matter where you go...that Dunkin Donuts service will almost always be consistently terrible."

It's true. As much as I loathe the 17 year old, shaggy haired kid at Starbucks cheerfully calling me "Boss" when I get a $19 coffee at 8am, I CAN appreciate the efforts that they go to in order to make everything perfect. The music is always playing, the stores are always clean, the help is always...er...helpful, and the experience is always pretty good. It almost makes the irony of paying the GNP of Ecuador for a cup of Ecuadorian coffee, a little less noticeable. I think that almost subconsciously, I have made a slight shift toward Starbucks this past several months for this very reason.

The simple theory behind *Broken Windows, Broken Business* . is as follows:

> ... that small things make a huge difference in business. The messy condiment area at a fast food restaurant may lead consumers to believe the company as a whole doesnÅƒÂ ¢Ã‚Â€Ã‚Â™t care about cleanliness, and therefore the food itself might be in question. Indifferent help at the counter in an upscale clothing store-even if just one clerk- can signal to

Stuff
- » del.icio.us links
- » Flickr Gallery
- » MoBlog
- » My Email List
- » Random Links
- » Who Is This Guy?

Categories
- » Advertising (31)
- » Bad Business (13)
- » Events (11)
- » General (76)
- » Music (32)
- » Photos (6)
- » Pop-Culture (35)
- » Random Links (5)
- » Sports (22)
- » Tech & The Net (85)

Archives
- » January 2006
- » December 2005
- » November 2005
- » October 2005
- » September 2005
- » August 2005
- » July 2005
- » June 2005
- » May 2005
- » April 2005
- » March 2005
- » February 2005
- » January 2005
- » December 2004
- » November 2004
- » October 2004
- » September 2004
- » May 2004

Affiliate Programs
- » Paid Tell-A-Friend
- » Reg Path Ads

to news article summarization. Furthermore, because blogs do not exhibit a stereotypical structure, some features such as position of sentence, or similarity with the first sentence, which have been shown to be useful for traditional news articles summarization ([DO08]) are not as useful for blog summarization (shown in [BG08]). As a result, for blogs, it is usually very difficult to identify which units are relevant to the query. On the other hand, news articles are more uniform in style and structure.

3.2 Related Work

To the best of our knowledge there has been little work carried out to compare the difference between blogs and news articles; however, none seems to have analyzed it at the linguistic level for a specific NLP application.

As described in Section 2.3.2, [KLC06] developed a language independent opinion summarization approach. They tested their approach with blogs and news articles for English and Chinese languages. From their evaluation, they found that blog summaries contain more question irrelevant information compared to news articles. Their results confirm our own results (see Section 3.3). [KLC06] also found that news articles use a larger vocabulary compared to blogs which makes the task of filtering non-relevant sentences harder for news articles. On the other hand, this larger vocabulary helps to determine sentiment polarity. Due to their limited vocabulary, the judgment of sentiment polarity of blogs was difficult.

[SWWS07] developed an opinion question answering approach for blogs and news articles. They exploited attitude information namely sentiment and argument types to answer opinion questions. They obtained comparable result with both text types.

[LKS06] developed the Lydia system to analyze blogs. They analyzed the temporal relationship between blogs and news articles. In particular, they analyzed how often bloggers report a story before newspapers and how often bloggers react to news that have already been reported.

Though both the work [LKS06] and [GSS07] handle news text and blogs, their application domains (temporal relationship and sentiment analysis) are different from ours. [SWWS07] tested their question answering approach for news articles and blogs. They compared their approach for both genres of text mainly on the basis of subjectivity information. On the other hand, we compared summaries of both text types on the basis of errors which mainly occurred due to the informal style

and structure of blogs. Our work is most similar to [KLC06]'s work. However, we identified a larger number of errors of summarization (see Section 3.3) and compared blog summaries with traditional news article summaries on the basis of these errors. As a result, our work will better enable us to pinpoint the difference between these two genres of texts for a summarization task.

3.3 Error Analysis

To analyze the different challenges posed by blog summarization as opposed to traditional news summarization in greater detail, we first tried to identify and categorize errors which typically occur in opinion summarization through an error analysis of the current blog summarizers. The goal was to identify the most frequently occurring errors. In this error analysis, we compared blog summaries with traditional news summaries to assess whether there is any information processing difference needed for these two genres of texts. For this analysis, we tried to find two tasks that were similar in nature but used two different datasets; news and blogs. We chose to use the summaries from participating systems at the TAC 2008 opinion summarization track and the first set of summaries from participating systems at the update summarization track. Summaries of the TAC 2008 opinion summarization track and update summarization track were generated from blogs and news articles, respectively.

As described in Section 2.3.2, at the TAC 2008 opinion summarization track, a set of target topics on various events or entities were given on which participating systems were evaluated. For each topic, a set of questions and a set of relevant blog entries were provided. For example, for the topic "*Jiffy Lube*", two questions were asked:

1. "*What reasons are given for liking the services provided by Jiffy Lube?*"

2. "*What reasons are given for not liking the services provided by Jiffy Lube?*"

and a set of IDs of related blogs were provided. Participating systems needed to extract answers to questions from these specified sets of blogs and summarize them. In this TAC 2008 opinion summarization track, 50 questions on 28 topics were distributed. For each question, from 9 to 39 relevant blogs, which are part of BLOG06, were provided. The TAC 2008 opinion summarization track provided a total 600 blogs as the input document set.

On the other hand, in the updated summarization track, the test dataset comprised 48 topics. In this track, on each topic, a topic statement and 10 relevant documents were distributed to create the first set of summaries[1]. These documents were collected from the AQUAINT-2 collection of news articles. For example, on the topic "*Airbus A380*", the topic statement was:

1. *Describe developments in the production and launch of the Airbus A380.*

In this task, participating systems needed to generate summaries of 100 words using 10 related documents to answer the topic statement.

The systems participating in the TAC opinion summarization track and in the update summarization track are quite different in several aspects, as they were designed to address two different tasks. The systems participating in the update summarization track were mainly required to find the answers to a given question, to summarize them, and detect redundant information; while the systems participating in the opinion summarization track were required to perform opinion mining and polarity classification (described in Section 2.3.2) in addition. Moreover, the systems

[1]In the update summarization track, on each topic, participants also needed to generate an update summary on another 10 documents, which chronologically followed the first set of documents, based on the assumption that readers already read the first summary. In our analysis, we considered only summaries generated on the first set of documents because participants of TAC 2008 opinion summarization track did not generate update summaries for blogs.

participating in the opinion summarization track were provided optional snippets (described in section 2.3.2) and were restricted to a maximum summary length which was much higher compared to the one for the update summarization track. Despite these differences, these two datasets were used in our comparison because both tasks are similar in nature in the sense that their goal is to generate query-based summaries and to our knowledge are the most similar pairs of datasets apart from their genres.

Figure 10: Sample Summary from TAC 2008 Opinion Summarization Track

Topic: Jiffy Lube

Question: What reasons are given for not liking the services provided by Jiffy Lube?

Summary

I didn t think it was the serpentine belt, because Jiffy Lube had just replaced that. THE BATTERY CABLES WERE UPSIDE DOWN. Apparently the connectors are tapered in one direction, and when the boys at Jiffy Lube put my new battery in a few months ago, they put them on upside down. When I worked for my last auto shop, our best oil change was 49. 99 to the same exact one for jiffy lube was damn near 100 bucks. Billy Comment by Billy| 01/ 16/ 06 at 7: 29 am Well, I suppose it is a little bit of a PITA to figure out what to do with the spent oil, but after some digging, I found out that every Jiffy Lube will take used oil for free! It seems that the What pissed me off was that the Jiffy Lube monkey lied to me and told me that there was no fuel filter in that car.. Or maybe he simply didn t know that some cars have an internal fuel filter. kids at Jiffy Lube screwed the 17mm oil drainage bolt in so hard that it would take a super human to unscrew it, or leverage. My bet is that Dallas will probably just go back to Jiffy Lube, since I' m not capable of a simple oil change. Remember, it s in the Quick Lube s best interest for it to break down so that way you can hurry up and GET BACK THERE FOR ANOTHER OIL CHANGE! IT was awesome seeing you!! I recommend Jiffy Lube for Oil changed. they ll like.. clean your windows, vacuum your floors and shit!!;)(Reply to this)(Thread) galaxyjen 2006- 01- 29 02: 35 pm UTC(link) Really?? I get my lube done by an expert that goes slow and deliberate; that s the only kind to fool with.

In this study, we have studied 50 summaries from participating systems at the TAC 2008 opinion summarization track and compared these to 50 summaries from the TAC 2008 update summarization tracks.

The average summary length of the opinion summarization track was 612 words, while that of the updated summarization track was 90 words. The average input documents length of the opinion summarization track was 1888 words, while that of the update summarization track was 505 words. Figure 10 and Figure 11 show a sample

Figure 11: Sample Summary from TAC 2008 Update Summarization Track

Title: Airbus A380

Narrative: Describe developments in the production and launch of the Airbus A380.

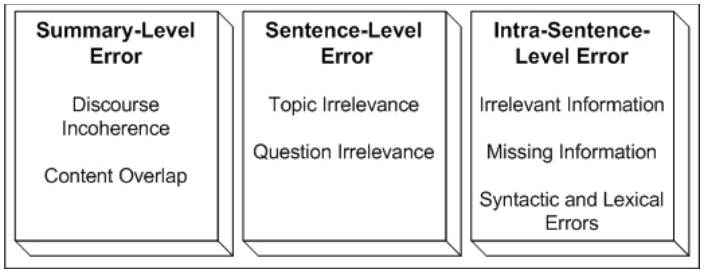
Summary

The double-decker airliner, capable of carrying up to 800 passengers, is a key factor in Airbus's battle with US aircraft maker Boeing for market dominance. The superjumbo Airbus A380, the world's largest commercial airliner, took off Wednesday into cloudy skies over southwestern France for its second test flight. December 19, 2000: Airbus officially launches the plane. January 23, 2002: Production starts of Airbus A380 components. Airbus has 154 firm orders for the A380, 27 of them for the freighter version. In the next 20 years the number of airports that could support A380 flights will grow substantially.

summary from TAC 2008 opinion summarization track and update summarization track, respectively. Summaries were randomly selected for the evaluation; however, we ensured that we selected summaries from all participating systems on all topics.

We have analyzed the most common types of errors in our 100-summary corpus

Figure 12: Types of Errors in Blog vs. News Summaries



and have categorized them in 3 main categories:

1. *Summary-Level Error*

2. *Sentence-Level Error*

3. *Intra-Sentence-Level Error*

The errors which we found under these three categories are shown in Figure 12 and discussed in the following sub-sections.

3.3.1 Summary-Level Errors

We define a Summary-Level Error as the textual contents which reduce the understandability and readability of the overall summary. There are two types of Summary-Level errors:

1. *Discourse Incoherence*

2. *Content Overlap*

Discourse Incoherence

A summary will exhibit a discourse incoherence if the reader cannot identify the communicative intentions of the writer from the clauses or if the clauses do not seem to be interrelated (discussed in [Hov93]). In the sample summaries that we have studied, discourse incoherence occurred both at the sentence level and at the clause level.

Consider the following summary (ID:T1004.20[2]) where a discourse incoherence occurs at the sentence level:

> **ID:** *T1004.20*
> **Topic:** *Starbucks coffee shops*
> **Question:** *Why do people like Starbucks better than Dunkin Donuts?*
>
> **Summary:** *I am firmly in the Dunkin' Donuts camp. It's a smooth, soothing cuppa, with no disastrous gastric side effects, very comforting indeed. I have a special relationship with the lovely people who work in the Dunkin' Donuts in the Harvard Square T Station in Cambridge. I was away yesterday and did not know. [...]*

In this summary, the underlined sentence is not coherent with the rest of the text because its communicative goal is not clear with respect to the rest of the text.

[2]All summaries numbered ID:Txxxx.xx are taken from the TAC 2008 opinion summarization track.

Consider the following summary where a discourse incoherence occurs at the clause level:

ID: *T1001.3*
Topic: *Carmax*
Question: *What motivated positive opinions of Carmax from car buyers?*

Summary: *At Carmax, the price is the price and when you want a car you go get one. Tyler Sink Says: December 7th, 2005 at 9:22 am, with Carmax you will generally always pay more than from going to a good used car dealer. [...]*

In the underlined sentence, the textual fragments *Tyler Sink ...* and *with Carmax ...* do not seem to be related to one another.

Extractive summarization approaches can deal with discourse incoherence which occurs at the sentence level. On the other hand, it is very difficult to address discourse incoherence which occur at the clause level because text generation or semantic interpretation is required for that.

In our work, we address discourse incoherence which occur at the sentence level.

Content Overlap

If semantically similar information exists in different units of a text, then we define it as content overlap. Content overlap can range from a simple duplication of text fragments to a more complex textual entailment problem. For example, consider the summary below:

ID: *T1019.35*
Topic: *China one-child per family law*
Question: *What complaints are made about China's one-child per family law?*

Summary:*[...] $6400 - a typical fine for having more than one child- in China is about 2-3 years salary.[...] Imagine losing your job, being fined 2-3 years salary for having a second child. [...]*

In this summary, the underlined sentences carry similar contents. So it may seem redundant to include both sentences in the final summary.

Table 7: Summary-Level Errors - Blogs vs. News

Error Type	**Blogs**	**News**	Δ
Discourse Incoherence	31%	11%	20%
Content Overlap	19%	15%	4%

Table 7 compares Summary-Level errors in our 50 blog summaries corpus and our 50 news articles summaries corpus. As the table shows, opinionated blog summarization and non-opinionated news articles summarization both exhibit an important number of discourse incoherence and content overlap errors. However, blog summarization have around 20% more discourse incoherence and about 4% more content overlap errors, than those of news article summarization. We suspect that the reason behind this is that because blogs are generally informal in nature, blog clauses themselves are often incoherent and contain redundant information. On the other hand, the formal nature of news articles reduces these errors for news articles summarization.

3.3.2 Sentence-Level Errors

If a summary sentence is irrelevant to the central topic of the input documents or to the user question, then the summary contains a Sentence-Level error. Two types of Sentence-Level errors were identified:

1. *Topic Irrelevance*

2. *Question Irrelevance.*

Topic Irrelevance

As mentioned earlier in this section, both in the TAC 2008 opinion summarization track (blogs) and the update summarization track (news texts), participating systems needed to generate a summary answering a set of questions on a specific target (topic). However, in both tasks, many systems generated summaries containing sentences that were not related to the specified topic. Here is an example of a topic irrelevance error:

ID: *T1004.33*
Topic: *Starbucks coffee shops*
Question: *Why do people like Starbucks better than Dunkin Donuts?*

Summary: *Well ... I really only have two. [...] I didn't get a chance to go ice-skating at Frog Pond like I wanted but I did get a chance to go to the IMAX theatre again where I saw a movie about the Tour de France it wasn't that good. [...]*

Question Irrelevance

Many of the system-generated summary sentences are not relevant to the question even though they are related to the topic. An example of a question irrelevance error is shown below:

ID: *T1004.3*
Topic: *Starbucks coffee shops*
Question: *Why do people like Starbucks better than Dunkin Donuts?*

Summary:*Posted by: Ian Palmer — November 22, 2005 at 05:44 PM Strangely enough, I read a few months back of a coffee taste test where Dunkin' Donuts coffee tested better than Starbucks. [...] Not having a Dunkin' Donuts in Sinless City I am obviously missing out... but Starbucks are doing a Christmas Open House today where you can turn up for a free coffee. [...]*

The underlined sentence is relevant to the topic but not to the question.

Table 8: Sentence-Level Errors - Blogs vs. News

Error Type	**Blog**	**News**	**Δ**
Topic Irrelevance	42%	6%	36%
Question Irrelevance	48%	17%	31%

Table 8 compares Sentence-Level errors for blog summaries and for news text summaries. Note that topic irrelevance is calculated based on the entire corpus. However, question irrelevance is calculated based only on the sentences which are related to the topic. Table 8 shows that a large number of sentences from blog summaries suffer from topic irrelevance and question irrelevance errors. In contrast, in news articles summarization, topic irrelevance errors occur only occasionally and question irrelevance errors are also not as frequently as in blog summaries. Blog summaries have around 30% more of these two errors than news article summaries.

We suspect that the main reason behind such a difference is brought about by the summary evaluation scheme. Indeed, many systems use the maximal summary length (7000 characters per question) allowed in TAC which results in many out of context sentences to be used as filler. As a result, the average summary length of the

opinion summarization track is much longer than that of the update summarization track (612 words versus 90 words). Another important reason for these errors is the informal style and structure of blogs. 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. In blogs, bloggers express various opinions about the topic which are not relevant to the question. Together all these issues may lead to a high number of topic and question unrelated sentences in the final summary.

3.3.3 Intra-Sentence-Level Errors

Intra-Sentence-Level errors occur within a sentence and involve irrelevant or missing information, grammatical errors, or lexical errors. Intra-Sentence-Level Errors include:

1. *Irrelevant Information*
2. *Missing Information*
3. *Syntactic and Lexical Errors*

Irrelevant Information

Under irrelevant information errors, a significant portion of a sentence is irrelevant to the summary topic or question. For example, consider the summary below:

ID: *T1003.9*
Topic: *Jiffy Lube*
Question: *What reasons are given for liking the services provided by Jiffy Lube?*

Summary: *They know it's fine cause Jiffy Lube sent them a little card in the mail and they have about a month before they need an oil change. [...] Well, they suppose it is a little bit of a PITA to figure out what to do with the spent oil, but after some digging, they found out that every Jiffy Lube will take used oil for free! [...]*

The underlined snippet above is irrelevant to the question even though it holds a coherent discourse relation with the last clause.

Missing Information

If a sentence does not contain all the necessary information to make it comprehensible for the reader and the required information to understand the sentence is also not available in the context then this error is defined as a missing information error. An example of missing information errors is shown below:

ID: *T1021.17*
Topic: *Sheep and Wool Festival*
Question: *Why do people like to go to Sheep and Wool festivals?*

Summary: *[...] i hope to go again this year and possibly meet some other knit bloggers this time around since i missed tons of people last year. I love going because of the tons of wonderful people, yarn, Sheep, rabbits, alpacas, llamas, cheese, sheepdogs, fun stuff to buy, etc., etc. [...]*

The underlined sentence contains incomplete information, which cannot be resolved from the context, making it incomprehensible.

Syntactic and Lexical Errors

Syntactic level errors such as grammatical incorrectness and incompleteness of a sentence or lexical level errors such as spelling errors, short forms, stylistic twists of informal writing . . . in a sentence are all included in syntactic and lexical errors.

For example, consider the following summary:

ID: *T1009.32*
Topic: *Architecture of Frank Gehry*
Question: *What compliments are made concerning his structures?*

Summary: *Central to Millennium Park in Chicago is the Frank Gehry-designed Jay Pritzker Pavilion, described as the most sophisticated outdoor concert venue of its kind in the United States. [...] Designing a right-angles-be-damned concert hall for Springfield, hometown of Bart et al.. [...]*

In this summary, the underlined sentence is an example of a syntactic and lexical error.

Table 9: Intra-Sentence-Level Errors - Blogs vs. News

Error Type	**Blog**	**News**	**Δ**
Irrelevant Information	30%	15%	15%
Missing Information	9%	2%	7%
Syntactic and Lexical Errors	19%	4%	15%

Table 9 compares Intra-Sentence-Level errors for blog summaries and for news article summaries. From Table 9, we can see that irrelevant information, missing information, and syntactic and lexical errors appear about 15%, 7%, and 15% more respectively in blog summarization. Here again, we believe that the informal nature of blogs explains these difference.

3.3.4 Discussion

Compared to a manual linguistic evaluation of a summary, our analysis tried to identify and quantify the differences in error types between two textual genres: blogs and news for the purpose of summarization.

Our error types incorporate both what the automatic and manual summary evaluation try to measure. Indeed, Sentence-Level errors (topic irrelevance and question irrelevance) evaluate the content and relevance of the summaries similarly to what an automatic metric tries to evaluate (see Section 2.5); whereas the remaining errors (Summary-Level errors and Intra-Sentence errors) evaluate more the linguistic quality of a summary.

It is not surprising to see that topic irrelevance, question irrelevance, and discourse incoherence are much more frequent in blogs than in news articles (from 36% to 20% more frequent). Content overlap and missing information, on the other hand, seem to be only slightly more frequent (5% and 7%) in blogs summaries than in news article summaries. These results give a clear idea of the challenges we face when dealing with blogs for summarization compared to news articles and where efforts should be made to improve such summaries.

3.4 Conclusion

The performance of blog summarization is generally much lower than for news article summarization. The purpose of this chapter was to analyze these differences and compare automatically-generated summaries for blogs with news texts based on the most common errors which occurred in summarization. The goal of our comparison was to assess whether these summary-related errors affect traditional news articles based non-opinionated summaries differently than opinionated blog summaries.

Our results show that all types of summary-related errors occur more often in blog summarization than in news article summarization. However, topic and question irrelevance as well as discourse incoherence pose a much greater problem for blog summarization than for traditional news articles; while content overlap and missing information seem to be only slightly more frequent in blogs than in traditional news articles. These results show how difficult it is to process blogs for summarization and show that different information processing techniques are required for these two genres of texts. Based on the results of this study and others (e.g. [CD08, DO08]) (described in Section 1.3), we focused our efforts to address question irrelevance and discourse incoherence errors which occur most frequently. It would be interesting to address all other summary related errors such as content overlap which we have identified in our error analysis but due to the limited scope of this work we could not address them all, so we only focused on the most frequent errors.

The next chapter will discuss how discourse relations can be utilized in a schema-based framework and how this approach will help to reduce question irrelevance and discourse incoherence of query-based blog summaries.

Chapter 4

A Schema-based Framework Utilizing Discourse Relations

The purpose of the present chapter is to show an overview of our proposed schema-based approach. Details of the predicate identification - the heart of the approach is explained in the next chapter and details of the implementation are provided in Chapter 6.

As described in Section 2.4.2, as early as 1985, [McK85] introduced a schema-based approach for text planning based on the observation that certain standard patterns of discourse organization (called schemata) are more effective to achieve a particular discourse goal. We also believe that to answer a particular type of question, certain types of sentences, if organized in a certain order, can meet the communicative goal more effectively and create a more coherent text. For example, to take [McK85]'s example, to define an entity or 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 another combination of sentences to be effective and coherent. When writing, humans also use predefined structures

to answer a particular type of question [McK85]. Based on these observations, we have developed a domain-independent schema-based approach for summarization that utilizes discourse relations to avoid question irrelevance and discourse incoherence in blog summarization.

In this chapter, we will show an overview how our schema-based approach works and how our approach can help in reducing question irrelevance and discourse incoherence. Details of the approach can be found in Chapter 5 and 6; whereas details of the evaluations can be found in Chapter 7.

4.1 Overview of Our Schema-based Approach

Given an initial question on a particular topic and a ranked list of sentences from the document set, our schema-based summarization approach identifies the most relevant sentences and their most effective order to include in the summary. Since highly ranked sentences could still be question irrelevant, we apply discourse-level and semantic-level analysis to remove question irrelevant sentences thus improving question relevance. To do so, from the ranked list of sentences, our approach selects a few most relevant sentences based on the rhetorical predicates that they contain using the appropriate schema for the given question type. Our approach also reorders these relevant sentences to improve discourse coherence. In this section, we will first briefly describe how our approach selects candidate sentences (Section 4.1.1) then we will describe in detail how it filters question-irrelevant sentences and reorders candidate sentences using schemata (Section 4.1.2).

4.1.1 Candidate Sentence Selection

To select and order sentences, our approach first needs a ranked list of candidate sentences. The candidate sentence extractor is expected to extract a list of sentences from the documents and rank them by relevance. In our current implementation, sentences are ranked based on question similarity, topic similarity, and subjectivity scores (details in Section 6.1.2). To select the initial candidate sentences, any sentence ranker, such as MEAD [RABG+04], can be used. In fact, Section 6.1.2 will describe an experiment with MEAD. Commonly used candidate sentence selection approaches for extractive summarization are described in Section 2.2.3.

Figure 13: Partial Candidate List Used as Input

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

Candidate Sentences	**Score**
(1) With Carmax you will generally always pay more than from going to a good used car dealer.	0.536
(2) We bought it at Carmax, and I continue to have nothing bad to say about that company.	0.449
(3) Carmax did split the bill (which made me happy).	0.416
(4) Not sure if you have a Carmax near you, but I've had 2 good buying experiences from them.	0.381
(5) have to say that Carmax rocks.	0.368
(6) At Carmax, the price is the price and when you want a car you go get one.	0.299
(7) Sometimes I wonder why all businesses can't be like Carmax.	0.278
(8) Arthur Smith, 36, has been living in a van outside the Carmax lot, 24 hours a day, for more than a month.	0.25

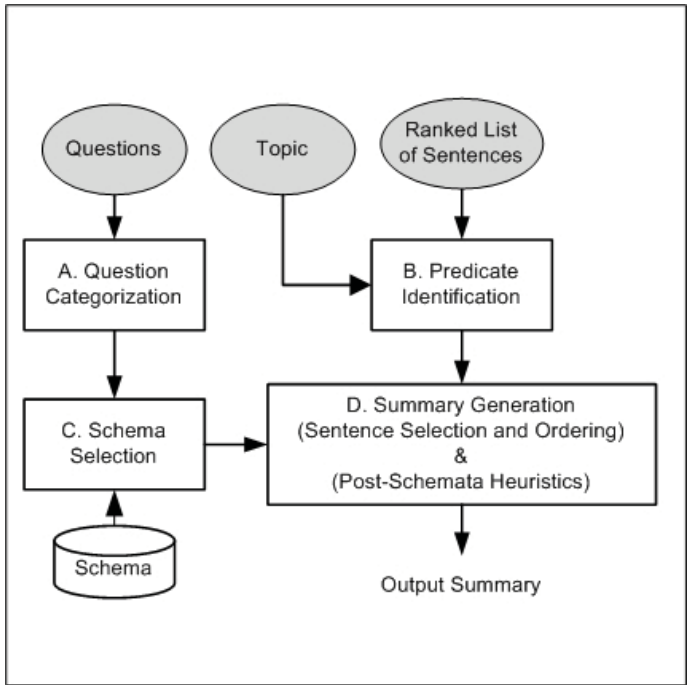
Figure 13 shows a partial candidate list to illustrate the output produced by the candidate sentence selection phase. The figure shows the 8 most relevant sentences along with the scores (out of 1) given the Topic: "*Carmax*", the Question: "*What motivated positive opinions of Carmax from car buyers?*", and a set of related blogs on the topic.

Section 6.1.2 will describe in detail how this has been implemented in our prototype system.

4.1.2 Content Filtering and Organization

An overview of the content filtering and organization is shown in Figure 14.

Figure 14: Architectural Design



As the figure shows, once we have the initial ranked list of sentences (as in Figure 13), schemata are used to organize the summary content by filtering question-irrelevant sentences from the candidate list and reorder the remaining sentences more

coherently. For content filtering and organization, our approach performs the following tasks (see Figure 14):

A. Question Categorization

B. Predicate Identification

C. Schema Selection from the Pre-designed Schemata

D. Summary Generation

Questions first need to be categorized based on their communicative goal (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 pre-designed schema needs to be selected for the specific question category (C) to incorporate the most relevant sentences into the summary. At the end of this process, a summary is generated by filtering and reordering sentences (D).

Let us describe the content filtering and organization steps in detail.

A. Question Categorization

As in our approach we want to answer different types of questions in different manners, our content organization approach first needs to categorize 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?*

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?*

Table 10 shows the question distribution of the TAC 2008 opinion summarization track dataset into the above three categories. The table shows that most of the questions were reason type (90%) and only 4% of the questions were comparison type and another 6% of the questions were suggestion type. This skewed distribution will imply that we have less data for development and testing for comparison and suggestion; however, results of Chapter 7 show that all 3 question types perform well.

Table 10: TAC 2008 Question Distribution

Question Category	**Distribution**
Comparison	4%
Reason	90%
Suggestion	6%

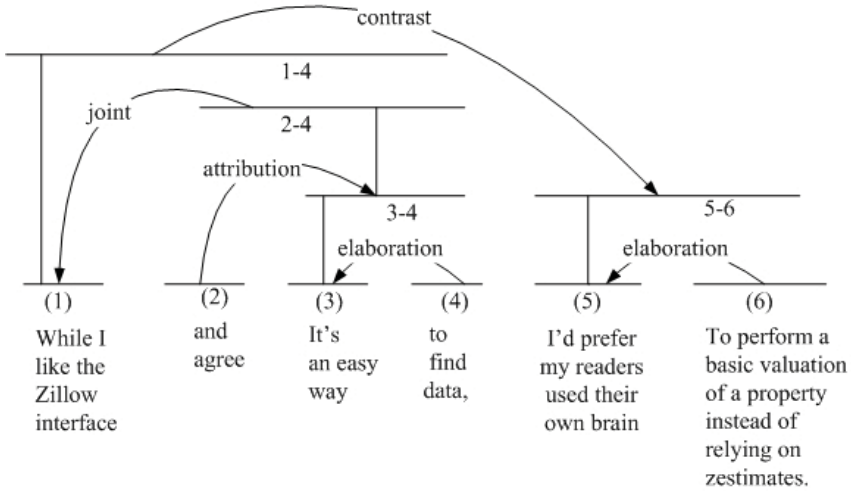
Section 6.2.1 will detail how question categorization has been implemented in our system.

B. Predicate Identification

In our schema-based approach, the basic units of a schema are rhetorical predicates (see Section 5.2). 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 (described in Chapter 5).

We considered six main categories of rhetorical predicates: *comparison*, *contingency*, *illustration*, *attributive*, *attribution*, and *topic-opinion*. Comparison, contingency, and illustration predicates can be sub-divided into sub-categories. In our approach, candidate sentences need to be tagged with these rhetorical predicates based on what discourse relations they contain. 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. In this process, 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 prefer 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, joint, attribution, and elaboration (shown in Figure 15).

Figure 15: Sample RST Tree



Given a set of candidate sentences as shown in Figure 13, the predicate identification module tags each sentence based on which rhetorical predicates it contains. This is shown in Figure 16.

Figure 16: Candidate Sentences along with Rhetorical Predicates

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

One of the most challenging tasks in our text schema-based approach for summarization is to identify which rhetorical predicate is communicated by a candidate sentence in order to figure out if it should be included in the summary and where. Because this step is crucial in our approach, Chapter 5 is dedicated to explaining our automatic predicate identification approach.

C. Schema Selection from the Pre-designed Schemata

To answer a specific question category, our content organization approach uses the associated schema (e.g. comparison) which is designed for that particular question category to select and order sentences for the final summary.

In order not to answer all questions the same way, we designed appropriate

schemata to generate a summary that answers specific types of questions. In human writing, writers often use stereotypical patterns to answer a specific type of question to make the answer relevant to the question. Based on this observation, our prototype uses three schemata, one for each question type that we have considered:

1) *Comparison*

2) *Reason*

3) *Suggestion*

To design these schemata, we have studied 15 articles of each type written by different authors. 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 (shown in Table 11).

Table 11: Corpus Analyzed to Design Schemata

Schema	**Dataset**	**Example**
Comparison	15 Compare/contrast essays, comparison review articles	"How to Write a Compare-and-Contrast Essay." "Gas vs. Diesel Comparison Review Article - Truck Trend."
Reason	15 Argumentative essays	"How to Write an Argumentative Essay." "Argumentative Essays - OWL - Purdue University."
Suggestion	15 Problem-solution essays	"A Problem-Solution Essay." "Problem Solving Essay Writing Techniques."

From our development essay corpus analysis, we have derived which question types should be answered by which type of predicates. Each schema is designed based on giving priority to its associated question type and subjective sentences as we are generating summaries for opinionated texts. Each schema specifies the types of predicates and the order in which they should appear in the output summary for a particular question type.

Figure 17 shows the reason schema that is used to answer *reason* questions. According to this schema, a sentence to be included at the beginning of the summary needs to contain either a topic-opinion predicate or an attribution predicate followed by contingency or comparison predicates then by attributive predicates. More formally, one or more topic-opinion or attribution predicates followed by zero or many contingency or comparison predicates followed by zero or many attributive predicates can be used.

Figure 17: The Reason Schema

Predicates & Constraints
Predicate: {*Topic-opinion/Attribution*}$^{+}$ Constraint: Predicates must have the same polarity as the question.
Predicate: {*Contingency/Comparison*}* Constraint: The topic of the sentence needs to be the focus of the sentence, predicates must contain the topic as one of the objects which are being compared.
Predicate: *Attributive** Constraint: The topic of the sentence needs to be the focus of the sentence.

Constraints for Schemata

In schema design, we have also defined constraints on the predicates, a novelty compared to [McK85]'s schemata. Figure 17 shows the constraints associated on each predicate of the reason schema. Constraints restrain the sentences that can fill the schema based on their semantic content. This is done to ensure that the sentences are topic-relevant and question-relevant. Constraints can be of different types:

1. Constraints on Sentence Polarity

This constraint ensures that the sentences included in the summary will have the correct polarity with respect to the question to be question-relevant. For example, if the question asks "*why do people like X?*", then sentences that discuss negative aspect of X should not be included. Figure 18 shows a more concrete example. The

polarity of the question shown in Figure 18 is positive. In this example, Sentence 1 and Sentence 2 are both categorized as topic-opinion sentences and their polarity is positive and negative, respectively. Since the question in Figure 18 is reason type, it will be answered by the reason schema. According to the reason schema shown in Figure 17, Sentence 1 will be added to the summary but Sentence 2 will not because its polarity is not same as the polarity of the question. This constraint is applied on topic-opinion and attribution predicates.

Figure 18: Example of Constraint on Sentence Polarity

	Predicate	Polarity
Question: Why do people like Subway Sandwiches?		Positive
Sentence 1: Subway used to be a haven, an oasis of lean sandwiches in a desert of fried meat products.	Topic-opinion	Positive
Sentence 2: Subway has bad food	Topic-opinion	Negative

2. Constraints on Sentence Focus

Constraints on sentence focus ensure that sentences included in the summary are topic relevant. According to this constraint, the topic of the sentence needs to be the focus of the sentence. Implementation details of this constraint is discussed in Chapter 6. In general, this constraint is applied on the attributive, contingency, and comparison predicates.

3. Constraints on the Compared Objects

Constraints on the compared objects ensure that sentences included in the summary are topic relevant as well as question relevant. This constraint is applied on the comparison predicate. This constraint on comparison predicates is varied based on its associated schema type. For example, as shown in Figure 17, the reason schema includes a constraint on comparison predicates that "they must contain the topic as

one of the objects which are being compared". For example, the sentence on the topic *Subway* "Overall I think Subway is one of the best restaurant on MM Alam road." contains a comparison predicate and also fulfils the constraint. However, the sentence "Obama is more ... than Bush." would not be included in a summary on the topic of *Subway*. On the other hand, the comparison schema includes a constraint on comparison predicates that "they must contain all objects or events which are being compared". For example, on the topic *Chrome, Firefox* "I like Chrome better than Firefox." contains a comparison predicate and also fulfils the constraint in this case. How these constrains are implemented are discussed in Chapter 6.

Now if we look at the reason schema of Figure 17, the topic-opinion and attribution predicates must satisfy constraints on sentence polarity; contingency and attributive predicates must fulfill the constraints on sentence focus; and comparison predicates must satisfy the constraints on the compared objects that they must contain the topic as one of the objects which are being compared and fulfill the constraints on sentence focus.

Figure 19: The Comparison Schema

Predicates & Constraints
Predicate: {*Comparison*/*Contingency*} $^{+}$ Constraint: Predicates must contain all objects or events which are being compared, the topic of the sentence needs to be the focus of the sentence. Predicate:{*Topic-opinion*/*Attribution*} * Constraint: Predicates must have the same polarity as the question. Predicate: *Illustration**

Figure 19 shows the comparison schema that we used to answer a comparison question. According to this schema, a sentence to be included in the beginning of the summary needs to be classified as either a comparison predicate or a contingency

predicate followed by topic-opinion or attribution predicates then by illustration predicates. More formally, one or more comparison or contingency predicates followed by zero or many topic-opinion or attribution predicates followed by zero or many illustration predicates can be used. From Figure 19, we can see that constraints are also defined on predicates based on their semantic content. In the comparison schema, the comparison predicates must contain all objects or events which are being compared and satisfy the constraints on sentence focus; contingency predicates must fulfill the constraints on sentence focus; and topic-opinion and attribution predicates must satisfy constraints on sentence polarity.

Figure 20: The Suggestion Schema

Predicates & Constraints
Predicate: *Contingency** Constraint: The topic of the sentence needs to be the focus of the sentence. Predicate:{*Topic-opinion/Attribution*} $^{+}$ Constraint: Predicates must have the same polarity as the question.

Figure 20 shows the suggestion schema used to answer suggestion question.

In order to answer a different type of questions (e.g. identification questions which can be used to provide a definition), a different schema would be more appropriate. It must be noted that the design of schemata is subjective and personal, just like writing a document is. The subjectivity and the personal writing styles of the author are important factors; however, our current content organization approach allows the generation of different summaries for particular question types by providing flexible sentence selection and reordering strategies.

In Section 6.2.3, we will discuss implementation details of these schemata.

D. Summary Generation

Once a schema is selected for a particular question type and sentences are tagged with rhetorical predicates, the most appropriate candidate sentences must be selected and ordered 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 similarity scores until the maximum summary length is reached.

Post-Schema Heuristics

While applying a schema, multiple sentences can be qualified to fill a specific position in a schema. For example, for the schema in Figure 17, there can be more than one candidate sentence that contains a comparison predicate and satisfies the constraints. Hence the use of schemata alone is not sufficient to achieve a total sentence order and several possible summaries may be produced. In order to produce the most coherent summaries, we have developed post-schema heuristics. These heuristics include: topical similarity, explicit discourse markers and aggregation, and context. At the end of the sentence ordering process, to create a linear sentence order, we finally use the rank of the sentences in the original list of candidates. Let us now describe the post-schema heuristics.

1. **Topical Similarity**: This heuristic tries to improve the final summary globally. [BEM02] demonstrated experimentally that even if human written summaries may have different discourse structures, topically similar sentences tend to stay together. They illustrate their point using a news article with the headline "China seeks solutions to its coal mine safety from the world" reproduced in Figure 21.

   In the article of Figure 21, we can see that topically similar sentences are placed

Figure 21: Example of Topical Similarity from a News Article

China seeks solutions to its coal mine safety from the world

With its coal mining safety a hot issue attracting wide attention from both home and overseas, China is seeking solutions from the world to improve its coal mining safety system.

The Conference on South African Coal Mining Safety Technology and Equipment held here Thursday provides an opportunity for Chinese coal mine production personnel to learn from their South African counterpart in the area of coal mining safety.

At the conference, experts in the coal mining industry from South Africa showcased their technology and equipment in mining safety and shared their expertise and experience in coal mine management.

Wang Shuhe, deputy director of the State Administration of Coal-Mine Safety (SACMS) said that the mining industry is well developed in South Africa. So China could learn much from the country which has a complete safety system and mines rescuing system.

...
Coal is the major energy resource in China, covering 67 percent in the country's consumption structure of all primary energy resources.

In recent years, The government has issued a series of regulations and measures to improve the country's coal mine safety situation.

together; paragraphs 2 and 3 are both discussing "the conference on South African Coal Mining Safety Technology and Equipment" and hence they are placed consecutively in the text. The next paragraph (paragraph 4) discusses a different, but related topic ("mining industry in South Africa"). Hence it is placed after all discussion on "the conference on South African Coal Mining Safety Technology and Equipment". 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 use topical similarity in order to group sentences that describe the same topic together. To find topically similar sentences, we used the cosine similarity using *tf.idf*. In principle, this should prevent the

summary from going back and forth on various topics and hence improve its coherence further. Our experimental results discussed in Section 7.5.2 support our assumption. Indeed, in a manual evaluation, 95% of the time, summaries generated using this heuristic were rated higher or equal compared to summaries generated without this heuristic (Section 7.5.2 will discuss this further).

2. **Explicit Discourse Markers and Aggregation**: This heuristic is meant to improve coherence at the local level by making explicit the discourse relations between consecutive clauses based on sentence similarity and polarity. This strategy has been used successfully by other researchers (e.g. [GS98, KD93]). The choice of the discourse marker is based on the sentences' topical similarity and polarity value (shown in Table 12).

Table 12: Discourse Markers

Topic Similarity	**Polarity**	**Discourse Marker**
High	Identical	*;*, *and*, *also*, *moreover*, *furthermore*, *in addition*
High	Opposite	*but*, *although*, *though*, *despite*, *however*, *while* *in contrast*

For example, for the question *"Why do people like Picasa?"*, three candidate sentences shown in Table 13, are extracted as candidate sentences and identified

Table 13: Sample Sentences

(1) Picasa is another Google product, that is almost enough to make it great (2) I really like Picasa as an image organizer application (3) One thing that I really like in Picasa is the ability to watch offline folders.
Summary: Picasa is another Google product, that is almost enough to make it great **and** I really like Picasa as an image organizer application. **Moreover**, one thing that I really like in Picasa is the ability to watch offline folders.

as topic-opinion sentences. Even though these sentences may not have been adjacent in the candidate list, they are topically similar and their polarity type

matches (shown in Table 12), then the first two sentences will be placed next to each other and made a single sentence out of them. As seen in Table 13, here our approach decided to use the discourse marker "and". Our approach finds the third sentence of Table 13 on this topic with the acceptable polarity value, it will place the third sentence next to this sentence using another discourse marker (e.g. moreover).

3. **Context:** This heuristic is also meant to improve coherence at the local level. To improve discourse coherence further, we try to address a frequent problem in extraction-based summarization: dangling anaphora. If a potential sentence contains a pronoun without having a potential antecedent, we include its previous sentence from the source document as a context from the original document. The phenomenon of dangling anaphora is a common one in summarization, therefore this heuristic should improve coherence substantially. More sophisticated approaches, such as probabilistic models (e.g. [BEM02, Lap03]) could be used, but we have found that this simple heuristic could be implemented with a very little processing cost. However, from the evaluation results, we have found that this heuristic does not have much effect on the summary quality (see Section 7.5.2).

To improve summary generation further, we could have used other widely used post-processing approaches such as using the sentence order of the original document set or chronological order. However, [BG08] experimentally showed that these features are not very effective for unstructured texts like blogs.

Section 6.2.4 discusses implementation details of these heuristics and Section 7.5 empirically shows the effect of our post-schemata heuristics rules on our summarization approach.

4.2 An Example to Demonstrate the Usability of Our Schema-based Approach

To better illustrate how our overall schema-based approach works, let us take the following example again:

Figure 22: Partial Candidate List Used as Input

Topic: *Carmax* **Question**: *What motivated positive opinions of Carmax from car buyers?*	
Candidate Sentences	**Score**
(1) With Carmax you will generally always pay more than from going to a good used car dealer.	0.536
(2) We bought it at Carmax, and I continue to have nothing bad to say about that company.	0.449
(3) Carmax did split the bill (which made me happy).	0.416
(4) Not sure if you have a Carmax near you, but I've had 2 good buying experiences from them.	0.381
(5) have to say that Carmax rocks.	0.368
(6) At Carmax, the price is the price and when you want a car you go get one.	0.299
(7) Sometimes I wonder why all businesses can't be like Carmax.	0.278
(8) Arthur Smith, 36, has been living in a van outside the Carmax lot, 24 hours a day, for more than a month.	0.25

Given the Topic: "*Carmax*", the Question: "*What motivated positive opinions of Carmax from car buyers?*", and a set of related blogs on the topic, our candidate sentence selector generates a ranked list of sentences. The 8 most relevant sentences along with their scores (out of 1) were shown in Figure 13, and are reproduced in Figure 22 for convenience.

The question categorization module classifies the above question as a *reason* type based on the question pattern matching (discussed in Section 6.2.1). Then the predicate identification module tags each of the candidate sentence with rhetorical predicates they contain. This is shown in Figure 23.

Figure 23: Candidate Sentences along with Rhetorical Predicates

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

For this question, the summary generation module used the *reason* schema to generate the final summary. The *reason* schema and the final order of the sentences are shown in Figure 24.

In this sample summary, we can see that the summary generation module did not include sentences 1 and 8 in the final summary. This is because these sentences did not fit within the *reason* schema. Though sentence 1 was classified as containing a comparison predicate, it did not fulfil the semantic constraint (shown in Figure 24)

Figure 24: Summary Generated using the Reason Schema

Schema	Summary
Predicate: $\{Topic\text{-}opinion/Attribution\}^{+}$ 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.

that the topic of the sentence (Carmax) be the focus of the sentence. 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 to remove question-irrelevant sentences.

We can see that since for the sentence 2, the antecedent of the pronoun *it* is missing, the post-schemata heuristic of "context" 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 discourse 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 predicate category they contained. This example shows intuitively that schemata can help filter question-irrelevant sentences and improve discourse coherence; however, Chapter 7 will provide a more formal evaluation.

4.3 Conclusion

This chapter has provided an overview of our schema-based approach and also explained how discourse relations are utilized in schema design. This chapter also demonstrated with an example that our schema-based approach can be effective in reducing question-irrelevant sentences and improving discourse coherence. A schema provides a partial ordering, therefore we also developed post-schema heuristics rules to improve coherence. Chapter 6 will provide implementation details of our approach.

This chapter has shown that rhetorical predicates are the building blocks of our schema-based approach. The next chapter will therefore discuss the set of rhetorical predicates that we have considered in our work and the automatic approaches we have used to identify these predicates.

Chapter 5

Rhetorical Predicate Identification

One of the most challenging task in our text schema-based approach for summarization is to identify which rhetorical predicate (e.g. comparison, contingency) is communicated by a candidate sentence in order to figure out if it should be included in the summary and where. In this chapter, we discuss our predicate identification approaches in detail. We focus on genre and domain independent intra-sentential rhetorical predicate identification approaches which can tag individual rhetorical predicates as opposed to performing a more complete discourse parse.

5.1 Introduction

According to [Hov93], a discourse, spoken or written, is a structured collection of clauses. The clauses are grouped into segments based on semantics or other grounds and the segments are nested to form larger segments that provide the discourse structure. Over decades, researchers have been studying the structure of discourse and facing questions such as: How do the segments relate? What inter-segment relations are there? How many relations are needed? [Hov93]. As presented in Chapter 2, over time, to utilize discourse structures in computational systems, different discourse

theories such as Rhetorical Predicates [Gri75, Hob85], Rhetoric [Ari54], Discourse Representation Theory [Kam81, Ash93], Rhetorical Structure Theory [MT88] and others (e.g. [Gro85, GL86, KD94, Hov93, HM93]) have been developed. Some theories are inclusive compared to others with respect to discourse structure definition and applicability. For example, Rhetorical Structure Theory (RST) [MT88] is comprehensive compared to its predecessors because it provides extensive definitions of various discourse relations which can connect different segments and [Mit93] showed that plan-based approaches can be used to apply these relations. However, even if the set of relations proposed by these theories are different, they are comparable.

In our research, out of the various discourse theories, we have followed rhetorical predicates theory [Gri75, Hob85] to model the discourse of our intended types of texts (blogs). According to this theory, rhetorical predicates describe the structural relations between propositions in a text where propositions can be clauses or sentences and describe different predicating acts a writer can use.

We have considered rhetorical predicates because they describe discourse structures by showing the relations between 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. Most of the discourse theories model the structures of a discourse by providing a set of relations which are used to relate clauses.

Moreover, 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 (or clause) "*Mary has a pink coat.*" provides details about an object.

Discourse structures which provide a definition or attributes of an object or a concept are also useful in many applications such as summarization and question answering. However, most of the discourse theories do not include discourse structures which are used to provide a definition or attributes of an object or a concept.

Rhetorical predicates were found useful in various computational applications. For example, [McK85] showed that rhetorical predicates can be used to select content and generate coherent text in question answering with the help of schemata. Rhetorical predicates have also been found useful for anaphora resolution ([McK85]) and machine translation ([Mit93]). However, even though rhetorical predicates are useful in many applications, their automatic identification remains a challenging task. Existing rhetorical predicate identification approaches (e.g. [McK85, Mit93]) are often domain or genre dependent. For example, in [McK85], predicates are identified based on the hierarchical structures and pre-stored relations in a knowledge base. In certain sub-languages, predicates are often identified by means of key words and other linguistic clues (e.g. *because*, *if*, *then*) or through verb frameworks [Mit93]. With verb frameworks, characteristics of a verb are defined for a specified sub-language and each verb is associated with possible rhetorical predicates. [Mit93] also used domain knowledge with verb frameworks to identify predicates.

In this chapter, we first introduce the set of rhetorical predicates which we have taken into consideration. Then we present different available approaches such as the SPADE parser [SM03] and Jindal et al.'s [JL06] work that are used to identify these rhetorical predicates. We also present our attributive and topic-opinion tagger which we developed to identify the attributive and topic-opinion predicates.

5.2 Rhetorical Predicates

As mentioned in the previous section, rhetorical predicates refer to the different predicating acts a speaker can use to communicate his/her thoughts and describe the structural relations between clauses or within a clause in a text. Some examples are *constituency* (that provides details about sub-parts), and *attributive* (that provides details about an entity or object).

Rhetorical predicates take clauses as arguments. Clauses represent the smallest units that stand in informational or interactional relationship with other parts of texts. In this framework, clauses are classified into rhetorical predicates based on their underlying information. Rhetorical predicates classify clauses into two broad categories:

1. *A clause that contains a relation with another clause.*

2. *A clause that provides information on its own.*

In the first case, rhetorical predicates describe the relation between clauses and thus express the relationship that unites them. For example, the *cause* predicate creates a relation with the stated fact in order to provide a reason; in the sentence "*[Previously, airlines were limiting the programs] [<u>because</u> they were becoming too expensive.]*" the two clauses (shown inside *[]*) are related with a *cause* relation.

In the second case, rhetorical predicates describe a relation between different objects or concepts within a clause. For example, 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. Within this category, rhetorical predicates can also provide a definition or an attribute of an object or a concept within a clause (e.g. the *attributive* predicate which describes the attributes of an object). Here, a single clause can characterize a

predicate. For example, in the sentence "*[Mary has a pink coat.]*", the single clause provides details of an object. This kind of discourse structure is not considered by most of the discourse theories except rhetorical predicates.

Our work is performed within the framework of developing a query-based summarizer for blogs. Hence, we need to consider the predicates that are most useful to our application. To find the set of the rhetorical predicates needed for our work, we have manually analyzed 50 summaries randomly selected from participating systems at the TAC 2008 opinion summarization track and 50 randomly selected blogs from BLOG06. From the corpus analysis, we have identified six types of rhetorical predicates, namely *comparison*, *contingency*, *illustration*, *attribution*, *topic-opinion*, and *attributive*. The comparison, contingency, and illustration predicates are also considered by most of the work in the field of discourse analysis such as the PDTB: Penn Discourse TreeBank research group [PMD+08] and the RST Discourse Treebank research group [CM01]. We considered three additional classes of predicates: attributive, attribution, and topic-opinion. In building our predicate model, we considered all main discourse structures listed in Mann and Thompson's Rhetorical Structure Theory (RST) taxonomy [MT88] (described in Section 2.4.1). These discourse structures are also considered in Grimes' [Gri75] and Williams' predicate lists [Wil83].

A description of these rhetorical predicates is given below:

1. **Comparison**: Gives a comparison and contrast among different situations. This predicate can be inter or intra clausal. For example, *"Its fast-forward and rewind work much more smoothly and consistently than those of other models I've had."* shows an intra-clausal comparison predicate. On the other hand, *"It said it expects full-year net of 16 billion yen, compared with 15 billion yen in the latest year."* shows an inter-clausal comparison predicate. The comparison

predicate also subsumes the contrast, analogy, and preference predicates according to the RST Discourse Treebank [CM01] and the Penn Discourse TreeBank [PMD+08].

2. **Contingency**: Provides cause, condition, reason, evidence for a situation, result or claim. This predicate mostly occurs in an inter-clausal situation. For example,

   i) *"Sears, Roebuck & Co. is struggling as it enters the critical Christmas season."*

   ii) *"The meat is good because they slice it right in front of you."*

   show two inter-clausal contingency predicates. The contingency predicate subsumes the explanation, evidence, reason, cause, result, consequence, background, condition, hypothetical, enablement, and purpose predicates according to the Penn Discourse TreeBank.

3. **Illustration**: Is used to provide additional information or detail about a situation. This predicate mostly occurs in an inter-clausal situation. For example,

   i) *"Allied Capital is a closed-end management investment company that will operate as a business development concern."*

   ii) "*The Xbox 360 and Vista both will use a new technology that makes games run at the fastest speed possible.*"

   show two inter-clausal illustration predicates. The joint, list, disjoint, and elaboration predicates are subclasses of the illustration predicate according to the RST Discourse Treebank and the Penn Discourse TreeBank.

4. **Attributive**: 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 predicate, also included in Grimes' predicates [Gri75], is considered because it describes attributes or features of an object or event and is often used in query-based summarization and question answering. This predicate mostly occurs within a clause.

5. **Attribution**: Provides instances of reported speech both direct and indirect which may express feelings, thoughts, or hopes. We considered the attribution predicate, also considered in [CM01], because by analyzing the BLOG06 dataset, we have found that the discourse structures captured by this predicate (e.g. feelings, thoughts) are often used in opinionated texts. This predicate mostly occurs between two clauses - e.g.

   i) *"The legendary GM chairman declared that his company would make "a car for every purse and purpose.""*

   ii) *"I said actually I think Zillow is great."*

6. **Topic-opinion**: We introduced topic-opinion predicates to represent opinions which are not expressed by reported speech. This predicate can be used to express an opinion; an agent can express internal feeling or belief towards an object or an event. This predicate also mostly occurs within a clause - e.g.

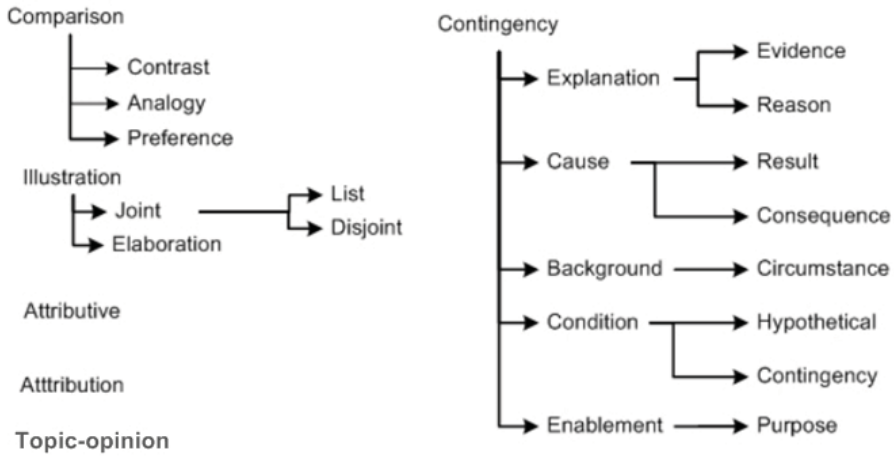
   i) *"Cage is a wonderfully versatile actor."*

   ii) *"The thing that I love about their sandwiches is the bread."*

The rhetorical predicates that we considered are summarized in Figure 25.

As stated earlier, our study focused only on these predicates as they were the most useful in our application however, other predicates would also be interesting to consider, for example *antithesis* - e.g. *"Although the legality of these sales is still an open question, the disclosure couldn't be better timed to support the position of export-control hawks in the Pentagon."*.

Figure 25: Rhetorical Predicates that we Considered



5.3 Approaches to Rhetorical Predicate Identification

Once we have defined our inventory of predicates, sentences now need to be classified into these predicates to fill the right slots of a schema. As described in Section 5.2, a rhetorical predicate can be inter-clausal or intra-clausal. As inter-clausal rhetorical predicates and discourse relations described in various theories (e.g. RST) are comparable, to identify inter-clausal rhetorical predicates - e.g. *evidence*, we have used the discourse parser SPADE [SM03] which is a RST-based sentence level discourse parser. To the best of our knowledge, this is the only publicly available discourse parser. Another discourse parser is HILDA[1] (HIgh-Level Discourse Analyzer); however it only supports a web interface and no library or API is available for this parser. As a result, it was difficult to use it for our work.

Currently, there is no existing approach to identify intra-clausal rhetorical predicates. To devise approaches to identify intra-clausal rhetorical predicates, we have performed an evaluation to calculate how often each intra-clausal rhetorical predicate

[1]HILDA: http://nlp.prendingerlab.net/hilda

of our interest occurred in a corpus. In this evaluation, we have manually analyzed 200 sentences of comparison, illustration, attribution, topic-opinion, contingency, and attributive types. In this study, the comparison corpus was built from [JL06], the topic-opinion corpus from [FHW06], and the illustration, attribution, contingency, and attributive corpora from the BLOG06 dataset. The results are shown in Table 14.

Table 14: Frequency of Intra-Clausal Rhetorical Predicates

Rhetorical Predicates	**Frequency**
Comparison	66%
Contingency	0%
Illustration	5%
Attributive	83%
Attribution	7%
Topic-opinion	67%

From Table 14, we can see that comparison and topic-opinion predicates occur about 65% of the time within a single clause. This table also shows that most of the time (83%) attributive predicate occur within a clause. From the table, we can also see that illustration, attribution, and contingency predicates rarely occur within a clause. Based on these results, we decided to design approaches to identify intra-clausal attributive, topic-opinion, and comparison predicates. We have used a classifier adapted from [JL06] to identify comparison predicates, we have designed a classifier to identify topic-opinion predicates using [FHW06]'s idea that the dependency relations of words defined by a dependency grammar are useful to find relations between a topic and subjective words, and our own classifier is based on dependency relations to identify attributive predicates. The rest of the predicates appear so insignificantly within a single clause, that we did not consider them. Table 15 summarizes the predicates we have considered and the main approaches used. The next sections will describe these in more detail.

Table 15: Rhetorical Predicates Considered and Identification Approaches Used

Rhetorical Predicates	**Inter Clause**	**Approaches**	**Intra Clause**	**Approaches**
Comparison	✓	SPADE	✓	Jindal et al.'s
Contingency	✓	SPADE	X	
Illustration	✓	SPADE	X	
Attributive	X		✓	Own
Attribution	✓	SPADE	X	
Topic-opinion	X		✓	Own

5.3.1 Tagging Inter-Clausal Rhetorical Predicates

We use SPADE (Sentence-level PArsing for DiscoursE) [SM03] to identify the inter-clausal predicates comparison, contingency, illustration, and attribution and their subclasses which occur between two clauses.

The SPADE Parser

The SPADE parser was developed within the framework of RST (see Section 2.4.1). The SPADE parser identifies discourse relations within a sentence by first identifying elementary discourse units (EDU)s, then identifying discourse relations between two EDUs (clauses) by following the RST theory. For example, in the sentence below, the SPADE parser identifies two clauses:

a. *[Previously, airlines were limiting the programs]*

b. *[because they were becoming too expensive.]*

and assigns the relation *cause* between these two clauses.

The SPADE parser consists of two components: the *discourse segmenter* and the *discourse parser*. The *discourse segmenter* divides sentences into clauses. It uses two components for this purpose namely a *statistical model*, which assigns a probability to

the insertion of a discourse boundary after each word in the sentence, and a *segmenter* which finds the most likely positions for inserting discourse boundaries. Given a sentence, this model first finds the syntactic parse tree of the sentence. Then using both lexical and syntactic features of the parse tree it determines a probability of inserting a discourse boundary. Once the most likely discourse boundaries of a sentence are determined the *discourse parser* creates a discourse tree for the sentence. The *discourse parser* also consists of two components: a *parsing model*, which assigns a probability to every potential candidate parse tree, and the *discourse parser*, which finds the most likely discourse tree using dynamic programming. In this process, if more than one discourse relations are candidates to relate two clauses then the relation with the highest probability score (that is calculated based on their syntactic and lexical information from the training corpus) is selected.

In this approach, each sentence processed by the SPADE parser will be labeled with its most likely discourse relation. We use these relations to classify a sentence into the corresponding rhetorical predicate. This may result in tagging a sentence with no or with multiple rhetorical predicates. For example, the sentence "*Starbucks has contributed to the popularity of good tasting coffee*" does not contain any rhetorical predicate according to SPADE. 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 4 predicates according to SPADE: contrast, joint, attribution, and elaboration.

According to [SM03], the SPADE parser achieved an F-measure score of 49% to tag 18 discourse relations using the RST Discourse Treebank corpus. A performance, that, to our knowledge, has not been beaten by any other system.

5.3.2 Tagging Intra-Clausal Rhetorical Predicates

Unfortunately, the SPADE parser can only identify discourse structures across clauses, and cannot identify predicates which occur within a clause. For example, in *"Its fast-forward and rewind work much more smoothly and consistently than those of other models I've had."* a comparison relation is used, but would not be identified by SPADE. However, recall from Table 14, that comparisons, attributive, and topic-opinion do occur frequently within a clause (66%, 83%, and 67% respectively).

The discourse taggers which we have used to identify rhetorical predicates including comparison, topic-opinion, and attributive that occur within clauses are described in the next sections.

Comparison Classifier

In order to label a clause as containing a comparison predicate, we have adapted Jindal et al.'s approach [JL06]. This approach uses a keywords and patterns which are learned from annotated text.

To build the pattern (or sequence) database, the classifier first considers sentences which contain at least one predefined keyword (such as comparative adjectives). Then it creates a sequence using words which occur within a window of 3 words around the keyword. In the next step, these words are replaced with their part of speech (POS) tags and a class is associated with the sequence based on whether this sentence is a comparison or non-comparison sentence. For example, the sentence "this/DT camera/NN has/VBZ significantly/RB **more**/JJR noise/NN at/IN iso/NN 100/CD than/IN the/DT nikon/NN 4500/CD." contains the keyword "more" and the following sequence will be stored in the database:

({NN}{VBZ}{RB}{more/JJR}{NN}{IN}{NN}) *comparison*

After the database is constructed, class sequential rules (CSR) are generated. A CSR is a rule with a sequence on the left and a class label on the right of the rule. The CSR rules are generated by combining sequences which are available in the sequence database. As CSR, those rules are accepted which meet a pre-specified *support* and *confidence* threshold value. The support and confidence of a rule are defined as follows:

$$Support\ of\ a\ rule = \frac{\#\ of\ instances\ containing\ this\ rule}{\#\ of\ instances\ in\ the\ sequence\ database}$$

$$Confidence\ of\ a\ rule = \frac{\#\ of\ instances\ containing\ this\ rule\ in\ this\ class}{\#\ of\ instances\ in\ the\ sequence\ database\ satisfying\ the\ rule}$$

A Naïve Bayes classifier is used with the CSR patterns as features to learn a 2-class classifier (comparison and non-comparison). This classifier achieved an F-measure score of 79%. Unfortunately, [JL06]'s comparison classifier is not publicly available. As a result, we have implemented it ourself using subset of their annotated dataset (see Section 7.4.1).

Topic-Opinion Classifier

The topic-opinion predicate indicates whether a sentence expresses an opinion towards a specific topic. It is useful to answer questions such as *"Do people like X?"*. To our knowledge, no parser is available to tag topic-opinion predicates. Fei et al. [FHW06] showed that the dependency relations of words defined by a dependency grammar are useful to find relations between a topic and subjective words. In light of this, we have adapted their approach to build our topic-opinion classifier.

Dependency relations of words are defined based on dependency grammars [dMM08]. They refer to the binary relations between two words where one word is the parent (or

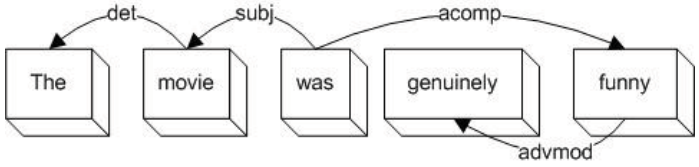
head) and the other word is the child (or modifier). In this representation, one word can be associated with only one parent but with many children. Therefore, when the dependency relations of a sentence is created it will be in the form of a tree (called a dependency tree [FHW06]). Typical dependency relations are showed in Table 16.

Table 16: Sample Dependency Relations between Words (taken from [FHW06])

Relation Name	**Description**	**Examples**	**Parent**	**Child**
subj	subject	I will go	go	I
obj	object	tell her	tell	her
mod	modifier (e.g. adj, adv, ...)	a nice story	story	nice

Dependency relations are useful to find relations (links) between subjective words and a topic. Different words of a sentence can be related using dependency relations directly or based on the transitivity of these relations. For example, the dependency relations of the sentence *"The movie was genuinely funny."* is shown in Figure 26.

Figure 26: Dependency Relations for the Sentence: *The movie was genuinely funny.*



The head of the arrow points to the child, the tail comes from the parent, and the tag on the arrow indicates the dependency relation type. For example, in Figure 26, both words *movie* and *funny* are modifiers of the word *was*. The word *movie* is the subject of the word *was* and the word *funny* is a direct adjectival complement (acomp) to the word *was*. With the help of dependency relations it is possible to find that the topic *movie* and the subjective word *funny* are related.

Recall from Table 14 that in an experiment we have found that 67% of time topic-opinion predicate occur within a single clause. As a result, our topic-opinion

predicate identification approach is based on the analysis of single clause. To develop our topic-opinion tagger, we have first manually selected 200 topic-opinion sentences from the BLOG06 corpus then parsed them using the Stanford parser[2]. Figure 27 shows three sentences from our development set.

Figure 27: Sample sentences from the Topic-opinion Dataset

Topic: Film

Sentence: Alright , what else was bad in this film ?

Topic: Movie

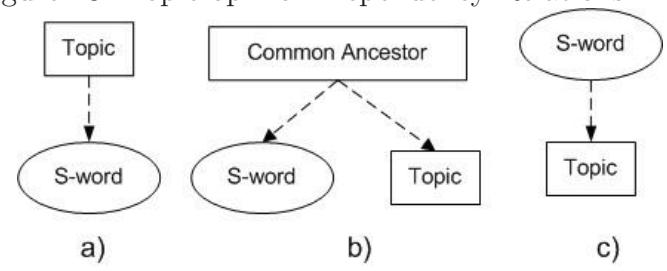
Sentence: I mean, showing melissa sagemiller running away from visions for about 20 minutes throughout the movie is just plain lazy!

Topic: Actors

Sentence: The actors are pretty good for the most part, although wes bentley just seemed to be playing the exact same character that he did in american beauty, only in a new neighborhood.

By manually analyzing the parse trees of these 200 topic-opinion sentences and using the work of Fei et al. [FHW06], we have found that 3 types of relations are typically used to indicate topic-opinion relations. These are shown in Figure 28.

Figure 28: Topic-opinion Dependency Relations Trees

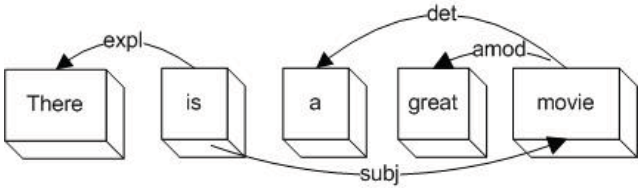


Heuristic 1: Subjective Words that Modify the Topic: Subjective words (S-word) that are in a modifier relation with the topic directly or based on transitivity relations are good indicators of a topic-opinion predicate. This is shown

[2]http://nlp.stanford.edu/software/lex-parser.shtml

in Figure 28 a). For example, in the sentence "*There is a great movie.*" the

Figure 29: Example of Topic-opinion Heuristic 1



subjective word *great* modifies the topic *movie* (shown in Figure 29). This is the most frequently encountered dependency relation in our topic-opinion development set and accounts for 45% of the development set.

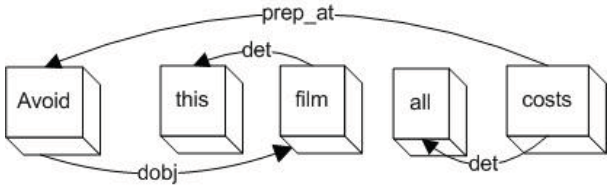
Heuristic 2: Subjective Words and the Topic that have the Common Ancestor: In this case, shown in Figure 28 b), [FHW06] accept instances where the same ancestor is the verb, but in our analysis we have found this heuristic to be too lenient. By analyzing our corpus of 200 topic-opinion sentences, we have restricted the dependency relations to link only:

- the topic and the ancestor verb
- the subjective word and the ancestor verb

For example, the topic could be related to the ancestor verb directly using the dependency relation *subj*. An example of this heuristic is shown in Figure 31. These dependency relations account for 34% of the development set.

Heuristic 3: Subjective Words that are Ancestors of the Topic: To classify in this category, according to [FHW06], the subjective word needs to be a verb, and the topic needs to be the subject or object of the verbs. For example, the sentence "*Avoid this movie at all cost.*" is an example of heuristic 3 (shown in Figure 30).

Figure 30: Example of Topic-opinion Heuristic 3



In this sentence, topic *film* is the object of the verb *avoid* which is a subjective word. For this type, we further constrained which set of dependency relations will be accepted as transitivity relations when the topic and subjective words are not directly connected. These relations account for 12% of the development set.

Table 17 shows the heuristics occurrence distribution in our development set. The table shows that 9% of the distribution dataset was not tagged by any of these heuristics because our dictionary-based approach was unable to find subjective words in those cases.

Table 17: Topic-opinion Heuristic Occurrence Distribution

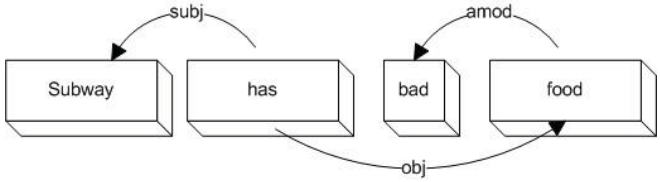
Heuristics	**Distribution**
Heuristic 1	45%
Heuristic 2	34%
Heuristic 3	12%
Total	91%

Our topic-opinion classifier works in two steps: first it identifies whether the sentence is opinion-bearing. To do that it uses a dictionary-based approach using the MPQA subjectivity lexicon[3](see Section 6.1.2 for the lexicon details). If the sentence contains any subjective word from the dictionary, it considers it as an opinion-bearing sentence. Once the opinionated sentence is found, the dependency classifier identifies whether the topic of the sentence is associated with any of the subjective word of the sentence using dependency relations. For example, for the sentence "*Subway has bad*

[3]available at http://www.cs.pitt.edu/mpqa/

food.", our classifier will first identify the word *bad* as a subjective word so it will recognize this sentence as an opinion bearing sentence.

Figure 31: Example of Topic-opinion Dependency Relations Tree



Then the classifier will find that the topic *Subway* and the subjective word *bad* are linked based on transitivity (see Figure 31). Using heuristic 2 above, this sentence will therefore be tagged as a topic-opinion predicate bearing sentence. The topic of a sentence is manually annotated in the dataset as shown in Figure 27.

Section 7.4 will describe the evaluation of these heuristics to tag topic opinion sentences.

Our Attributive Tagger

As mentioned in Section 5.2, an attributive predicate provides details about an entity or an event - e.g. "*Mary has a pink coat.*" In this example, the sentence contains an attributive predicate because it provides details about the entity *coat*. Attributive predicates can also be used to illustrate a particular feature about a concept or an entity - e.g. " *iPad 2. will support full touchscreen HD display with a screen resolution of 2048 x 1536.*" The sentence of this example also contains an attributive predicate since it is describing a particular feature of the entity *iPad 2*. Even though attributive predicates are often used in query-based summarization (e.g. [MK10]) and question answering systems (e.g. [McK85, Par85]), to our knowledge, no previous work has focused on tagging attributive predicates automatically. We therefore propose an automatic domain and genre-independent approach to tag attributive predicates by utilizing dependency relations of words based on dependency grammars [dMM08].

Similarly to our topic-opinion tagger, to develop our method, we have first created a development set containing 200 attributive sentences by tagging them manually from the BLOG06 corpus. Figure 32 shows three sentences from our development set. A first analysis of the development set showed that 83% of the time, attributive

Figure 32: Sample Sentences from the Attributive Dataset

Topic: Jiffy Lube

Sentence: Jiffy Lube promotes the short , 3,000 mile oil drain interval.

Topic: Zillow

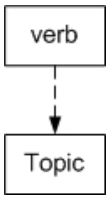
Sentence: Zillow priced the land 30K higher than what I thought it was worth...

Topic: Picasa

Sentence: Picasa has this bizzare setup where instead of editing the image itself, it edits a copy of it someplace else.

relations occur within a clause (see Table 14 in Section 5.3); as opposed to many other discourse relations that span across clauses. Due to this, our approach is based on the analysis of single clauses. To identify attributive predicates automatically, we have used dependency relations of words based on dependency grammars [dMM08].

Figure 33: Attributive Dependency Relations Tree

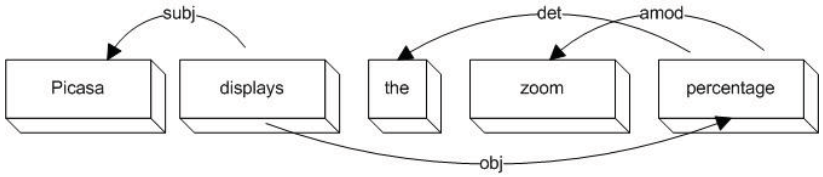


In order to develop our classifier, we have first parsed the sentences of our development set using the Stanford parser. A manual analysis of these parses showed that to be classified as an attributive sentence, the topic of the sentence needs to be the descendant of a verb (shown in Figure 33) and be in a subject or object relation with

it. However, the topic and the verb can be related in several ways; which we describe by 3 heuristic rules:

Heuristic 1: The Topic is a Direct Nominal Subject: The *topic* is a direct nominal subject, a noun phrase that is the syntactic subject of the *verb* (e.g., `subj` in the Stanford parser).

Figure 34: Example of Heuristic 1 to Tag the Attributive Predicate



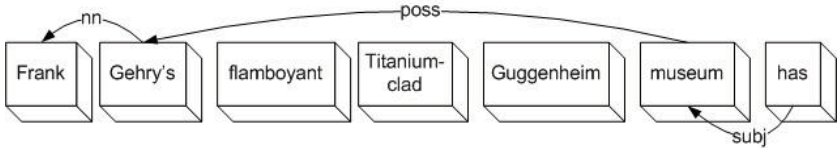
For example, the sentence *"Picasa displays the zoom percentage"* contains an attributive relation where the topic "*Picasa*" is directly related to the verb "*displays*" using the dependency relation `subj` (shown in Figure 34). This is the most frequently encountered dependency relation in our attributive development set and accounts for 42% of the development set.

Heuristic 2: A Noun is the Syntactic Subject and the Topic is a Modifier of the Noun: A noun is the syntactic subject of the *verb* and the *topic* is a modifier of the noun. Under this heuristic rule, a modifier can be a noun compound modifier (e.g., `nn` in the Stanford parser), a propositional modifier (e.g., `prep` in the Stanford parser) or a possession modifier (e.g., `poss` in the Stanford parser).

For example, the sentence *"Frank Gehry's flamboyant, titanium-clad Guggenheim Museum has a similar relationship to the old, masonry city around it."* contains an attributive relation where the noun "*Museum*" is the subject of the verb "*has*" and the topic "*Frank Gehry*" is a possession modifier of the noun

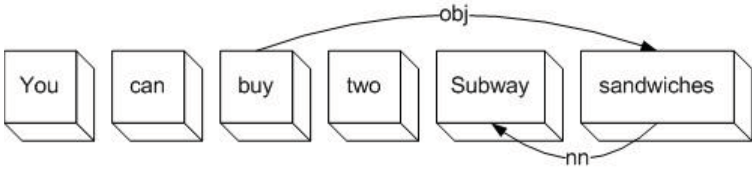
"*Museum*" (a partial dependency tree is shown in Figure 35). These dependency relations account for 38% of the development set.

Figure 35: Example of Heuristic 2 to Tag the Attributive Predicate



Heuristic 3: A Noun is the Syntactic Direct Object and the Topic is a Modifier of the Noun: A noun is the syntactic direct object of the *verb* (e.g., `obj` in the Stanford parser) and the *topic* is a modifier of the noun. Under this heuristic rule, a modifier can be a noun compound modifier (e.g., `nn` in the Stanford parser).

Figure 36: Example of Heuristic 3 to Tag the Attributive Predicate



For example, the sentence *"You can buy two Subway sandwiches for $7.99 on sunday."* contains an attributive relation where the noun "*sandwiches*" is the object of the verb "*has*" and the *topic* "*Subway*" is a modifier of the noun '*sandwiches*" (a partial dependency tree is shown in Figure 36). These relations account for 16% of the development set.

Table 18 shows the heuristics occurrence distribution in our development set. The table shows that 4% of the development dataset was not tagged by any of these heuristics that was due to the parser errors.

Table 18: Attributive Heuristic Occurrence Distribution

Heuristics	**Distribution**
Heuristic 1	42%
Heuristic 2	38%
Heuristic 3	16%
Total	96%

This chapter has presented the approaches we have taken to tag rhetorical predicates. Chapter 7 includes a full evaluation of these and the performance of all rhetorical predication taggers including the SPADE parser, the comparison classifier, the topic-opinion tagger, and the attributive tagger. We have also calculated a baseline and human performance to identify various predicates and have used them to compare the performance of the predicate identification approaches. The experimental setup and the evaluation results are discussed in Section 7.4.

5.4 Conclusion

In this chapter, we have identified a set of intra-sentential rhetorical predicates which can be expressed in texts and have analyzed domain and genre-independent automatic approaches to identify these rhetorical predicates. As much as possible, we tried to use off-the-shelf approaches which have been developed for discourse analysis or for other purposes to identify intra-sentential rhetorical predicates. However, to identify the attributive predicate and topic-opinion predicate, we have introduced our own automatic approaches based on dependency relations. Table 15 summarizes the rhetorical predicates considered and the approaches used.

The next chapter will describe our prototype system called BlogSum which we have developed to validate our summarization approach, while Chapter 7 will evaluate the system including the predicate tagging approaches presented in this chapter.

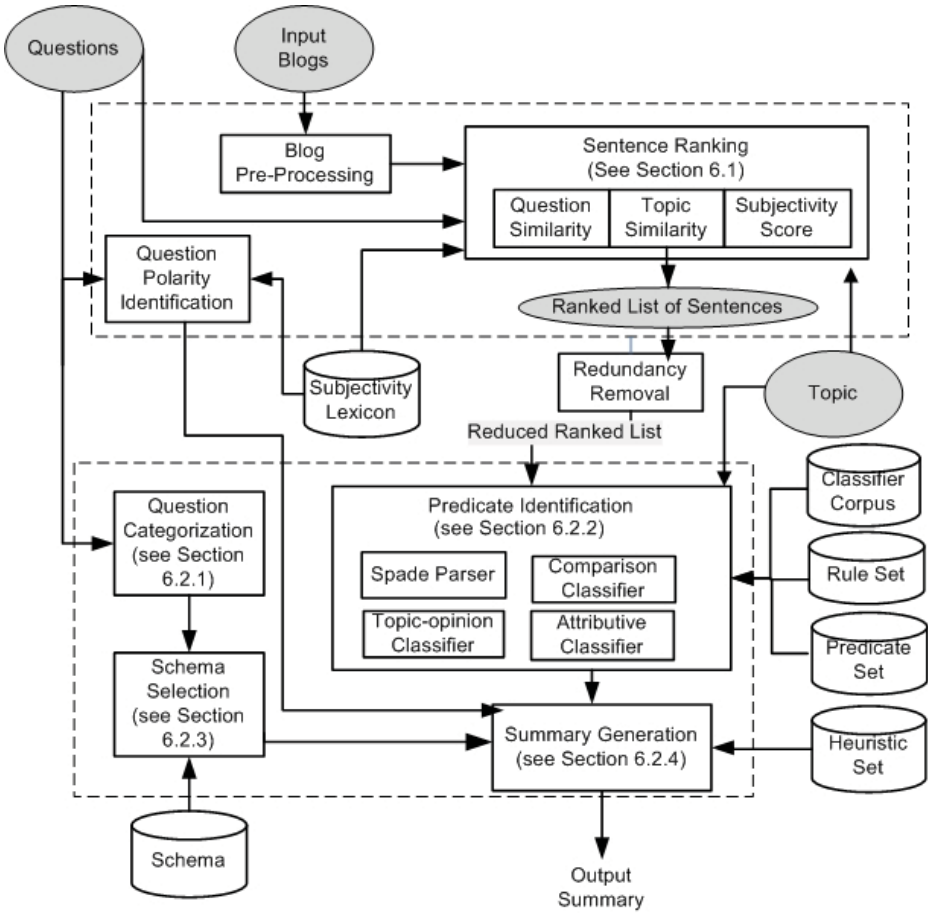
Chapter 6

BlogSum: Our Prototype Blog Summarizer

In order to evaluate the general model described in Chapter 4, and the predicate identification approaches of Chapter 5, we have developed a prototype system called BlogSum. In this chapter, we will first briefly describe how our approach selects candidate sentences from the document collections then we will describe implementation details of how our approach filters question irrelevant sentences and reorders candidate sentences using schemata to create the final summary

Figure 37 shows the detailed architecture of BlogSum. The figure implements the approach presented in Figure 14 of Section 4.1.2. It shows that given an initial topic and question and a set of related blogs, BlogSum first creates a ranked list of sentences that could potentially be included in the final summary. To create this ranked list of sentences, BlogSum performs pre-processing such as filtering textual content from other non-textual elements such as html tags (see Section 6.1.1) and then creates a preliminary candidate list using question similarity, topic similarity and subjectivity scores (see Section 6.1.2). In the next step, BlogSum removes redundant sentences from the candidate list to address content overlap errors using the cosine similarity

Figure 37: Detailed Architecture of BlogSum



(see Section 2.2.4). To remove redundant sentences, the cosine value is calculated for each pair of sentences. Before inserting a sentence into the list of candidate sentences, it is checked for similarity with the sentences already in the list. If the sentence is similar to any of the sentence in the list then it is not inserted. Through this process, candidate sentences are checked for redundancy. Then BlogSum generates summaries by categorizing the initial question (see Section 6.2.1), identifying the rhetorical predicates that each candidate sentence conveys (see Section 6.2.2), selecting the appropriate schema (see Section 6.2.3), and generating the summary (see Section 6.2.4). In Figure 37, the dotted box in the upper section shows the steps involved in candidate sentence selection while the dotted box in the lower section

shows the processes involved in summary generation. BlogSum is implemented using java and using third-party tools such as the Stanford parser[1], the SPADE parser[2], WordNet lemmatizer[3], and uses the Weka toolbox[4].

The next sections will describe the implementation of these steps.

6.1 Candidate Sentence Selection

In order to extract the initial candidate sentences from the original blogs, BlogSum needs to perform some pre-processing on the blogs to retrieve their textual content.

6.1.1 Blog Pre-processing

The blogs distributed in the BLOG06 corpus contain many non-textual elements such as html tags, scripting codes, and unwanted links (e.g. image link). Figure 38 shows a partial original input blog from BLOG06. The input blog contains many html tags such as , <a>,
. The input blog also contains JavaScript codes written inside the tag <script>. All these tags and codes are not part of the textual content. BlogSum removes these tags and codes to retrieve the main text. To remove these unrelated contents, BlogSum uses rule-based patterns and regular expressions. These patterns were designed by manually analyzing 50 input blogs from BLOG06. For example, to remove all text within the <script> tag, we use the pattern: *<script.*?</script> replace by nothing.* All texts inside the <script> tag are not part of the textual content; they are JavaScript codes that should be deleted. Figure 38 also shows the cleaned blog text extracted by BlogSum for the input blog. Once the textual content is extracted, BlogSum can extract candidate sentences from the cleaned blogs.

[1]The Stanford parser: http://nlp.stanford.edu/software/lex-parser.shtml
[2]The SPADE parser: http://www.isi.edu/licensed-sw/spade
[3]WordNet: http://wordnet.princeton.edu
[4]Weka: http://www.cs.waikato.ac.nz/ml/weka/

Figure 38: Blog Pre-processing

Input Blog:

Actually, I am still learning and have serious problems with many of these arguments, but I continue to investigate them because they fascinate me. I, against all logic, want this world-view to be true. It appeals to me on a gut level.

So there is doubt. There is disbelief. But I still like it.

On a policy level, I (clearly) have my leanings but am willing to discuss, argue, and have my mind changed. No one has ever called me dogmatic.

So when I was directed to this article in today's WP, I was shocked. Shocked that a sitting President has managed to promote an economic fallacy of such Freshman/Economic 101 proportions that it defies explanation.
 ...

Posted by: Carl / 1:50 PM

<script type="text/javascript">postCount('113821713738842982'); </script> | <script type="text/javascript">postCountTB('113821713738842982'); </script>

Cleaned Blog:

Actually, I am still learning and have serious problems with many of these arguments, but I continue to investigate them because they fascinate me. I, against all logic, want this world-view to be true. It appeals to me on a gut level.

So there is doubt. There is disbelief. But I still like it.

On a policy level, I (clearly) have my leanings but am willing to discuss, argue, and have my mind changed. No one has ever called me dogmatic.

So when I was directed to this article in today's WP, I was shocked. Shocked that a sitting President has managed to promote an economic fallacy of such Freshman/Economic 101 proportions that it defies explanation.

6.1.2 Candidate Sentence Selection

In order to have a complete system, in BlogSum, we have developed our own sentence extractor to retrieve the initial list of candidate sentences based on question similarity, topic similarity, and subjectivity scores. However, any other sentence ranker could have been used (such as MEAD [RABG$^+$04]). Later in this section, we will describe an experiment comparing our approach with that of MEAD.

In our current candidate sentence selection approach to calculate the score of a sentence, we have considered question similarity and topic similarity to give priority to topic and question relevant sentences. Since we are interested in query-based summarization, we gave priority to topic and question relevant sentences. To calculate the score of a sentence, we have also considered their subjectivity scores because blogs are subjective in nature and the questions we are dealing with are also mostly subjective - e.g. "*Why do people like x?*".

To rank sentences, BlogSum calculates a score for each sentence using the following features shown in Equation 1 (Eq1):

$$Sentence\ Score = Question\ Similarity + Topic\ Similarity + |SubjectivityScore| \quad \text{(Eq1)}$$

Let us see how each feature is computed.

Question Similarity and Topic Similarity

To compute the similarity between two sentences, different approaches are available. To choose the best approach for our application, we first conducted an experiment to compare the performance of various similarity calculation approaches which are commonly used. We considered 5 measures: cosine *tf.idf* uni-gram, cosine *tf.idf* bi-gram, cosine *tf.idf* bi-gram skip 4, word overlap, and *idf* overlap. The cosine *tf.idf* bi-gram measure is based on pairs of juxtaposed lemmas; the cosine *tf.idf* bi-gram skip 4 measure evaluates the similarity based on pairs of lemmas, but the notion of pair is more flexible; word overlap is defined as the proportion of words that appear in both sentences normalized by the sentence length; *idf* overlap is defined as the proportion of words that appear in both sentences weighted by their inverse document frequency (*idf*).

To compare these sentence similarity measures, we have used the data from TAC 2008 opinion summarization track. The data set consists of 50 questions on 28 topics; on each topic one or two questions are asked and 9 to 39 relevant documents are given. In this experiment, we used the ROUGE metric, using answer nuggets (provided by TAC), which had been created to evaluate participants' summaries at TAC, as gold standard summaries. For this evaluation, using each similarity approach, a list of 15 sentences was produced for each question and compared against the gold standard summaries using F-scores of ROUGE-2 and ROUGE-SU4 (see Section 2.5).

Table 19: Comparison of Various Similarity Measures

Similarity Measure	**ROUGE-2 (F-Measure)**	**ROUGE-SU4 (F-Measure)**
Cosine uni-gram	0.080	0.118
Cosine bi-gram	0.073	0.112
Cosine bi-gram Skip 4	0.068	0.110
Word overlap	0.082	0.119
idf overlap	0.002	0.014

From Table 19, we can see that all approaches except idf overlap perform similarly using both ROUGE-2 and ROUGE-SU4. Actually word overlap gives the best result, although the difference may not be significant. In BlogSum, we used the cosine uni-gram because it is the commonly used similarity metric and it gave good results in similarity measure evaluation.

To calculate similarity between a sentence and the question, we used the cosine similarity using word (lemma) uni-gram matching above a predefined threshold value. We have experimentally set the threshold value 0.2 because we have achieved the best results with the threshold value of 0.2 for the TAC 2008 sample dataset. To calculate similarity, sentences and questions are represented as a weighted word vector based on *tf.idf* (for sentences) and *tf* (for questions). The similarity between a sentence and

the topic is calculated as with the question similarity using the words in the topic instead of the question. The weight of a word in the similarity measure is weighted by its *tf.idf* value in the document set.

Subjectivity Score

To keep the sentiment analysis process simple, BlogSum uses a dictionary-based approach to calculate the subjectivity score instead of using machine learning. However, it is possible to use any approach such as the combined sentiment analysis approach developed in [And09] to identify sentence-level sentiment to use the benefit of lexical-based and corpus-based approaches for sentiment analysis. To calculate the subjectivity scores, BlogSum uses the MPQA subjectivity lexicon[5], which contains more than 8000 entries of polarity words. This dictionary is commonly used in this area and is the basis for the work of many, such as [MJCN08, Sek08]. In the lexicon, for each subjective word, the prior polarity and subjectivity strength (*weak*, *strong*) are provided. Four types of prior polarity values are used namely, *positive*, *negative*, *both*, and *neutral*. Table 20 shows the prior polarity and subjectivity types of a few words from the MPQA lexicon.

Table 20: Examples of Word Polarity and Subjectivity in the MPQA Lexicon

Word	**Subjectivity Type**	**Prior Polarity**
Congratulation	Strong	Positive
Ability	Weak	Positive
Hate	Strong	Negative
Complication	Weak	Negative
Covet	Strong	Both
Eat	Not Applicable	Neutral

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

To assign a polarity to a word in a sentence, we used the polarity value *positive*, *negative*, *both* or *neutral* and assigned the score 1, -1, 0.25, and 0, respectively. Moreover, if a word is tagged as weakly subjective then we reduce the subjectivity strength by a factor (0.25 in our current prototype); on the other hand, if a word is tagged as strongly subjective then we increase the subjectivity strength by a factor (0.25 again). To calculate polarity, we also considered a predefined set of valence shifters such as *not*, *rarely* which occur in a window of size 3 on both sides of a subjective word. These valence shifters reverse the polarity class of the subjective word. The subjectivity score of a sentence is then calculated based on the match of the sentence words with the subjective words listed in the subjectivity lexicon. The subjectivity score of a sentence is calculated in the following manner:

$$Subjectivity\ score\ of\ a\ sentence = \frac{sum\ of\ the\ polarity\ score\ of\ all\ sujective\ words\ found\ in\ the\ sentence}{\#\ of\ subjective\ words\ in\ the\ sentence}$$

For example, the sentence "*I love SECOND CUP, because I love the convenience of a good cup of coffee almost anywhere I am.*" contains 4 subjective words: *love* (twice), *good*, and *convenience*. The prior polarity of all these words is positive (1+1+1+1) and the subjectivity type of the word *love* is strong while other two words are weak (.25+.25-.25-.25). Therefore, the subjectivity score of the sentence is calculated as 1.

$$Subjectivity\ score\ of\ the\ sentence = \frac{1+1+1+1+.25+.25-.25-.25}{4} = 1$$

The sentence will be considered as a positive sentence. This approach is similar to [KC08].

Four types of polarity values including *positive*, *negative*, *both*, and *neutral* are used to classify a sentence. The subjectivity score of a sentence is used to determine

its polarity class in the following manner:

- **Positive**: If subjectivity score ≥ 0.5
- **Negative**: If subjectivity score $\leq$ -0.5
- **Both**: If subjectivity score < 0.5 to $>$ -0.5
- **Neutral**: If subjectivity score $= 0$ OR no subjective word

The polarity of all sentences in the candidate list is identified and this information is used by the summary generation module during sentence selection. In general, the polarity of a sentence needs to be identical with the polarity of the question to be considered as a summary sentence. For example, if the polarity of the question is positive then the polarity of a sentence also needs to be positive to be considered as a summary sentence. However, to answer a positive question, we also accept sentences with neutral polarity. For example, the question "*What features do people like about iPhone 4S?*" contains only one subjective word (*like*) and the prior polarity of this word is positive. As a result, this question will be considered as a positive question and according to our schema, positive questions will be answered only using positive or neutral sentences.

Table 21: Sentence Polarity Corresponding to Question Polarity

Question Polarity	**Sentence Polarity**
Positive	Positive or Neutral
Negative	Negative
Both	Positive, Negative, Both or Neutral
Neutral	Positive, Negative, Both or Neutral

On the other hand, if the polarity of the question is negative then the polarity of a sentence also needs to be negative to be considered as a summary sentence. However,

to answer a question containing polarity class both or to answer a neutral question, we accept sentences with any type of polarity. This is shown in Table 21.

The polarity of a question is calculated the same way as the polarity of a sentence. However, the subjectivity score of a question is only used to identify its polarity class but the subjectivity score of a sentence is used to identify its polarity class as well as calculating its rank.

We have also evaluated the accuracy of our polarity identification approach. For this experiment, we used a set of about 1200 product review sentences extracted from the annotated corpus[6] made available by Bing Liu [HL04]. This dataset contains 403 positive sentences, 403 negative sentences, and 403 neutral sentences. The accuracy of our polarity identification approach using this dataset is shown in Table 22. From this evaluation, we have found that in many cases, our approach tagged positive or negative sentences as neutral because of missing words in the MPQA lexicon to find the appropriate polarity class. However, this result is satisfactory against the baseline established for the overall accuracy by [AB08] for this dataset which was 59.3% for the lexicon-based approach and 60.7% for the supervised approach.

Table 22: Accuracy of Our Polarity Identification Approach

Polarity Class	**Accuracy**
Positive	73%
Negative	61%
Neutral	66%
Overall	67%

Figure 39 shows a partial candidate list to illustrate the output produced by the candidate sentence selection phase (also shown in Chapter 4). Given the Topic: "*Carmax*" and the Question: "*What motivated positive opinions of Carmax from car buyers?*". BlogSum extracted the 8 most relevant sentences are shown in Figure 39

[6]http://www.cs.uic.edu/ liub/FBS/sentiment-analysis.html

along with their score (out of 1) calculated based on Eq1.

Figure 39: Partial Candidate List Used as Input

Topic: *Carmax* **Question**: *What motivated positive opinions of Carmax from car buyers?*	
Candidate Sentences	**Score**
(1) With Carmax you will generally always pay more than from going to a good used car dealer.	0.536
(2) We bought it at Carmax, and I continue to have nothing bad to say about that company.	0.449
(3) Carmax did split the bill (which made me happy).	0.416
(4) Not sure if you have a Carmax near you, but I've had 2 good buying experiences from them.	0.381
(5) have to say that Carmax rocks.	0.368
(6) At Carmax, the price is the price and when you want a car you go get one.	0.299
(7) Sometimes I wonder why all businesses can't be like Carmax.	0.278
(8) Arthur Smith, 36, has been living in a van outside the Carmax lot, 24 hours a day, for more than a month.	0.25

To validate our initial candidate list, we compared it to MEAD [RABG+04]. We have conducted an experiment to verify whether MEAD-generated summaries are better than our candidate list (called OList). 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 the cosine similarity based on the *tf-idf* value. In this evaluation, we used the TAC 2008 opinion summarization dataset and summaries were evaluated using the ROUGE-2 and ROUGE-SU4 scores. Table 23 shows the results of the automatic evaluation using ROUGE based on summary content.

Table 23 shows that MEAD-generated summaries achieved weaker ROUGE scores

Table 23: Automatic Evaluation of MEAD based on Summary Content

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

compared to that of our candidate list (OList). The table also shows that MEAD performs weaker than the average performance of the participants of TAC 2008. We suspect that the poor results of MEAD are due to two issues 1) in MEAD, we cannot use opinionated terms or polarity information as a sentence selection feature. On the other hand, most of the summarizers, which deal with opinionated texts, use opinionated terms and polarity information for this purpose; 2) we have found that in this experiment, for some of the TAC 2008 questions, MEAD was unable to create any summary. This evaluation results justified the use of our own candidate sentence selector.

6.2 Content Filtering and Organization

Once we have our ranked list of candidate sentences, we need to select which sentences will be part of the summary and where, and which will be discarded.

As discussed in Section 4.1.2, for content filtering and organization, our approach performs the following tasks:

A. Question Categorization

B. Predicate Identification

C. Schema Selection from the Pre-designed Schemata

D. Summary Generation

Section 4.1.2 explained the purpose of these modules and the methodology used to develop them. The next sections will detail how they have been implemented in a working system.

6.2.1 Question Categorization

As described in Section 4.1.2, 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. Recall from Section 4.1.2 that we have identified 3 categories of questions based on their communicative goals, namely: *comparison*, *suggestion*, and *reason*.

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. [JL06]). 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 (randomly selected 50 comparison questions) by [JL06]. 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. Datasets were used to create question patterns are shown in Table 24.

Table 24: Datasets used to Question Pattern Analysis

Question Type	**Development Set**		**Test Set**	
	Source	Size	Source	Size
Reason	TAC 2008 Development Dataset	16 Questions	TAC 2008 Opinion Summarization Dataset	45 Questions
Comparison	Jindal et al.'s Dataset	50 Questions	Jindal et al.'s Dataset	100 Questions
Suggestion	TAC 2008 Opinion Dataset	3 Questions	X	X

By analyzing our development question set, we have designed 4 patterns for the *comparison* question, 4 patterns for the *suggestion* question, and 6 patterns for the *reason* question.

To analyze part of speech tags in order to design patterns for question categories, we have used the Stanford POS tagger[7] which uses the Penn Treebank tag set[8]. 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.

For example, Figure 40 shows a sample question on the target "Carmax" and provides ids (doc id) of the input documents. In this sample, we used the text of the target which is "Carmax" as the topic.

The patterns as well as their accuracy for a particular question category are shown in Table 25. We have calculated the accuracy of 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% (shown in Table 25). Recall from Table 10 of Section 4.1.2, that the TAC 2008 opinion summarization track contains only 4% of *comparison* questions and 6% of *suggestion* questions.

[7]http://nlp.stanford.edu/software/tagger.shtml

[8]http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQP-HTMLDemo/PennTreebankTS.html

Figure 40: Sample from the TAC 2008 opinion summarization Dataset

```
<target id = "1001" text = "Carmax">

        <q id = "1001.2" type= "SquishyList">
           What motivated negative opinions regarding purchasing a car from CARMAX?
        </q>

        <q id = "1001.3" type= "SquishyList">
           What motivated positive opinions of CARMAX from car buyers?
        </q>

        <doc id = "BLOG06-20060115-018-0013919728" />
        <doc id = "BLOG06-20051227-047-0021802750" />
        <doc id = "BLOG06-20051210-053-0002549820" />
        <doc id = "BLOG06-20051227-047-0021813992" />
        <doc id = "BLOG06-20060122-013-0025101599" />
        <doc id = "BLOG06-20060105-006-0012467028" />
        <doc id = "BLOG06-20051225-065-0006640208" />
        <doc id = "BLOG06-20060120-021-0032989196" />
        <doc id = "BLOG06-20060120-019-0002130175" />
        <doc id = "BLOG06-20060113-014-0001529470" />

</target>
```

Thus, 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 [JL06]'s dataset and achieved an accuracy of 97% (shown in Table 25). Datasets were used to calculate the accuracy of the question patterns are shown in Table 24. Due to the lack of data, we could not evaluate patterns of the *suggestion* question type. However, the results of Chapter 7 with the review dataset show that all 3 question types perform well.

Table 25: Lexico-syntactic Patterns for Question Categorization

Patterns	**Accuracy**
Comparison: Pattern 1: [...]NP VB(opinionated terms) NP RBR(comparison terms) PP(containing topics) Pattern 2: Compare NP(containing all topics) Pattern 3: Compare CC Contrast NP(containing all topics) Pattern 4: What NNS VBP VBN IN VBG(containing topic) NNS RBR(comparison terms) IN NNP Example: Why do people like Starbucks better than Dunkin Donuts?	97%
Suggestion: Pattern 1: What NNS VBP suggested/advised [...] Pattern 2: What VBP NP VP(containing topic) Pattern 3: What VBP NP suggest/recommend/advice PP(containing topic) Pattern 4: What suggestions/recommendation/advices VP(containing topic) Example: What steps are being suggested to correct this problem?	N/A
Reason: Pattern 1: What reasons [...] VB(opinionated terms) NP(containing topic) Pattern 2: Why do/don't NNS VB(opinionated terms) NP(containing topic) Pattern 3: What NNS VBP NNS VB(opinionated terms) PP(containing topic) Pattern 4: What NNS(opinionated terms) VP(containing topic) Pattern 5: What NP(containing topic) VB(opinionated terms) Pattern 6: What motivated positive/negative opinions VP(containing topic) / PP(containing topic) [optional PP] Example: Why do people like Picasa?	96%

where, [...] refers to any lexico-syntactic pattern and *NP*, *RBR*, *NNS*, ... refer to the parts of speech categories (noun, adverb, ...) using the Penn Treebank tag set.

6.2.2 Predicate Identification

In our schema-based approach, sentences need to be classified and organized based on what rhetorical predicates they convey. Candidate sentences are therefore classified as containing none or any of the 6 predefined rhetorical predicates presented in

Section 5.2 to fill the various slots of the matched schema. Our approach to predicate identification was detailed in Chapter 5; in this section, we will explain how our approach was implemented within BlogSum.

As specified in Section 5.2, predicates can describe a single clause or a relation between clauses. In order to identify predicates within a single clause - e.g. attributive, we have used the three classifiers presented in Section 5.3.2: the comparison classifier, the topic-opinion classifier, and our attributive tagger. The comparison classifier is used to identify intra-clause comparison predicate; the topic-opinion classifier is used to identify topic-opinion predicates and we proposed an approach to tag attributive predicates. In our prototype, to identify predicates between clauses - e.g. *evidence*, we have used the SPADE discourse parser [SM03].

As discussed in Section 5.3.2, we have adapted Jindal et al.'s approach [JL06] to develop the comparison classifier. This classifier is developed by using a set of keywords and patterns which are learned from annotated text using the dataset from [JL06]. This 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 from different sources. We have randomly selected 1500 sentences for test and the rest of the dataset was used for training where the training and the testing set were mutually exclusive. To build this classifier, we have extracted 130 patterns which are used as features to train a 2-way Naïve Bayes classifier. To do this, we have used the Weka toolbox[9]. Given a sentence, the comparison classifier labels it either comparison or non-comparison.

Recall from Section 5.3.2 that to design the topic-opinion classifier, we used word dependency relations. By manually analyzing dependency relations of 200 topic-opinion sentences from the BLOG06 corpus generated using the Stanford parser,

[9]http://www.cs.waikato.ac.nz/ml/weka/

we devised 3 heuristic rules to design this classifier (details in Section 5.3.2). This classifier first checks whether a sentence is opinion-bearing based on a dictionary-based approach using the MPQA subjectivity lexicon. If the sentence contains any of the subjective words from the lexicon, the sentence is considered as opinion-bearing. If the sentence is opinionated then the classifier checks whether any of the dependency-based heuristic rule applies. If so then the sentence is tagged as a topic-opinion sentence. To check if a heuristic applies, the classifier uses word dependency relations from the Stanford parser, the polarity information from the MPQA lexicon, and the topic terms which were manually annotated in the (development and test) dataset.

To build our attributive classifier, we used a similar approach as for the topic-opinion classifier. We used the Stanford parser to identify dependency relations. Here again, given a sentence, this rule-based classifier tries to determine if any of the 3 dependency relations shown in Section 5.3.2 match. If this is the case, it tags the sentence as attributive.

As discussed in Section 5.3.1, SPADE is a discourse parser that works at the sentence level. SPADE takes a one-sentence-per-line file as input, and outputs one discourse parse tree per sentence. The algorithm uses the syntactic parse trees generated by Charniak's syntactic parser[10]. Both Charniak's parser and SPADE use SUN executables; SPADE also needs Perl 5.0.

To make the SPADE parser suitable for our application, we have added a filter to classify its output to consider only topic relevant sentences (topics were manually annotated in the dataset). Indeed, in query-based summarization or question answering, questions are asked on a particular topic (e.g. event or object). To handle topic-oriented questions, we need to consider a sentence if it is classified as a certain type of rhetorical predicate and contained the topic. For example, in the sentence,

[10]Charniak parser: http://www.cs.brown.edu/people/ec/#software

a. *[perhaps that's why for my European taste Starbucks makes great espresso]*

b. *[while Dunkin's stinks.]*

the topic of the sentence is *Starbucks* and the topic is present in the clauses related by a *contrast* predicate. As a result, we will consider it in the summary generation. On the other hand, in the following two clauses a *cause* predicate is identified by SPADE but the two clauses do not contain the topic (*Subway*), so we will not consider this sentence in summary generation.

a. *[which rocks]*

b. *[because those sandwiches are awesome]*

Moreover, as discussed in Section 5.2, the attribution predicates carry important information for opinionated texts because they express opinions. However, the attribution predicates identified by SPADE can be either opinionated or non-opinionated. For example, "*[I said] [actually I think Zillow IS great.]*" and "*[The legendary GM chairman declared] [that his company would make a car for every purse and purpose.]*" are opinionated and non-opinionated attribution predicates, respectively. In our application, we are only interested in opinionated attribution, hence we kept only attribution sentences which contained opinionated terms. To determine whether the sentence contained any opinionated term, we used a dictionary-based approach using the MPQA lexicon.

In our prototype, to tag a sentence, we use all 4 taggers: the SPADE parser, the comparison classifier, the topic-opinion classifier, and our attributive tagger. By combining these 4 approaches, a sentence is tagged with all possible predicates that it may contain and is ready to be used in a schema. For example, the sentence "*While visiting a Boston Dunkin Donuts this past Sunday morning, I noticed that it was un-surprisingly dingy, and that the counter help spoke very little English, and*

was anything but helpful." contains 3 predicates: contrast, attribution, and joint predicates.

Once our predicate taggers were working, we performed an experiment to see what proportion of sentences were tagged by which predicate taggers. This information serves two purposes: first, all predicate taggers do not perform equally well (see Section 7.4.2) hence knowing which tagger was more used would influence the overall performance of the system; and second we wanted to make sure that the summarization process was receiving various types of predicates. Table 26 shows the result of an

Table 26: Predicate Tagging Distribution based on Classifier Used

Classifier Used	**Distribution**
The SPADE Parser	70%
The Comparison Classifier	6%
The Topic-opinion Tagger	18%
The Attributive Tagger	32%

analysis of 221 random summary sentences from the BLOG06 corpus. It shows that 70%, 6%, 18%, 32% of the sentences were tagged by the SPADE parser, the comparison classifier, the topic-opinion tagger, and the attributive tagger, respectively.

With the same dataset, we also tried to investigate what proportion of sentences are not tagged by any of the taggers to make sure that we are not discarding many sentences in summary generation. Table 27 shows that only 5% of the sentences were not tagged at all and 31% were tagged with multiple predicates.

Hence, we were satisfied that all sentences were given an opportunity to appear in the final summary.

Table 27: Predicate Tagging Distribution based on Number of Tags Occurs

Tag Occurrence	Distribution
No Tag	5%
Single Tag	64%
Multiple Tags	31%

6.2.3 Schema Selection from the Pre-designed Schemata

To use the associated schema for a question category, we have designed three schemata. These schemata are implemented according to their designs shown in Section 4.1.2. These schemata are implemented as an ordered list of rules, where each rule is a combination of predicates that a candidate sentence must satisfy to be selected. In addition, as discussed in Section 4.1.2, constraints may also be specified on predicates. Recall from Section 4.1.2 that 3 types of constraints are used: 1) constraints on the sentence polarity, 2) constraints on the sentence focus, and 3) constraints on the compared objects.

To implement "constraints on the sentence polarity" where the predicate must have the same polarity as the question, we used the polarity information of a sentence and the question calculated in the candidate sentence selection phase. To implement "constraints on the sentence focus" where the topic of the sentence needs to be the focus of the sentence, we verified whether the topic is the subject or object of the sentence using dependency relations from the Stanford parser. To implement the third type of constraint "constraints on the compared objects" we used comparison question patterns to know which objects are compared. In these tasks, which word is the topic is annotated in the dataset.

Figure 41 shows the partial code for the reason schema implementation. In the *schemeQTypeReason* method, constraints for different predicates are shown in bold

text.

Figure 41: Sample Code for Reason Schema

```
private List<Sentence> schemeQTypeReason(List<Sentence> candindates, Sentence query)
  {
    List<Sentence> summary = new ArrayList<Sentence>();
    List<RSTRelationType> relations = new ArrayList<RSTRelationType>();

    relations.add(RSTRelationType.ATTRIBUTION);
    relations.add(RSTRelationType.TOPIC_OPINION);
    List<Sentence> subsummary = extractRelMatchWithSamePol(candindates, query, relations, true);
    summary.addAll(sentenceReOrdering(subsummary));
    candindates.removeAll(subsummary);

    relations.clear();
    relations.add(RSTRelationType.CONTINGENCY);
    relations.add(RSTRelationType.COMPARISON);
    relations.add(RSTRelationType.CONTRAST);
    subsummary = extractRelMatchWithTopicFocus(candindates, query, relations, false);
    summary.addAll(sentenceReOrdering(subsummary));
    candindates.removeAll(subsummary);

    relations.clear();
    relations.add(RSTRelationType.ATTRIBUTIVE);
    subsummary = extractRelMatchWithFocus(candindates, query, relations, false);
    summary.addAll(sentenceReOrdering(subsummary));
    candindates.removeAll(subsummary);

    List<Sentence> post_summary = new ArrayList<Sentence>();

    subsummary = extractSamePolarity(candindates, query);

    ...

    summary.addAll(sentenceReOrdering(post_summary));

    summary = sentenceReOrderingForContrast(summary);

    return summary;
  }
}
```

6.2.4 Summary Generation

Recall from Section 4.1.2 that once a schema is selected for a particular question type and sentences are tagged with rhetorical predicates, this module attempts to use 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 score until the maximum summary length is reached.

To generate the final summary, this module recursively consumes sentences from the candidate sentence list based on the rules implemented in the schema (e.g. reason). At this stage of the content filtering and organization, each candidate sentence containing an extraction score, is labeled with a polarity class (e.g. negative), and is tagged which rhetorical predicates it contains. This module uses all of this information of a candidate sentence to check whether the sentence can fill the current position of the schema. This process is iterative for a given rule and the selected candidate sentences are moved from the candidate sentence pool to the summary pool. The remaining sentences are processed by the next rules in the list. This process is continued until there is no more sentences left in the candidate sentence pool or there is not any rule to match.

Post-Schemata Heuristics

Recall from Section 4.1.2 that we have devised 3 heuristics to achieve a linear sentence order. As described in Section 4.1.2, to implement the heuristic called "Topical similarity" we used the similarity scores between sentences using the cosine similarity based on *tf.idf* above the threshold value. To implement heuristic "Explicit Discourse Marker and Aggregation", we used sentence similarity scores, sentence polarity class, and a set of predefined discourse markers. According the heuristic named "Context" - if a potential sentence contains a pronoun without having a potential antecedent, we include its previous sentence from the source document as a context from the original document. To determine if a sentence contains a pronoun, we used the Stanford POS tagger and look for the PRP (personal pronoun) and PRP$ (possessive pronoun) tags.

6.3 Sample Summaries

To have a better overall view of the results, Figure 42 shows two sample summaries generated by BlogSum. The first summary was rated as "good" and the second summary was rated as "very good" in the human evaluation (see Section 7.3). Appendix-A provides other examples.

Figure 42: Sample Summaries Generated by BlogSum

Topic: Nancy Grace

Question: What reasons did people give for disliking Nancy Grace?

Summary

It also seems people are getting pretty sick of Nancy Grace and her big mouth. Nancy Grace was called a whore this past week and it seems people agree. (I would give my left nut to punch Nancy Grace in face on live TV. It also seems people are getting pretty sick of Nancy Grace and her big mouth. I recently suffered the same Nancy Grace shock you are now experiencing. Any one that would stalk Nancy Grace is guaranteed to be found insane by any court. For the record, the reason Nancy Grace has continually refused to accept this offer is because she knows I will wipe the floor up with her ignorant butt. Alright, I am curious...How do expose Nancy Grace for the cyborg that she really is? Oh crap now Nancy Grace the ghoul of murder is now on CNN HN. "Nancy Grace is one of the few people who has the guts to take on the criminal justice system and expose the judges whose incompetent decisions result in predators being back on the streets. Nancy Grace is a angry, angry woman, who has been mentioned specifically in three appeals cases. And as for Nancy Grace she needs to find the real evidence before she starts yapping her yap and rolling her eyes acting like she knows exactly what has happened!

Topic: A Million Little Pieces

Question: Why did people enjoy "A Million Little Pieces?"

Summary

James Frey wrote a gripping and truthful book about his battle with alcohol and drug addiction that has inspired millions of people to turn their live around and what is more amazing than the conman James Frey is the fact that Millions of people actually thought it was a good book! It still sold millions of copies and readers assumed there was some truth behind it and i can assure you there are millions of people who have and always will stand behind you and your book! So to everyone who claims that "millions" of people have been helped by Frey, I ask you how many have been killed? What is more amazing than the conman James Frey is the fact that Millions of people actually thought it was a good book and he has taken the money and accepted his Oprah given role as "inspiration" for millions of people. Moreover, finally, having read the description of A Million Little Pieces, as well as some of the fawning praise given to it, I have to wonder: "Why are people so enamored with this book and its author?". Absent the Oprah seal of approval, one wonders if the book would have enjoyed the millions of sales it subsequently garnered. It takes an awesome soul to scam a couple million "Oprah brainwashed" housewives out of millions of dollars.

6.4 Conclusion

This chapter provided a detailed description of how our schema-based approach was implemented in the prototype called BlogSum. BlogSum requires a ranked list of candidate sentences as initial input. Then it filters out question irrelevant sentences from this candidate list and reorders them using the most appropriate schema. To achieve this, our approach performs four main tasks namely: question categorization, predicate identification, schema selection, and summary generation.

In Section 4.2, we have demonstrated with an example that our approach seems effective to reduce question irrelevance and discourse incoherence; however, Chapter 7 will provide a more formal evaluation to assess BlogSum's performance with respect to question relevance and coherence and will also analyze the performance of each rhetorical predicate identification approach and the effect of the post-schemata heuristic on sentence ordering.

Chapter 7

Evaluation

As described in Section 2.5, a summary needs to be evaluated for both content and linguistic quality. BlogSum-generated summaries have been evaluated for content and linguistic quality, specifically discourse coherence. The goal of the content evaluation was to measure how effective our approach is at reducing question irrelevance in the final summary. On the other hand, the goal of the evaluation of discourse coherence was to quantify how successful our approach is at decreasing discourse incoherence in the summary. The evaluation of the content was done both automatically and manually and the evaluation of the discourse coherence was done manually. This is because, today, no tool exists to evaluate discourse coherence automatically. Table 28 summarizes the evaluation performed on the overall system.

This chapter also discusses the evaluation of the predicate identification approaches and the effects of post-schema heuristics.

Table 28: Evaluation Performed to Measure Summary Content and Coherence

	Content	**Discourse Coherence**
Automatic	✓ (see Section 7.2.1)	X
Manual	✓ (see Section 7.2.2)	✓ (see Section 7.3.2, 7.3.3)

7.1 Baseline

In our evaluation, BlogSum-generated summaries were mostly compared with the original candidate list (called OList) generated by our approach without the discourse re-ordering (see Section 6.1.2). In order to verify how OList compared with other possible baselines, we have compared it to MEAD [RABG+04], a widely used publicly available summarizer[1]. 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 in the next section) and summaries were evaluated using the ROUGE-2 and ROUGE-SU4 scores. Table 29 shows the results of the automatic evaluation using ROUGE scores based on summary content.

Table 29: 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

Table 29 shows that MEAD-generated summaries achieved weaker ROUGE scores compared to that of our candidate list (OList). The table also shows that MEAD 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 reasons, the rest of the evaluation was performed using OList as a baseline in order to be more strict on our approach.

[1]MEAD: http://www.summarization.com/mead

7.2 Evaluation of Content

The evaluation of BlogSum-generated summary content was done both automatically and manually.

7.2.1 Automatic Evaluation of Content

First, we have automatically evaluated the summaries generated by our approach for content. As described earlier, we used the original ranked list of candidate sentences (see Section 6.1.2), called OList as a baseline, and compared them to the final summaries (BlogSum). We have used the data from the TAC 2008 opinion summarization track and the DUC 2007 dataset for the automatic evaluations of content.

Automatic Evaluation of Content with the TAC 2008 Dataset

Recall that the TAC 2008 opinion summarization Dataset consists of 50 questions on 28 topics; on each topic one or two questions are asked and 9 to 39 relevant documents are given. For each question, one summary was generated by OList and one by BlogSum and the maximum summary length was restricted to 250 words. This length was chosen because in the DUC conference from 2005 to 2007, in the main summarization task, the summary length was restricted to 250 words. In addition, [CS08] created summaries of length 250 words in their participation in the TAC 2008 opinion summarization task and performed well. Furthermore, [CS08] also pointed out that if the summaries were too long this adversely affect summary scores. Moreover, according to the same authors, shorter summaries are easier to read. Based on these observations, we have restricted the maximum summary length to 250 words. However, in the TAC 2008 opinion summarization track, the allowable summary length is very long (the number of non-whitespace characters in the summary must not exceed 7000 times the number of questions for the target of the summary). In this experiment,

we used the ROUGE metric, which is a standard automatic summary content evaluation metric (see Section 2.5). We also used 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. Recall from Section 2.5 that ROUGE-2 is based on the overlap of bi-grams (using words as tokens) between the automatically generated summaries and human-generated gold standard summaries (or reference summaries) while the ROUGE-SU4 score is 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-bigram), and includes uni-gram co-occurrence statistics as well. 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 30. Note that in the table *Rank* refers to the rank of the system compared to the other 36 systems. Table 30 shows that BlogSum (based on OList) 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.

Table 30: 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
Best TAC	0.130	0.139	1
BlogSum based on OList	0.125	0.128	3
BlogSum based on TAC Best	0.138	0.151	<1

Compared to the other systems that participated to the TAC 2008 opinion summarization track, BlogSum performed very competitively; it ranked third and its F-Measure score difference from the TAC best system is very small. Both BlogSum and OList performed better than the TAC average systems.

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 further improve those summaries. The results of this evaluation, shown in Table 30, indicate that BlogSum (based on TAC Best) can further improve the output of a high performing summarizer. This result indicates that our approach does improve the state of the art.

A further manual analysis shows that BlogSum (based on OList) reduced the number of question irrelevant sentences from OList by 21%. However, BlogSum still contains a large number of question irrelevant sentences. We suspect that the reason behind this is that incorrect results of other intermediate tasks such as predicate identification, polarity identification, or design of the schema result in these irrelevant sentences. This is why Sections 7.4 and 7.5 will further evaluate these intermediate steps. From the results, 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 relevant sentence for the question "*What reasons are given as examples of UN commission's ineffectiveness?*" on the topic "UN commission on human right". But BlogSum missed the sentence because it does not attempt to identify the referent for the pronoun "*it*".

Automatic Evaluation of Content with the DUC 2007 Dataset

In this experiment, we used the DUC 2007 Main Task dataset. In this task, given a topic (title) and a set of 25 relevant documents, participants created an automatic summary from the input documents. The documents for summarization came from the AQUAINT corpus (described in Chapter 2), comprising newswire articles. Figure 43 shows an example of the information provided in one DUC 2007 topic.

Figure 43: Sample DUC 2007 Dataset

<num> D0722E </num>

<title> US missile defense system </title>

<narr>

Discuss plans for a national missile defense system. Include information about system costs, treaty issues, and technical criticisms. Provide information about test results of the system.

</narr>

<docs>

XIE19960217.0145

...

</docs>

In this task, the generated summaries needed to be brief, well-organized, and fluent which answers the need for information expressed in the topic statement (<narr> in Figure 43). The summary length was restricted to 250 words. In the dataset, there were 45 topics. In this shared task, 30 participants participated. Participants summaries were evaluated manually for linguistic quality and responsiveness and ROUGE scores were computed automatically to evaluate summary content.

We used this dataset to generate summaries and evaluated BlogSum-generated summaries (based on OList) using ROUGE scores[2]. We have also compared our

[2]I would like to thank Prasad Perera, Masters student of Dr. Leila Kosseim, who conducted this experiment.

results with DUC 2007 participants' results (shown in Table 31). With this dataset, we have also generated summaries using MEAD and evaluated those using ROUGE.

Table 31: Automatic Evaluation of BlogSum based on Summary Content with the DUC 2007 Dataset

System Name	**ROUGE-2 (F-Measure)**	**ROUGE-SU4 (F-Measure)**
BlogSum	0.0930	0.1315
MEAD	0.0988	0.1415
Best	0.1245	0.1771
Average	0.0955	0.1570
Baseline 1	0.0604	0.1051
Baseline 2	0.0938	0.1464

Table 31 shows that BlogSum performed very close to the average performance of the DUC 2007 participant systems. BlogSum performs better than the Baseline 1 and similar to the Baseline 2. It must be noted that the average performance and the performance of the Baseline 2 (the best single document summarizer at DUC 2004) are very similar. These results show that even though BlogSum is designed for opinionated texts, it performs quite satisfactorily on news articles. Interestingly, we can see from the results that BlogSum is good for what it was designed for (blogs), but works quite well for news articles as well.

A further analysis of the results of Table 30 shows that there is no significant difference between BlogSum-generated summaries (based on OList) 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, [CD08, DO08] showed that the performance gap between human-generated summaries and system-generated summaries 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.

7.2.2 Manual Evaluation of Content

Based on an analysis of the 2005-2007 Document Understanding Conference data, [CD08] showed that the ROUGE evaluation and a human evaluation can significantly vary due to the fact that ROUGE ignores linguistic quality of summaries, which has a huge influence in human evaluation. In addition, [DO08] pointed out that automatic evaluation is rather different than the one based on manual assessment. In an automatic evaluation, not only is there no significant gap between models and systems, but in many cases, automatic systems scored higher than some human models.

Table 32: Automated vs. Manual Evaluation at TAC 2008

	Automated		**Manual**		
	ROUGE-2	**ROUGE-SU4**	**Pyramid**	**Linguistics**	**Resp.**
Human Mean	0.12	0.14	0.66	4.79	4.62
System Mean	0.08	0.12	0.26	2.35	2.32
Human Best	0.13	0.17	0.85	4.91	4.79
System Best	0.12	0.14	0.36	3.25	2.79

Table 32 shows the performance of human and automated systems (participants) using automated and manual evaluation in the TAC 2008 update summarization track. Table 32 shows that there is no significant difference between human and participants in automated evaluation but that there is a significant performance difference between them in the manual evaluation.

Given this, we have conducted two manual evaluations using two different datasets to better quantify BlogSum-generated summary content.

Manual Evaluation of Content with the TAC 2008 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 restricted to 250 words. To evaluate content, we asked 3 participants to manually rate 50 summaries from

Figure 44: Sample Summary Generated by BlogSum Used in the Manual Evaluation

Topic: UN Commission on Human Rights

Question: What reasons are given as examples of their ineffectiveness?

Summary-1

Sins of Commission and recall that back in 2001 we were kicked off the commission, and this certainly didn't have any salutary effects on the commission -- recall that a Libyan representative was selected to chair the commission shortly thereafter. Moreover, at the Commission half of all the resolutions that censure states are targeted against Israel. On the table were also 10 resolutions condemning Israel for human rights abuses and in so doing, it has been actively destructive to the interests of human rights. In addition, u.N. claims U.S. social system violates human rights. The problem is, this way, the notion of "human rights" which are supposedly violated by the "dictatorships", loses any practical meaning and it is blatantly obvious that the UN as it currently stands is not capable of protecting human rights. "The presence of the permanent five on any U.N. body makes it more serious and more likely to succeed over the long term, and that includes in the field of human rights and aP said there is some concern that Russia and China would be on the commission permanently. Moreover, six of the fifty-three members of the commission in 2005 were considered among the world's 'worst of the worst' abusers of human rights by Freedom House. Sadly, the commission has devolved into a feckless organization that human-rights abusers use to block criticism or action promoting human rights and in the past two years, Israel has had 101 human rights resolutions passed against it.

Question Relevance

1. **Very Poor**
2. **Poor**
3. **Barely Acceptable**
4. **Good**
5. **Very Good**

OList and 50 summaries from BlogSum using a blind evaluation. In this evaluation, participants also evaluated 50 summaries generated by the top ranked system at the TAC 2008 opinion summarization track. In this evaluation, summaries were rated on a likert scale of 1 to 5 where 1 refers to “very poor” and 5 refers to “very

good"[3]. 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.

Figure 45: Sample OList Summary Used in the Manual Evaluation

Topic: UN Commission on Human Rights

Question: What reasons are given as examples of their ineffectiveness?

Summary-2

Commission on Human Rights. It's a far cry from the organization that gave the world a universal declaration of human rights. As you know, rapporteurs are appointed by the Commission on Human Rights -- not by the High Commissioner and not by the Secretary-General. Commission on Human Rights (UNCHR), whose membership has included regimes bent on heading off resolutions critical of their own records. The UN Human Rights Commission report urges the Bush Administration to put the more than 500 detainees on trial or release them. Sins of Commission. From the Jerusalem Center for Public Affairs: One of the greatest violators of the UN Charter's equality guarantee has been the UN Commission on Human Rights. Six of the fifty-three members of the commission in 2005 were considered among the world's 'worst of the worst' abusers of human rights by Freedom House. This commission has no credibility. A lot of people in the Bush administration don't trust the UN Human Rights Commission; they think it's politically motivated (possible), that it indudes major human rights abusers (true), and that it's out to embarrass Bush (possible). Given that many initiatives of democratic states at the commission pass by only one or two votes, this redistribution is a major blow. In its 2006 world human rights report, the New York-based human rights group Freedom House downgraded Russia from "partly free" to the status of "unfree" (freedomhouse.org). Sadly, the commission has devolved into a feckless organization that human-rights abusers use to block criticism or action promoting human rights.

Question Relevance

1. **Very Poor**
2. **Poor**
3. **Barely Acceptable**
4. **Good**
5. **Very Good**

Figures 44, 45, and 46 show three summaries given to the annotators for evaluation. The summary shown in Figure 44 was generated by BlogSum while the summary shown in Figure 45 is the corresponding OList (restricted to 250 words).

[3]We did not calculate Pyramid scores of BlogSum-generated summaries because to calculate Pyramid scores, 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.

The summary shown in Figure 46 was generated by the TAC best system.

Figure 46: Sample TAC Best System Summary Used in the Manual Evaluation

Topic: UN Commission on Human Rights

Question: What reasons are given as examples of their ineffectiveness?

Summary-3

High health care costs and lack of low-cost housing exacerbate poverty and this can be seen as a human rights abuse. Â™ debating clubÃƒÂƒÃ‚Â¢Ã‚Â‚Ã‚Â€Ã‚Â‚Ã‚Â? esthetic that some ascribe to the whole organization . Like clockwork , the Commission issues regular resolutions condemning Israel while overlooking real offendersÃƒÂƒÃ‚Â¢Ã‚Â‚Ã‚Â€Ã‚Â‚Ã‚Â”such as many of its members . To require that member states participating in the human rights bodies at the U.N. be democracies . To ensure this new body would be no facsimile of its predecessor , the legislation prohibits membership to countries that violate human rights or are subject to specific human rights resolutions . This includes the election of states committed to human rights so that the likes of Lybia cannot serve (let alone chair the Council as Lybia did the current Human Rights Commission) , having the Council be a standing body with the right to investigate specific human rights abuses and to meet more often than the current short spring session in Geneva . While many people say that the Commission is somehow discredited because of recent memberships being granted to Sudan , Cuba , Zimbabwe , Libya , and other perceived pariah countries (at least where human rights are concerned) , some bloggers are not sure that closing the Commission off from such countries is the best idea . The UK and its allies have consistently failed in their desire to get the UN Human Rights Commission to condemn Zimbabwe for the alleged human rights violations . At the Commission half of all the resolutions that censure states are targeted against Israel . It systematically singles out Israel for discriminatory treatment . Another is the replacement of the U.N.'s Commission on Human Rights -- which mocks its own mission by allowing in countries such as Sudan -- with a more credible body . One of the greatest violators of the UN Charter's equality guarantee has been the UN Commission on Human Rights . Candidates for the Human Rights Commission are now selected by a system of regional rotation that makes no distinction between rights advocates and abusers . John R. Bolton , the U.S. ambassador to the United Nations , said he will start the new year by reinvigorating stalled efforts to restructure management of the world body , beginning with a controversial push to seek assurances that the Security Council's five major powers will be guaranteed posts on a new Human Rights Council . Sudan chaired the U.N. Human Rights Commission . In the past two years , Israel has had 101 human rights resolutions passed against it . In releated news , a UN Human Rights Council is proposed to replace the existing 53-member Human Rights Commission , which has been criticized for routinely granting membership to governments with questionable rights records , including Cuba , Libya , Sudan and Zimbabwe. ...

Question Relevance

1. Very Poor
2. Poor
3. Barely Acceptable
4. Good
5. Very Good

In this evaluation, we have calculated the average scores of all 3 annotators' ratings to a particular question to compute the score of each system for a particular question. Table 33 shows the performance comparison between OList and BlogSum; and performance comparison between the TAC best system and BlogSum. The results show that 58% of the time BlogSum summaries were rated better than OList summaries. This implies that 58% of the time, our approach has improved the question relevance compared to that of the original candidate list (OList). The table

also shows that 30% of the time both approaches performed equally well and 12% of the time BlogSum was weaker than OList. From Table 33, we can see that 36% of the time BlogSum performs better than the TAC best system, 23% of the time they perform equally and 41% of the time BlogSum performs weaker than the TAC best system.

Table 33: Comparison of the TAC Best System, OList, and BlogSum based on the Manual Evaluation of Summary Content with the TAC 2008 Dataset

Comparison	%
BlogSum Score > OList Score	58%
BlogSum Score = OList Score	30%
BlogSum Score < OList Score	12%
BlogSum Score > TAC Best Score	36%
BlogSum Score = TAC Best Score	23%
BlogSum Score < TAC Best Score	41%

Table 34 shows the performance of BlogSum versus OList and BlogSum versus the TAC best system on each likert scale; where Δ_1 shows the difference in performance between BlogSum and OList; and Δ_2 shows the difference in performance between BlogSum and the TAC best system. Table 34 demonstrates that 52% of the time, BlogSum summaries were rated as "very good" or "good", 26% of the time they were rated as "barely acceptable" and 22% of the times they were rated as "poor" or "very poor". From Table 34, 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. Results also show that BlogSum performs very competitively compared to the TAC best system.

In this evaluation, we have also calculated whether there is any performance gap

Table 34: Manual Evaluation of the TAC Best System, OList and BlogSum based on Summary Content with the TAC 2008 Dataset

Category	**OList**	**TAC Best**	**BlogSum**	Δ_1	Δ_2
Very Good	6%	9%	14%	8%	5%
Good	16%	41%	38%	22%	-3%
Barely Acceptable	38%	36%	26%	-12%	-10%
Poor	26%	14%	18%	-8%	4%
Very Poor	14%	0%	4%	-10%	4%

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. We have also calculated whether there is any performance gap between BlogSum and the TAC best system. The t-test results show that in a two-tailed test, there is no significant difference between them with a p-value of 0.872.

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 [LK77] with the kappa-value of 0.58.

From Table 33, we can see that our approach improves the original candidate list summaries 58% of the times using the TAC 2008 dataset. Both Tables 33 and 34 demonstrate that our approach outperforms OList-generated summaries and performs competitively compared to the TAC best system. However, from Table 34, we can see that about 22% of the time BlogSum-generated summaries were rated as "poor" or "very poor". A further error analysis of these summaries revealed that

since the initial candidate sentences are selected using simple cosine similarity without applying a deeper semantic analysis, sometimes sentences that contain question or topic terms get higher scores even though they have low question relevance. In some cases, question irrelevance is an outcome of a wrong polarity or predicate identification. Furthermore, in some cases, BlogSum missed question-relevant sentences because a) sometimes relevant sentences do not contain any topic or question term, b) since we did not consider anaphora resolution, sentences which contain anaphoric reference to the topic instead of the direct topic terms, were missed by our approach. For example, the question "*What reasons do people give for liking Zillow?*" was asked on the topic "Zillow". BlogSum missed the following two relevant sentences:

Sentence 1: *It seems to run very smooth and has great overlay graphics.*
Sentence 2: *Very educational.*

because Sentence 1 contains a pronoun that refers to the topic *Zillow* and Sentence 2 contains an ellipsis that also refers to *Zillow*, and contain no other topic or question term.

From Table 33, we can see that 12% of the time, BlogSum-generated summaries were ranked lower than OList-generated summaries. The reason behind this was a wrong predicate identification (see Section 7.4 for an evaluation of predicate identification) and schema design.

We can see that even though there was not any significant performance gap between BlogSum and OList-generated summaries in the automatic evaluation of Section 7.2.1, the manual evaluation shows that BlogSum and OList-generated summaries significantly vary. Moreover, in the automatic evaluation, statistically, there is no significant difference in performance among all participants at TAC 2008. Our

results support [CD08, DO08]'s findings and points out for a better automated summary evaluation tool.

Manual Evaluation of Content with the Review Dataset

Because some of our development was based on the TAC 2008 dataset, we wanted to make sure that our approach had not been tailored to that dataset. To verify this we have conducted a second evaluation using the OpinRank dataset[4] and [JL06]'s dataset on reviews to evaluate BlogSum-generated summary content.

Unfortunately, apart from the TAC 2008 dataset, no publicly available dataset existed for our query-based opinionated summarizer. We therefore had to build our own dataset. To do so, we have used a subset (41,534 reviews) of the OpinRank dataset and [JL06]'s dataset. The OpinRank dataset contains reviews on cars and hotels collected from Tripadvisor and Edmunds. It contains 42,230 reviews on cars for different model-years and 259,000 reviews of 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.

[JL06]'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, 500 to 900 review sentences were provided as input documents to create the summary. Some sample questions and BlogSum-generated summaries are included in Appendix A.

For each question, one summary was generated by OList and one by BlogSum and the maximum summary length was restricted to 250 words again. To evaluate question relevance, we used the same methodology as with the TAC 2008 dataset: 3

[4]OpinRank Dataset: http://kavita-ganesan.com/entity-ranking-data

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 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.

Table 35: Comparison of OList and BlogSum based on the Manual Evaluation of Summary Content with the Review Dataset

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

Here again, 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 35 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. These results are inline with those found for the TAC 2008 dataset (see Table 33).

Table 36: Manual Evaluation of OList and BlogSum based on Summary Content with 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%

Table 36 shows the performance of BlogSum versus OList on each likert scale;

where Δ shows the difference in performance. Table 36 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 "very poor". From Table 36, we can also see that BlogSum outperformed OList in the scale of "very good" by 34% and improved the performance in "poor" and "very poor" categories by 23% and 10%, respectively. These results are somewhat different than those found for TAC 2008 (see Table 34) but in both cases, overall, 30% of the time BlogSum increased the very good or good.

In this evaluation, 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 0.00236.

We have found that the average pair-wise inter-annotator agreement is substantial according to [LK77] with the kappa-value of 0.77.

Figure 47: Results Comparison of the TAC and Review Dataset for Content Evaluation

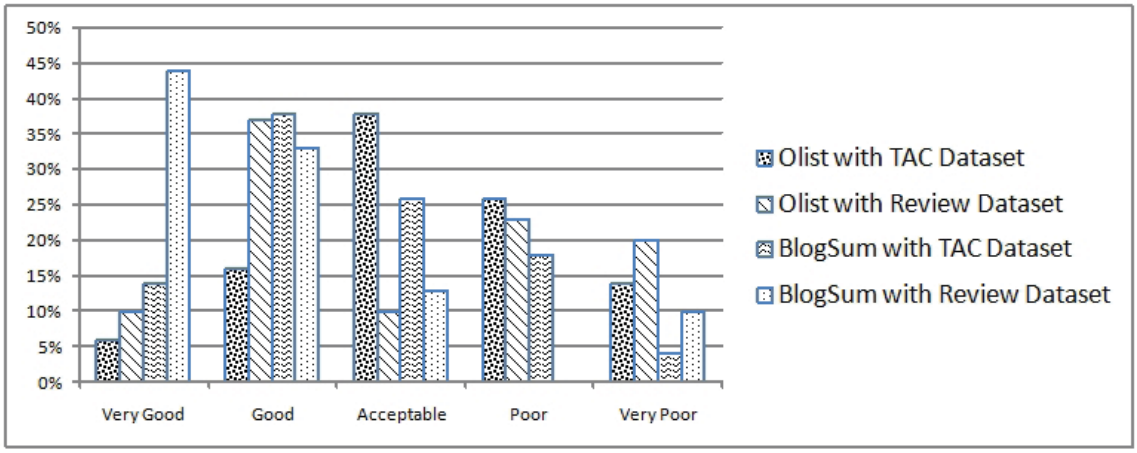


Figure 47 compares the results of the two manual experiments for content using the TAC 2008 dataset and the review dataset. In the experiment with the review dataset, 44% of times BlogSum-generated summaries were rated as "very good" whereas 14% of times BlogSum-generated summaries were rated "very good" for the TAC 2008

dataset. For the review dataset, only 23% of times BlogSum-generated summaries rated as "acceptable", "poor" and "very poor". On the other hand, for the TAC 2008 dataset, 48% of times BlogSum-generated summaries rated as "acceptable", "poor" and "very poor". These results indicate that blogs contain more question irrelevant sentences compared to reviews.

7.3 Evaluation of Discourse Coherence

Recall from Section 1.2 that one of our goals was to reduce discourse incoherence. To this end, the second type of evaluation that we performed was geared at measuring generated summaries for coherence and overall readability. As a baseline, we used the original ranked list of candidate sentences (OList) (restricted to 250 words), and we again compared them to the final summaries generated by BlogSum.

7.3.1 Automatic Evaluation of Discourse Coherence

As shown in Table 28, to evaluate coherence, we did not use an automatic evaluation because, [BGM06] 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 nuggets were provided which had been used as summarization content units (SCUs) in pyramid evaluation (see Section 2.5) to evaluate TAC 2008 participants' summaries but no complete summaries are 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.

7.3.2 Manual Evaluation of Discourse Coherence with the TAC 2008 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 (based on OList) and the maximum summary length was again restricted to 250 words. Four participants manually rated 50 summaries from OList, 50 summaries from BlogSum, and 50 summaries generated by the top ranked system at the TAC 2008 opinion summarization track for coherence using a blind evaluation. These summaries were again rated on a likert scale of 1 to 5 where 1 refers to "very poor" and 5 refers to "very good".

To compute the coherence score of each system for a particular question, we calculated the average scores of all annotators' ratings to that question. Table 37 shows the performance comparison between BlogSum and OList; and the performance comparison between BlogSum and the TAC best system. 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). The table also shows that 77% of the time BlogSum summaries were rated better than the TAC best system summaries; 14% of the time both performed equally well; and 9% of the time BlogSum was weaker than the TAC best system summaries.

Table 38 shows the performance of BlogSum versus OList and the performance of BlogSum versus the TAC best system on each likert scale; where Δ_1 shows the difference in performance between BlogSum and OList and Δ_2 shows the difference in performance between BlogSum and the TAC best system. From Table 38, we can

Table 37: Comparison of the TAC Best System, OList, and BlogSum based on the Manual Evaluation of Discourse Coherence with the TAC 2008 Dataset

Comparison	**%**
BlogSum Score > OList Score	52%
BlogSum Score = OList Score	30%
BlogSum Score < OList Score	18%
BlogSum Score > TAC best Score	77%
BlogSum Score = TAC best Score	14%
BlogSum Score < TAC best Score	9%

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 38 also shows that most of the summaries of the TAC best system were rated as poor or barely acceptable. Further analysis of the results revealed that the summaries of the TAC best system contain lots of question irrelevant sentences which caused their low ranks in coherence.

Table 38: Manual Evaluation of the TAC Best System, OList, and BlogSum based on Discourse Coherence with the TAC 2008 Dataset

Category	**OList**	**TAC Best**	**BlogSum**	Δ_1	Δ_2
Very Good	8%	0%	24%	16%	24%
Good	22%	0%	30%	8%	30%
Barely Acceptable	36%	23%	24%	-12%	1%
Poor	22%	73%	8%	-14%	-65%
Very Poor	12%	4%	14%	2%	-10%

To be noted that the TAC best system achieved a score of 1.98 (out of 5) in the TAC 2008 opinion summarization track's manual evaluation of coherence and in our manual evaluation of coherence, it also achieved a similar score of 2.13 (out of 5). On the other hand, in our manual evaluation of coherence, BlogSum achieved a score of 3.78 (out of 5).

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, BlogSum performed significantly better than OList with a p-value of 0.0223. We have also evaluated if the difference in performance between BlogSum and the TAC best system was statistically significant. The t-test results show that in a two-tailed test, BlogSum performed significantly better than the TAC best system with a p-value of 0.0001. In this experiment, the average pair-wise inter-annotator agreement is substantial according to [LK77] with the kappa-value of 0.76.

The results of Table 37 show that 52% of the time our approach has improved the coherence over the original candidate list (OList). However, in 18% of the time (9 summaries), our approach was weaker than OList. We have analyzed these 9 summaries and found that in 4 cases, some sentences were tagged with the wrong polarity; as a result when the post-schemata heuristics were applied (e.g. discourse markers) they made the summaries weaker. In 3 cases, sentences were tagged with the wrong predicates thus they were included in the final summaries yet they should not have and in 2 other cases, BlogSum excluded sentences which were actually potential sentences again because of a wrong polarity identification and predicate tagging.

7.3.3 Manual Evaluation of Discourse Coherence with the Review Dataset

In this evaluation, we have again used the same dataset (OpinRank dataset and [JL06]'s dataset; described in Section 7.2.2) to conduct the second evaluation of coherence. 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 using a blind evaluation. These summaries were again rated

on a likert scale of 1 to 5 where 1 refers to "very poor" and 5 refers to "very good".

To compute the score of BlogSum and OList for a particular question, we calculated the average scores of all annotators' ratings to that question. Table 39 shows the performance comparison between BlogSum and OList. We can see that 57% of

Table 39: Comparison of OList and BlogSum based on the Manual Evaluation of Discourse Coherence with the Review Dataset

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

the time BlogSum 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. This means that 57% of the time, our approach has improved the coherence compared to that of the original candidate list (OList). Again these results are inline with those found with the TAC 2008 dataset. Hence, showing that our system is not tailored specifically for the TAC 2008 dataset.

Table 40: Manual Evaluation of BlogSum and OList based on Discourse Coherence with 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%

Table 40 shows the performance of BlogSum versus OList on each likert scale;

where Δ shows the difference in performance. From Table 40, 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, BlogSum performed significantly better than OList with a p-value of 0.0371. In addition, the average pair-wise inter-annotator agreement is substantial according to [LK77] with the kappa-value of 0.74.

Results from Table 39 show that in 23% of the time, our approach was weaker than OList. We believe reason behind these are wrong polarity identification, wrong predicate identification, and wrong usage of post-schemata heuristics.

Figure 48: Results Comparison of the TAC and Review Dataset for Discourse Coherence Evaluation

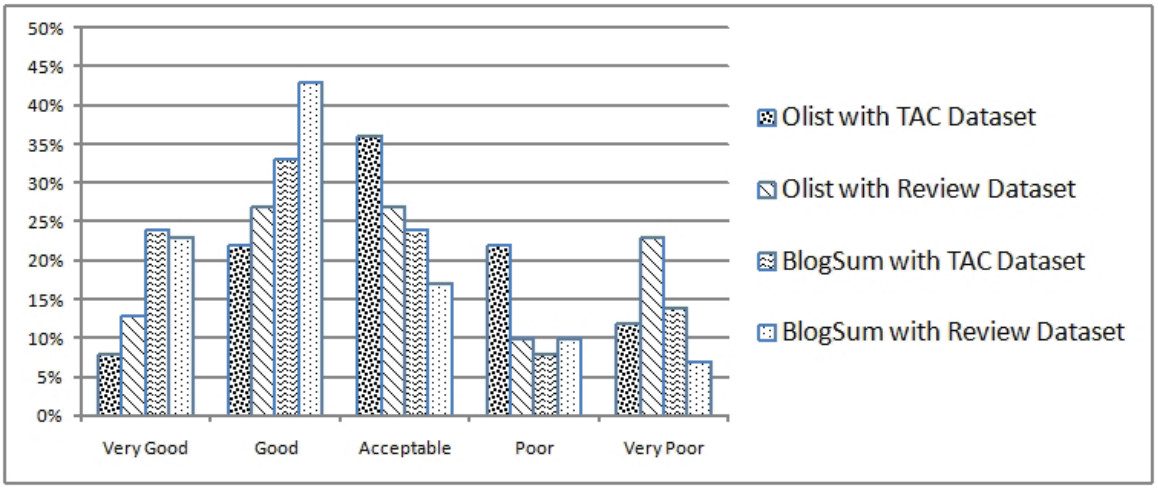


Figure 48 compares the results of the two manual experiments for discourse coherence using the TAC 2008 dataset and the review dataset. The evaluation of coherence shows that for blog dataset, BlogSum-generated summaries were rated as "good" and "very good" 57% of the time compared to 66% of the time 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.

7.4 Evaluation of the Rhetorical Predicate Identification

Both previous evaluations (summary content and discourse coherence) have highlighted the importance of the effectiveness of the intermediate steps of BlogSum; therefore, we tried to measure the effectiveness of a crucial step: the predicate identification. This section describes the corpora and the methodology used to evaluate the predicate identification approaches. This section also provides a comparison with a baseline and human performance for each predicate.

7.4.1 Corpora and Experimental Design

Because BlogSum uses four distinct approaches for predicate identification, we have evaluated these independently. Since the evaluation required annotated corpora, to evaluate each rhetorical predicate identification approach, four different corpora have been used as shown in Table 41. The descriptions of these corpora are given below.

The SPADE Parser Corpus

To evaluate the SPADE parser, the publicly available RST Discourse Treebank 2002[5] was used. This corpus contains 385 Wall Street Journal articles from the Penn TreeBank. It is divided into a training set of 347 articles (6132 sentences) and a testing

[5]Distributed by the Linguistic Data Consortium (http://www.ldc.upenn)

set of 38 articles (991 sentences). In the corpus, for each document, a discourse tree was manually created by following Rhetorical Structure Theory (RST) (see Section 2.4). In the evaluation, only discourse subtrees over individual sentences were used.

The SPADE parser identifies rhetorical predicates which describe relation between two clauses. On the other hand, the other three classifiers (the comparison classifier, the topic-opinion classifier, and the attributive classifier) are used to identify rhetorical predicates that occur within a clause as described in Section 5.3.2. Since the RST Discourse Treebank shows relations between clauses only, we used this corpus to evaluate the SPADE parser but we used three different corpora which show relations within a clause to evaluate the rest of the classifiers.

The Comparison Corpus

To evaluate the comparison classifier, the dataset developed by [JL06] was used. This corpus 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 from different sources. We have randomly selected 1500 sentences for test and the rest of the dataset was used for training where the training and the testing set were mutually exclusive.

The Topic-opinion Corpus

To evaluate the topic-opinion classifier, the corpus developed by [FHW06] from the polarity dataset[6] was used. The original polarity dataset includes 1000 positive and 1000 negative reviews on films. From this polarity dataset, [FHW06] have randomly annotated 400 sentences that contain both film terms and opinionated expressions containing terms from the General Inquirer[7]. The General Inquirer is a publicly

[6]http://www.cs.cornell.edu/people/pabo/movie-review-data
[7]http://www.wjh.havard.edu/∼ inquirer

available opinion dictionary where prior polarity of subjective words are listed. In this corpus of 400 sentences, 262 sentences have an opinion attached to the topic. To annotate this corpus, 86 popular film terms from the dataset and online film glossary[8] were collected by [FHW06].

The Attributive Corpus

Since no standard dataset was available for the attributive predicates, we have developed our own test set containing 400 sentences from the BLOG06 corpus. Two annotators manually tagged 200 sentences as attributive and 200 sentences as non-attributive. Discrepancy between annotators was settled through discussion to arrive at a consensus.

7.4.2 Results

For the evaluation, each approach was evaluated with respect to its associated dataset and the performance was evaluated using precision, recall, and F-Measure scores (see Section 2.5). In [SM03], the SPADE parser's performance was evaluated on 18 discourse relations identification because they group all discourse relations into 18 high-level relations, where each of these relations also contains sub-relations. On the other hand, the performance evaluation of the other three classifiers was binary (e.g. attributive versus non-attributive) because for each predicate we developed a separate classifier.

Table 41 shows the results of the evaluation. The table indicates: a) the rhetorical predicates which have been identified; b) at what level these predicates occurred (within a clause or across two clauses); c) which datasets were used in evaluation; d) which classifier was used to identify the specified predicate; e) the evaluation results using precision, recall, and F-Measure. Note that in the table, RST D. TB refers to

[8] http://www.filmsite.org/filmterms.html

Table 41: Performance of Different Predicate Identification Approaches

Rhetorical Predicate	**Clause Level**	**Dataset**	**Classifier**	**Precision**	**Recall**	**F-Measure**
Comparison	Inter	RST D. TB	SPADE	58%	31%	40%
Comparison	Intra	[JL06]'s	Jindal et al.'s	77%	81%	79%
		[JL06]'s	Our Imp.	66%	68%	67%
Contingency	Inter	RST D. TB	SPADE	85%	76%	80%
Illustration	Inter	RST D. TB	SPADE	79%	93%	85%
Attribution	Inter	RST D. TB	SPADE	52%	83%	64%
Topic-opinion	Intra	[FHW06]'s	Ours	66%	68%	67%
Attributive	Intra	Ours	Ours	77%	76%	77%

RST Discourse Treebank.

To identify inter-clause comparison, contingency, illustration, and attribution predicates, the SPADE parser was used (as explained in Section 5.3.1). As the evaluation of the SPADE parser conducted by [SM03] was executed on 18 relations and the performance for a specific predicate identification is not mentioned in [SM03], we have computed ourselves the performance of the SPADE parser for contingency, comparison, illustration, and attribution predicates using the same corpus used by [SM03]. The performance of the SPADE parser to identify each of these predicates is shown in Table 41.

The table also shows the evaluation results of Jindal et al.'s approach (as published in [JL06]) and our implementation of their approach to identify the comparison predicate which occur within a clause. Jindal et al's [JL06] comparison classifier achieved an F-Measure of 79%. Our development of this classifier (Our Imp.) was much weaker (67%). One possible reason could be that [JL06] used 13 hand crafted rules (not available in their paper), in addition to the keywords and patterns, which we did not use.

Table 41 also shows the evaluation results of our topic-opinion classifier. The topic-opinion classifier achieved an F-Measure of 67%. Recall from Section 5.3.2, that this classifier first identifies whether a sentence is opinion-bearing. If the sentence contains any subjective word then it verifies whether the topic of the sentence is associated with the identified subjective words. We have seen that our polarity identification approach achieved an accuracy of 67%. We suspect this causes the low F-Measure scores of the topic-opinion classifier. This table also shows the evaluation results of our approach to identify the attributive predicate.

Overall, the classifiers achieved an F-measure between 64% to 84%, except for the inter-clause comparison tagged by the SPADE parser which achieved an F-measure of 40%. This shows that in many cases, poor evaluation scores of summary content and discourse coherence of our approach were the result of a wrong predicate identification. In a manual evaluation of discourse coherence, we had found that in some cases the wrong predicate identification was identified as the reason of poor coherence (see Section 7.3.2). The results of the predicate identification evaluation support our previous finding. A wrong predicate identification can lead an inappropriate organization of summary sentences. In a manual analysis of content evaluation, we had found that our approach incorporated many question-irrelevant sentences in the final summary (see Section 7.2.2) and we found that one core reason was the wrong predicate identification because this step plays a key role in our approach for filtering question-irrelevant sentences. The results of the predicate identification evaluation again support our previous finding.

Baseline and Human Performance

To see how each of the classifier performed compared to the baseline and human performance, we have calculated these two measures for each predicate. The human

performance gives us an idea of how difficult or how easy it is to identify a rhetorical predicate and allows us to calibrate our appreciation of the results of the automatic taggers.

To evaluate the predicate tagging approaches, the baseline and the human performance figures were computed as described below:

Baselines:

Inter-Comparison, Contingency, Illustration, Attribution: The SPADE baseline described in [SM03] was used. The baseline algorithm builds right branching discourse trees and labels them with the most frequent relation learned from the training set.

***Intra-Comparison*:** The baseline algorithm considers a sentence as a comparison if it contains any of the keywords of Jindal et al. [JL06].

***Topic-opinion*:** We used the baseline proposed by [FHW06], which considers sentences as topic-opinion if they follow one of the two patterns below:

(RB)+JJ+(NN)+Target

((RB)+JJ)+NN+Target

where, RB, JJ, and NN are parts of speech tags (adverb, adjective, noun) and Target is the topic of the sentence.

***Attributive*:** The baseline system tags a sentence as attributive if heuristic 1 (the topic is the direct subject of the sentence) described in Section 5.3.2 applies. Recall that our approach used three heuristics to tag attributive sentences based on their dependency relations (see Section 5.3.2). Heuristic 1 was chosen as baseline because

it accounts for most cases of our attributive development set (42% of the time).

Human Performance:

Inter-Comparison, Contingency, Illustration, and Attribution: As the human performance, we considered the gold standard described in [SM03]. It was computed as the agreement between two human annotators who independently annotated 53 articles of the RST Discourse Treebank corpus.

Intra-Comparison, Topic-opinion, and Attributive: To evaluate the human performance to tag intra-clause comparison, topic-opinion, and attributive predicates, we asked two human participants to annotate 100 sentences of each type. These 100 sentences were randomly selected from each corpus (e.g. the attributive corpus) where 50 sentences are positive examples (e.g. attributive) and 50 sentences are negative examples (e.g. non-attributive) for a particular predicate (corpora are described in Section 7.4.1). At the end, for each predicate, human performance was compared with the gold standard using precision, recall and F-measure.

Table 42: Baseline and Human Performance of the Rhetorical Predicate Identification Approaches

		Baseline			**Human Performance**		
Rhetorical Predicate	**Clause Level**	**P**	**R**	**F**	**P**	**R**	**F**
Comparison, Contingency, Illustration, Attribution	Inter	unknown	unknown	23%	unknown	unknown	77%
Comparison	Intra	94%	32%	48%	91%	86%	89%
Topic-opinion	Intra	70%	21%	32%	77%	77%	77%
Attributive	Intra	39%	67%	49%	79%	88%	83%

Table 42 shows the baseline and human performance for identifying these rhetorical predicates using precision (P), recall (R), and F-Measure (F). In this evaluation, the baseline and the human performance for inter-clause predicates (e.g. comparison, contingency) are shown from [SM03]. On the other hand, the baseline and the human performance for intra-clause predicates (e.g. comparison, topic-opinion) were computed by us. From Table 42, we can see that human performance is higher than the baseline to tag all predicates. The results also demonstrate that human performance is around 77% to 89% depending on the predicates.

In general, the predicate identification methods used in our work do much better at tagging rhetorical predicates compared to the baseline and do respectably well compared to the human performance (compare Tables 41 and 42). The evaluation shows that these approaches are effective to identify some rhetorical predicates (e.g. illustration) compared to others (e.g. attribution). As Table 41 shows, currently, the state of the art systems have difficulty tagging the rhetorical predicate topic-opinion - achieving an F-Measure of 67%. However, human performance is also rather low (77%), leading us to believe that this predicate is hard to identify. We suspect that because this predicate is not marked explicitly in the text, or may be marked in a variety of ways, it is hard to identify. Moreover, sentiment identification, which is a sub-task of topic-opinion predicate tagging, is a complex task on its own. As a result, the F-Measure scores of the attribution predicate tagging, which also requires sentiment analysis, is also low. On the other hand, the rhetorical predicate intra-comparison is tagged satisfactorily by the state of the art systems, and the human performance is high too. We believe that this rhetorical predicate is more explicitly marked linguistically and in a more stereotypical manner. From the evaluation results, we can see that the precision and the overall F-Measure score of human participants to tag attributive predicates are not very high (around 83%). We suspect that the

reason behind this is that even though attributive relations are found useful in natural language research, this relation is not easy to recognize.

Table 43 shows the inter-annotator agreement using the Cohen's Kappa statistic. Inter-annotator agreement to tag inter-clausal predicates was computed by [SM03] while we have conducted experiments to calculate the inter-annotator agreement to tag intra-comparison, topic-opinion, and attributive predicates. Table 43 shows that inter-annotator agreement to tag comparison, contingency, illustration, attribution, and intra-comparison predicates is substantial according to [LK77]. On the other hand, inter-annotator agreement to tag topic-opinion, and attributive predicates is moderate according to [LK77].

Table 43: Inter-Annotator Agreement on Predicate Tagging

Rhetorical Predicate	**Kappa Value**	**Strength of Agreement**
Inter-Comparison, Contingency, Illustration, Attribution	0.77	Substantial
Intra-Comparison	0.73	Substantial
Topic-opinion	0.52	Moderate
Attributive	0.51	Moderate

7.5 Effects of the Post-Schema Heuristics

The last evaluation we have conducted was to evaluate the effect of the post-schema heuristic rules on our summarization approach. Recall from Section 4.1.2 that in this evaluation, we have tried to compare the difference in performance between summaries generated with and without applying the heuristics rules.

7.5.1 Corpora and Experimental Design

In this evaluation, we have again used the TAC 2008 opinion summarization track data. For each question, two summaries were generated by BlogSum, one with and

one without using the heuristics rules. In this experiment, we restricted the summary length again to 250 words. In our experiment, two participants manually rated 20 summaries from both approaches 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". We have conducted 3 separate experiments for all 3 heuristics rules under the same experimental setting.

In order to identify the effect of each parameter (heuristic), one normally performs several evaluations with different configurations - turning only one parameter on for each configuration, and possibly do all permutations of the 3 parameters as shown in Table 44.

Table 44: Possible Configurations to Evaluate the Post-Schema Heuristics

	C_1	C_2	C_3	C_4	... C_n
Topical Similarity	on	on	on	off	
Discourse Marker	on	off	on	on	...
Context	on	on	off	off	

However, a manual evaluation is very expensive and since the automatic evaluations are not relevant here, we have only calculated the effect of individual heuristics separately on the summarization and only used the configurations shown in Table 45.

Table 45: Configurations Used to Evaluate the Post-Schema Heuristics

	C_1	C_2	C_3	C_4
Topical Similarity	off	on	off	off
Discourse Marker	off	off	on	off
Context	off	off	off	on

7.5.2 Results

Table 46 shows for each heuristic rule, out of 20 summaries, how many summaries received a score of 1, how many summaries received score of 2, and so on, on the

scale of 1 to 5 where 1 refers to "very poor" and 5 refers to "very good". This table shows the scores of BlogSum-generated summaries with (W) and without (W/O) each heuristic.

Table 46: Details of the Effect of the Post-schema Heuristic Rules on Summary Quality

	Topical Similarity		**Discourse Marker**		**Context**	
	W/O	**W**	**W/O**	**W**	**W/O**	**W**
Scale 1 "very poor"	3	2	2	1	3	1
Scale 2	1	1	1	2	4	9
Scale 3	6	4	7	4	6	4
Scale 4	6	7	5	10	5	4
Scale 5 "very good"	4	6	5	3	2	2
Total	20	20	20	20	20	20

Figure 49: Effect of the Post-schema Heuristic Rules on the Summary Quality

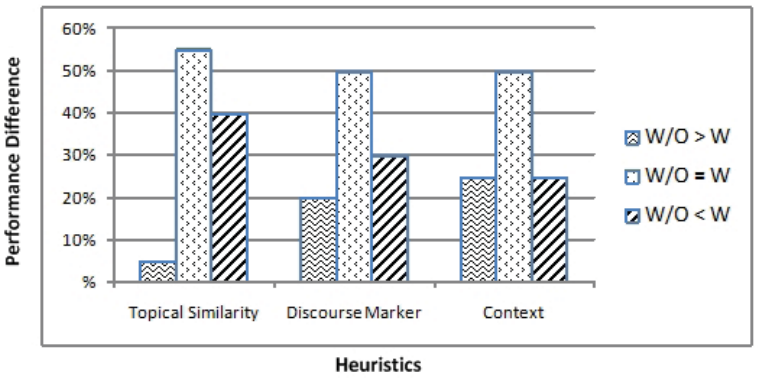


Figure 49 compares the results graphically. In Figure 49, the x-axis shows the heuristics rules and the y-axis shows what percentage of summaries generated using heuristic rules have a lower, equal or higher performance compared to summaries generated without heuristic rules. In this figure, *W/O* and *W* represents BlogSum's performance without and with heuristics rules, respectively. We can see that most of the time, the summaries generated using the heuristics rules are ranked equally

as those without the heuristics (light middle column). In addition, the t-test results show that for all 3 heuristics rules there is no significant performance difference between summaries generated with or without using these heuristics rules. These results indicate that the significant improvement that BlogSum achieved over the OList evaluated in the manual content evaluation (Section 7.2.2) and the manual coherence evaluation (Section 7.3.2, 7.3.3) was not because of the usage of the heuristics but because of the effects of the schemata.

Figure 49 also shows that on a few occasions, summaries generated with the heuristic rules are ranked lower than those without heuristics rules. With the introduction of explicit discourse marker, in 20% of the time, summaries are ranked lower because if two sentences are topically similar and contain the same polarity values then we join them together using discourse markers. However, due to the wrong polarity identification and the bag-of-words approach of similarity calculation, occasionally these sentences degrade the coherence. For the heuristic rule named Context, 25% of the time, summaries using the heuristic rule are ranked lower than summaries without the heuristic. Recall from Section 4.1.2 that to improve context, if a candidate sentence contains a pronoun, we add the previous sentence from the original document. However, since, blogs are very unpredictable and unstructured, sometimes this heuristic creates incoherence instead of reducing it.

7.6 Conclusion

This chapter discussed the performance of BlogSum compared to the original candidate list, TAC 2008 best system, and the performance of shared tasks participants using 3 datasets: the TAC 2008 opinion summarization dataset, the review dataset, and the DUC 2007 dataset. Satisfactory performance with different datasets shows

that our approach behaves well regardless of the dataset. We have calculated BlogSum's performance for question relevance automatically using ROUGE scores and also manually on a likert scale 1 to 5. We have also evaluated BlogSum's performance for discourse coherence manually again on a likert scale 1 to 5. The automatic evaluation of content shows that BlogSum achieved a performance gain over the original candidate list. The manual evaluation shows that our approach performs significantly better than the original candidate list for summary content and coherence. Manual evaluations with the TAC 2008 opinion summarization dataset show that BlogSum performs very competitively compared to the TAC 2008 best system for summary content and significantly better for summary coherence. Evaluation results of this chapter also show that even though our approach was designed for opinionated texts, it can still handle news article summarization. In an experiment, we have also found that if we feed BlogSum the summaries generated by the best system at the TAC 2008 opinion summarization track, BlogSum does improve the state of the art.

This chapter also discussed the evaluation of the rhetorical predicate identification approaches. The evaluation shows that most of the approaches perform better than the baseline and are rather competitive to human performance. These approaches are more effective to identify some rhetorical predicates (e.g. illustration) compared to others (e.g. attribution).

Finally, the evaluation results of post-schema heuristic rules show that they do not have a significant effect on summary quality. Hence the bulk of the improvement is attributable to the use of the schema.

These evaluations show that the goals we had set in our research were achieved - question irrelevance has decreased and discourse coherence has improved significantly.

The next chapter will present some conclusions and highlight future directions of our research.

Chapter 8

Discussion and Conclusion

To identify the challenges involved in blog summarization, we have first identified and categorized errors which typically occur in opinion summarization through an Error Analysis of the current summarization systems (see Chapter 3). The goal of our research was to develop an efficient blog summarization approach that addressed these most frequently occurring errors. We targeted the two most important issues in blog summarization: *question irrelevance* and *discourse incoherence*, which have been identified as the most frequently occurring errors for automatic summaries by various studies as well as from our Error Analysis.

To resolve these errors, we aimed at selecting content properly and organizing it coherently. For this purpose, we have exploited discourse relations of texts and introduced a schema-based approach. To overcome the domain dependency and unavailability of automatic approaches to identify discourse relations across sentences of previous schema-based and discourse relation-based approaches (e.g. [McK85, Mar97a, Bos04]), in our approach, we have utilized intra-sentential discourse relations. We proposed a discourse relation identification approach which is domain and genre independent. To identify discourse relations, we have used a combination of

the RST-based parser SPADE along with three other classifiers. We have also developed an attributive classifier to identify attributive relations. This new tagger has a performance of 77% F-measure, well above the baseline of 49% and not far below human performance.

We have also built a prototype system called BlogSum to test our approach and evaluated BlogSum performance with respect to question relevance and discourse coherence. Evaluation results show that our approach performs significantly better than the original candidate list.

8.1 Main Findings and Contributions of the Thesis

We believe that our work makes the following contributions to the current state of the art in automatic summarization.

8.1.1 Theoretical Contributions

Analysis of Summary-Specific Errors

With the goal of developing an effective summarization approach for blogs, we have performed a systematic analysis and comparison of the current state of the art blog summaries and news summaries. In this analysis (described in Chapter 3), we have identified frequently occurring errors in blog summaries and quantified the information processing difference between the two genres. Our results show that all types of summary-related errors occur more often in blog summaries than in news article summaries. However, topic and question irrelevance as well as discourse incoherence pose a much greater problem for blog summarization than for traditional news articles; while content overlap and missing information seem to be only slightly more frequent in blogs than in traditional news articles. These results show how difficult it

is to process blogs for summarization and show that different information processing techniques are required for these two genres of texts. Details of this work were published in [MK09].

Development of a Schema-based Summarization Approach

Question irrelevance and discourse incoherence are important and typical problems in multi-document summarization. To address these two issues, we have developed a schema-based summarization approach for query-based blog summaries that utilizes discourse structures. Our schema-based approach, takes a ranked list of sentences as input, then categorizes questions to select the appropriate schema which helps to answer different types of questions in different manners with the goal of better meeting the communicative goal. In this process, candidate sentences are tagged based on what rhetorical predicates they convey. Once the schema is selected and candidate sentences are tagged with rhetorical predicates, a summary is generated by filtering question irrelevant sentences and reordering them to be coherent. The schema defines what types of sentences can be used based on rhetorical predicates they convey and in which order these sentences should appear in the final summary. To be question relevant, the schema also specifies semantic constraints on sentences. Details of this work were published in [MK10, MK11b].

Analysis of Current Predicate Tagging Approaches

To evaluate the current state of the art, we methodically analyzed and compared four rhetorical predicate tagging approaches. We have also computed the baseline and human performance for each predicate. In general, the predicate identification methods used in our work do much better at tagging rhetorical predicates compared to the baseline and do respectably well compared to the human performance. However,

the performance varies from F-Measure scores of 85% to 40% (inter-clause comparison is as low as of 40%). This was published in [MK11a].

Other evaluations show that in many cases, poor evaluation scores of summary content and discourse coherence of our approach were the result of a wrong predicate identification. In a manual evaluation of discourse coherence, we had found that in some cases the wrong predicate identification was identified as the reason for poor coherence. Results of the predicate identification evaluation support our previous finding. The wrong predicate identification can lead to an inappropriate summary organization. In a manual analysis of content evaluation, we have found that our approach incorporates many question irrelevant sentences in the final summary and we have found that one core reason was the wrong predicate identification because this step plays a key role in our approach to filter out question irrelevant sentences.

Identification and Development of a Predicate Identification Approach

We have introduced a predicate tagging approach for the attributive predicates, included in Grimes' relation list [Gri75]. An attributive predicate provides details about an entity or an event or can be used to illustrate a particular feature about a concept or an entity. Since attributive predicates describe attributes or features of an object or an event, they are often used in query-based summarization and question answering systems. However, to our knowledge, no previous work has focused on tagging attributive predicates automatically. We proposed an automatic domain and genre-independent approach to tag attributive predicates by utilizing dependency relations of words based on dependency grammars [dMM08]. By using a subset of the BLOG06 corpus, we have evaluated the accuracy of our attributive classifier and compared it to a baseline and human performance using precision, recall, and F-Measure. It achieved an accuracy of 77% F-measure which is well above the baseline of 49% and compares

favorably with human performance of 83%. This was published in [MK11a].

8.1.2 Practical Contributions

Design of a Prototype

We have developed a prototype called BlogSum to validate our schema-based summarization approach. Given an initial topic and question and a set of related blogs, BlogSum first creates a ranked list of sentences that could potentially be included in the final summary. To create this ranked list of sentences, BlogSum performs preprocessing such as filtering non-textual content from the blogs and then creates a preliminary candidate list using question similarity, topic similarity and subjectivity scores (see Section 6.1.2). In the next step, BlogSum removes redundant sentences from the candidate list to address content overlap errors. To remove redundant sentences, the cosine value is calculated for each pair of sentences. Before inserting a sentence into the list of candidate sentences, it is checked for similarity with the sentences already in the list. If the sentence is similar to any of the sentences in the list above a threshold then it is not inserted. In this process, candidate sentences are checked for redundancy. Then BlogSum generates summaries by categorizing the initial question based on its communicative goal (see Section 6.2.1), identifying the rhetorical predicates that each candidate sentence conveys (see Section 6.2.2), selecting the appropriate schema for the identified question category (see Section 6.2.3), and generating the summary by filtering and reordering sentences (see Section 6.2.4).

Evaluation of the Approach

The evaluation of BlogSum empirically supports our theoretical developments. BlogSum-generated summaries have been evaluated for content and coherence using several datasets. The content evaluation gives an indication of the question relevance of the summary as well as the usefulness of our approach and the coherence evaluation gives an indication of the coherence of the summary. In these evaluations, BlogSum-generated summaries mostly have been compared with the original candidate list called OList.

An automatic evaluation using the ROUGE metric has been performed to assess BlogSum-generated summary content using the TAC 2008 opinion summarization dataset. The evaluation results show that our approach has a positive effect on content selection and our approach performed very competitively (positioned at rank 3) compared to all 36 participants in TAC-2008. Our approach also achieved a performance gain of about 18% in F-Measure over the original ranked list in removing question irrelevant sentences. We have conducted another automatic experiment to evaluate BlogSum-generated summary content using the ROUGE metric with the DUC 2007 dataset on news articles. Evaluation results show that even though BlogSum was designed for opinionated texts, it performed quite satisfactorily with news articles; very close to the average performance of the DUC 2007 participants. In an automatic evaluation, we have also found that if we feed BlogSum the summaries generated by the best system at the TAC 2008 opinion summarization track, BlogSum does improve the state of the art.

We have also conducted two manual evaluations using two different datasets to quantify BlogSum-generated summary content. In the first evaluation, three human annotators manually rated 50 summaries generated by the TAC 2008 best system, 50 summaries from BlogSum, and 50 summaries from OList from the TAC 2008 opinion

summarization dataset in a likert scale of 1 to 5. The results show that 58% of the time BlogSum summaries were rated better than OList summaries which implies that our approach has improved question relevance compared to that of the original candidate list (OList). The results show that BlogSum performed competitively compared to the TAC best system. In this evaluation, 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. In this evaluation, we have also calculated whether there is any performance gap between BlogSum and the TAC best system summaries and the t-test results show that there is no significant difference in performance. In the second evaluation, three human annotators manually rated 30 summaries generated by BlogSum and 30 summaries from OList from the OpinRank dataset[1] and [JL06]'s dataset in a likert scale of 1 to 5. The results show that 67% of the time BlogSum summaries were rated better than OList summaries. The t-test results show that in a two-tailed test, BlogSum performed very significantly better than OList with a p-value of 0.0023.

We have also conducted two manual evaluations using two different datasets to quantify BlogSum-generated summary coherence. The performance of BlogSum was evaluated using the TAC 2008 opinion summarization dataset for coherence manually using four human participants in a likert scale of 1 to 5. The results indicate that about 54% of BlogSum summaries are rated as "very good" or "good" as opposed to 30% for the summaries generated by OList. The evaluation results also show that our approach has significantly improved summary coherence compared to that of the original candidate list with a p-value of 0.0223 in a t-test. The evaluation results also show that BlogSum performed significantly better than the TAC 2008 best system with a p-value of 0.0001. In a second evaluation, BlogSum-generated summaries

[1]OpinRank Dataset: http://kavita-ganesan.com/entity-ranking-data

using OpinRank dataset and [JL06]'s dataset were evaluated for coherence manually by 3 annotators in a likert scale of 1 to 5. The results show that 57% of the time BlogSum performs better than OList. The t-test results show that in a two-tailed test, BlogSum performed significantly better than OList with a p-value of 0.0371.

We have also conducted an evaluation to evaluate the effect of the post-schema heuristics rules on our summarization approach using the TAC 2008 opinion summarization dataset. In this evaluation, we have tried to compare the difference in performance between summaries generated with and without applying the heuristics rules. To do so, two annotators manually evaluated 20 summaries generated by BlogSum with heuristics and 20 summaries from BlogSum without heuristics in a likert scale of 1 to 5. However, the t-test results show that for all 3 heuristics rules there is no significant performance difference between summaries generated with or without using these heuristics rules. This result indicates that the significant improvement BlogSum achieved over OList evaluated in the manual coherence evaluation is not due to the usage of heuristics but because of the effects of the schemata.

Identification of Summary Evaluation Issues

Based on the DUC and TAC evaluation results, [CD08, DO08] showed that the performance gap between humans and systems, which is clearly visible in the manual evaluation is often not reflected in automated evaluations using ROUGE scores. In our content evaluation, we have used the standard automated measure ROUGE (ROUGE-2 & ROUGE-SU4) and the t-test results show that there is no significant difference between BlogSum-generated summaries and OList summaries with a *p-value* of 0.228 and 0.464 for ROUGE-2 and ROUGE-SU4, respectively. 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 two manual evaluations for content using two different datasets. In both evaluations, we have also calculated whether there is any performance gap between BlogSum and OList. The t-test results for both datasets show that in a two-tailed test, BlogSum performed significantly better than OList with a p-value of 0.00281 and 0.00236. We can see that even though there was no significant performance gap between BlogSum and OList-generated summaries in the automatic evaluation, the manual evaluation results clearly show that BlogSum summaries are significantly better than OList. Our results supports [CD08, DO08]'s findings and points out for a better automated summary evaluation tool. This discrepancy has been reported in [MKP12].

8.2 Directions for Future Research

My thesis supervisor told me one day that there are "complete" thesis and that there are "thesis that are finished". Work on this thesis could go on and on as many questions are left without answers; and even if we answer these, it will only create more questions. So for the sake of having a "thesis that is finished", we will describe here our current open questions, in the hopes that some day they will be answered.

8.2.1 Extensions

Test BlogSum with Different summary Lengths

As discussed in Chapter 7, currently BlogSum-generated summary length is 250 words, in the future, it will be also interesting to check whether summary length has any effect on the summary quality.

Address all Summary Related Errors

In our summary-related error analysis (Chapter 3), we have identified seven main types of errors. However, currently, we are only addressing question irrelevance and discourse incoherence which are most frequent errors. In the future, all types of errors should be addressed.

Evaluate the Effect of Schema Design

Currently, we have designed three schemata for three different question types namely comparison, reason, and suggestion. Schemata specify which types of sentences can fill the schema and in which order they should appear. In other words, a schema determines the final summary content and organization. In the future, 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.

Add More Linguistic Tools

From the evaluation results, we have found that sometimes BlogSum missed question-relevant sentences because it does not perform anaphora resolution. In the future, we would like to incorporate anaphora resolution in BlogSum. In addition, we would like to include other NLP tools in BlogSum such as a named entity identifier to give importance to proper nouns found in the questions.

Evaluate the Effect of Polarity Identification

In our schema-based approach, polarity information is used at several critical points: in candidate sentence selection, predicate identification, and summary generation. From a manual analysis of question irrelevance and discourse coherence, we have

found that in many cases, a wrong polarity identification has been identified as the cause of poor evaluation results. Currently, we are using a simple dictionary-based polarity identification approach. In the future, it would be interesting to use more sophisticated polarity identification approach such as the combined sentiment analysis approach developed in [And09] to identify sentence-level sentiment to use the benefit of lexical-based and corpus-based approaches for this. In the future, it would be also interesting to evaluate the effect of polarity identification on summary quality.

Analyze and Improve Predicate Identification Approaches

From the evaluation results of Chapter 7, we have seen that some predicates are very difficult to identify. In the future, it would be helpful to identify reasons for their wrong identification and try to improve their performance. Since predicate identification plays a key role in our approach, if we can improve their performance, the overall summary quality of our approach should be improved. We believe that the use of discourse relations would be investigated more often in NLP if more reliable tools were available. As stated in Chapter 5, our study focused only on six main categories of rhetorical predicates, in the future, it would be also interesting to consider other predicates, for example *antithesis* and evaluate the effect of predicate identification on summary quality. In the future, it would be also helpful to use rhetorical predicates which occur across sentences.

Analyze and Improve Post-schema Heuristics

The evaluation results of Chapter 7 show that post-schema heuristics do not have significant effect on BlogSum-generated final summaries. In the future, it would be interesting to investigate the reason behind that and try to improve them. Thus would improve the coherence of the final summaries.

8.2.2 Future Directions

When putting our work into perspective, one can also consider larger directions of future work. A few are listed here.

Apply Our Approach to Other Applications

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.

Apply Our Approach to News Summarization

Some News articles such as editorials could also contain opinionated content. Thus, it would be interesting to apply our approach on such news articles. Since news articles are more structured than blogs, we expect that our approach will work even better for such texts. However, news articles might contain more fine grained discourse relations compared to blogs. As a result, predicate identification for news articles will be an interesting challenge. To be mentioned that we have already tested our approach with factual news articles (DUC 2007) and on that dataset, our approach performed close to the average performance of the DUC 2007 participants.

In the beginning of this dissertation, we raised two questions 1) "*How discourse relations and text schemata can be utilized to reduce question irrelevance and discourse incoherence?* and 2) "*How we can identify different types of discourse relations automatically for any given domain?*" We finish this thesis with the hope that our work has provided some answers to these questions.

Bibliography

[AB08] A. Andreevskaia and S. Bergler. When Specialists and Generalists Work Together: Overcoming Domain Dependence in Sentiment Tagging. In *Proceedings of ACL-08: HLT*, pages 290–298, Columbus, Ohio, USA, June 2008.

[ABU07] A. Andreevskaia, S. Bergler, and M. Urseanu. 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)*, Boulder, Colorado, March 2007.

[AKS05] S. Afantenos, V. Karkaletsis, and P. Stamatopoulos. Summarization from medical documents: A survey. *Artificial Intelligence in Medicine*, 33(2):157–177, 2005.

[And09] A. Andreevskaia. *Sentence-level Sentiment Tagging Across Different Domains and Genres*. PhD thesis, Department of Computer Science and Software Engineering, Concordia University, Montreal, Canada, 2009.

[Ari54] Aristotle. *The Rhetoric. Translation in The Rhetoric and the Poetics of Aristotle*. Random House, New York, 1954.

[Ash93] N. Asher. *Reference to Abstract Objects in Discours*. Kluwer Academic Press, Boston, 1993.

[BBW08] S. Bellemare, S. Bergler, and R. Witte. ERSS at TAC 2008. In *Proceedings of the Text Analysis Conference*, Gaithersburg, Maryland, USA, 2008.

[BEM02] R. Barzilay, N. Elhadad, and K.R. McKeown. Inferring Strategies for Sentence Ordering in Multidocument News Summarization. *Artificial Intelligence Research*, 17:35–55, 2002.

[BG07] S.J. Blair-Goldensohn. *Long-answer Question Answering and Rhetorical-semantic Relations*. PhD thesis, Department of Computer Science, Columbia University, Columbia, USA, 2007.

[BG08] A. Bossard and M. Genereux. Description of the LIPN Systems at TAC 2008: Summarizing Information and Opinions. In *Proceedings of the Text Analysis Conference*, Gaithersburg, Maryland, USA, November 2008.

[BGM06] S.J. Blair-Goldensohn and K. McKeown. 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, June 2006.

[BL04] R. Barzilay and L. Lee. Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization. In *Proceedings of HLT-NAACL*, pages 113–120, Boston, USA, 2004.

[Bos04] W. Bosma. Query-Based Summarization using Rhetorical Structure Theory. In *15th Meeting of Computational Linguistics in the Netherlands CLIN*, pages 29–44, Leiden, Netherlands, December 2004.

[Bos09] W. Bosma. Contextual salience in query-based summarization. In *Proceedings of the International Conference RANLP-2009*, pages 39–44, Borovets, Bulgaria, September 2009.

[CD08] J.M. Conroy and H.T. Dang. Mind the Gap: Dangers of Divorcing Evaluations of Summary Content from Linguistic Quality. In *Proceedings of the the 22nd International Conference on Computational Linguistics Coling*, pages 145–152, Manchester, UK, 2008.

[CHJ08] Y. Chali, S.A. Hasan, and S.R. Joty. UofL at TAC 2008 Update Summarization and Question Answering. In *Proceedings of the Text Analysis Conference*, Gaithersburg, Maryland, USA, 2008.

[CKK+08] T. Copeck, A. Kazantseva, A. Kennedy, A. Kunadze, D. Inkpen, and S. Szpakowicz. Update Summary Update. In *Proceedings of the Text Analysis Conference*, Gaithersburg, Maryland, USA, 2008.

[CM01] L. Carlson and D. Marcu. Discourse Tagging Reference Manual. Technical Report ISI-TR-545, University of Southern California Information Sciences Institute, 2001.

[CN94] B.E. Cline and J.T. Nutter. Kalos - A System for Natural Language Generation with Revision. In *AAAI'94: Proceedings of the Twelfth National Conference on Artificial Intelligence (vol. 1)*, pages 767–772, Seattle, Washington, USA, July-August 1994.

[CS08] J.M. Conroy and J.D. Schlesinger. CLASSY and TAC 2008 Metrics. In *Proceedings of the Text Analysis Conference*, Gaithersburg, Maryland, USA, 2008.

[CSS05] J.M. Conroy, J.G. Stewart, and J.D. Schlesinger. CLASSY query-based multi-document summarization. In *Proceedings of the Document Understanding Conference. Workshop at the Human Language Technology Conference on Empirical Methods in Natural Language Processing HLT/EMNLP*, 2005.

[Dan08] H.T. Dang. Overview of the TAC 2008 Opinion Question Answering and Summarization Tasks. In *Proceedings of the Text Analysis Conference*, Gaithersburg, Maryland, USA, 2008.

[DM07] D. Das and A.F.T. Martins. A Survey on Automatic Text Summarization, 2007. Available from: http://www.cs.cmu.edu/∼nasmith/LS2/das-martins.07.pdf, Literature Survey for the Language and Statistics II course at Carnegie Mellon University.

[dMM08] M.C. de Marneffe and C.D. Manning. The Stanford Typed Dependencies Representation. In *Coling 2008: Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8, Manchester. UK, 2008.

[DO08] H.T. Dang and K. Owczarzak. Overview of the TAC 2008 Update Summarization Task. In *Proceedings of the Text Analysis Conference*, Gaithersburg, Maryland, USA, 2008.

[DOCS07] D.M. Dunlavy, D.P. O'Leary, J.M. Conroy, and J.D. Schlesinger. QCS: A System for Querying, Clustering and Summarizing Documents. *Information Processing Management*, 43(6):1588–1605, 2007.

[Dub11] J. Dubuc. Modeling the Evolving Structure of Social Text for Information Extraction and Topic Detection. Master's thesis, Department

of Computer Science and Software Engineering, Concordia University, Montreal, Canada, 2011.

[FHW06] Z. Fei, X. Huang, and L. Wu. Mining the Relation between Sentiment Expression and Target Using Dependency of Words. In *PACLIC20: Coling 2008: Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation*, pages 257–264, Wuhan, China, November 2006.

[FR08] S. Fisher and B. Roark. Query-focused Supervised Sentence Ranking for Update Summaries. In *Proceedings of the Text Analysis Conference*, Gaithersburg, USA, 2008.

[GKV06] G. Giannakopoulos, V. Karkaletsis, and G. Vouros. Automatic Multi-document Summarization and Prior Knowledge: Past, Present and Vision. Technical Report DEMO-2006-2, NCSR Demokritos, Greece, 2006.

[GL86] B.J. Grosz and Sidner. C. L. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, 12(3):175–204, 1986.

[GL10] V. Gupta and G.S. Lehal. A Survey of Text Summarization Extractive Techniques. *Emerging Technologies in Web Intelligence*, 2(3):258–268, 2010.

[GLNW09] P. Genest, G. Lapalme, L. Nerima, and E. Wehrli. A Symbolic Summarizer for the Update Task of TAC 2008. In *Proceedings of the Text Analysis Conference*, Gaithersburg, USA, 2009.

[GLYM09] P. Genest, G. Lapalme, and M. Yousfi-Monod. HEXTAC: the Creation of a Manual Extractive Run. In *Proceedings of the Text Analysis Conference*, Gaithersburg, USA, 2009.

[GMCK00] J. Goldstein, V. Mittal, J. Carbonell, and M. Kantrowitz. Multi-Document Summarization By Sentence Extraction. In *Proceedings of the ANLP/NAACL Workshop on Automatic Summarization*, pages 40–48, 2000.

[Gri75] J.E. Grimes. The Thread of Discourse. Technical Report NSF-TR-1, NSF-GS-3180, Cornell University, Ithaca, New York, 1975.

[Gro85] B.J. Grosz. Discourse Structure and the Proper Treatment of Interruptions. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pages 832–839, 1985.

[GS98] B. Grote and M. Stede. Discourse Marker Choice in Sentence Planning. In *Proceedings of the 9th International Workshop on Natural Language Generation*, pages 128–137, Niagara-on-the-Lake, Canada, 1998.

[GSS07] N. Godbole, M. Srinivasaiah, and S. Skiena. Large-Scale Sentiment Analysis for News and Blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM'2007)*, pages 219–222, Boulder, Colorado, USA, March 2007.

[HB08] I. Hendrickx and W. Bosma. Using Coreference Links and Sentence Compression in Graph-based Summarization. In *Proceedings of the Text Analysis Conference*, Gaithersburg, Maryland, USA, 2008.

[HCGL08] T. He, J. Chen, Z. Gui, and F. Li. CCNU at TAC 2008: Proceeding on Using Semantic Method for Automated Summarization Yield. In *Proceedings of the Text Analysis Conference*, Gaithersburg, USA, 2008.

[HL04] M. Hu and B. Liu. Mining and Summarizing Customer Reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177, 2004.

[HM93] E.H. Hovy and E. Maier. Parsimonious or Profligate: How many and which discourse structure relations? Unpublished Manuscript. Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.141.2760, 1993.

[Hob85] J.R. Hobbs. On the Coherence and Structure of Discourse. Technical Report CSLI-85-37, Center for the Study of Language and Information, Stanford University, 1985.

[Hov93] E.H. Hovy. Automated Discourse Generation using Discourse Structure Relations. *Artificial Intelligence*, 63(1-2):341–385, 1993.

[Jag06] J. Jagarlamudi. Query-Based Multi-Document Summarization using Language Modeling. Master's thesis, International Institute of Information Technology, Hyderabad, India, 2006.

[JHZ09] F. Jin, M. Huang, and X. Zhu. A Query-specific Opinion Summarization System. In *Proceedings of the 8th International Conference on Cognitive Informatics*, pages 428–433, Kowloon, Hong Kong, 2009.

[JKN10] K. Jaidka, C.S.G. Khoo, and J. Na. Imitating Human Literature Review Writing: An Approach to Multi-document Summarization. In *Proceedings of the 12th International Conference on Asia-Pacific Digital Libraries*, pages 116–119, Gold Coast, Australia, 2010.

[JL06] N. Jindal and B. Liu. Identifying Comparative Sentences in Text Documents. In *SIGIR'06: Proceedings of the 29th Annual International*

ACM SIGIR Conference on Research and Development in Information Retrieval, pages 244–251, Seattle, Washington, USA, August 2006.

[JM00] D. Jurafsky and J.H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistic, and Speech Recognition*. Prentice Hall, 2000.

[Kam81] H. Kamp. A theory of truth and semantic representation. In J. A. G. Groenendijk, T. M. V. Janssen, and M. B. J. Stokhof, editors, *Formal Methods in the Study of Language*, volume 1, pages 277–322. Mathematisch Centrum, Amsterdam, 1981.

[KC08] S. Kumar and D. Chatterjee. IIIT Kharagpur at TAC 2008: Statistical Model for Opinion Summarization. In *Proceedings of the Text Analysis Conference*, Gaithersburg, USA, 2008.

[KD93] A. Knott and R. Dale. 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*, pages 47–67, Pisa, Italy, 1993.

[KD94] A Knott and R Dale. Using Linguistic Phenomena to Motivate a Set of Coherence Relations. *Discourse Processes*, 18(1):35–62, 1994.

[KLC06] L. Ku, L. Lee, and H. Chen. 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, 2006.

[KPVZ08] H. D. Kim, D. H. Park, V. G. V. Vydiswaran, and C. Zhai. 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, Maryland, USA, 2008.

[Lap03] M. Lapata. Probabilistic Text Structuring: Experiments with Sentence Ordering. In *Proceedings of the Annual Meeting of ACL*, pages 545–552, Sapporo, Japan, 2003.

[Lin04] C.-Y. Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July 2004.

[LK77] R.J. Landis and G.G. Koch. A One-way Components of Variance Model for Categorical Data. *Biometrics*, 33(1):671–679, 1977.

[LKS06] L. Lloyd, P. Kaulgud, and S. Skiena. Newspapers vs. Blogs: Who Gets the Scoop? In *Proceedings of the AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs, 2006*, California, USA, March 2006.

[LLWH07] M. Liu, W. Li, M. Wu, and Hu. H. 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*, pages 117–122, Henan, China, 2007.

[LN08] A. Louis and A. Nenkova. Automatic Summary Evaluation without Human Models. In *Proceedings of the First Text Analysis Conference (TAC 2008)*, Gaithersburg, Maryland, USA, November 2008.

[LOHW08] W. Li, Y. Ouyang, Y. Hu, and F. Wei. PolyU at TAC 2008. In *Proceedings of the First Text Analysis Conference (TAC 2008)*, Gaithersburg, Maryland, USA, November 2008.

[LTHZ09] F. Li, Y. Tang, M. Huang, and X. Zhu. Answering Opinion Questions with Random Walks on Graphs. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 737–745, Suntec, Singapore, August 2009.

[Luh58] H.P. Luhn. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, 2(2):159–165, 1958.

[Mar97a] D. Marcu. From Discourse Structures to Text Summaries. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarizatio*, pages 82–88, Madrid, Spain, July 1997.

[Mar97b] D. Marcu. *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*. PhD thesis, Department of Computer Science, University of Toronto, Toronto, Canada, 1997.

[MBB98] I. Mani, E. Bloedorn, and Gates. B. Using Cohesion and Coherence Models for Text Summarization. In *Proceedings of the Spring Symposium on Intelligent Text Summarization (AAAI 98)*, pages 69–76, Stanford, CA, USA, March 1998.

[McK85] K.R. McKeown. Discourse Strategies for Generating Natural-Language Text. *Artificial Intelligence*, 27(1):1–41, 1985.

[ME02] D. Marcu and A Echihabi. An Unsupervised Approach to Recognizing Discourse Relations. In *ACL'02: Proceedings of the 40th Annual*

Meeting on Association for Computational Linguistics, pages 368–375, Philadelphia, Pennsylvania, USA, July 2002.

[MG06] G. Mishne and N. Glance. Predicting Movie Sales from Blogger Sentiment. In *Proceedings of the AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006)*, 2006.

[Mit93] R. Mitkov. 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*, pages 86–89, Columbus, Ohio, June 1993.

[MJCN08] G. Murray, S. Joty, G. Carenini, and R. Ng. The University of British Columbia at TAC 2008. In *Proceedings of the Text Analysis Conference*, Gaithersburg, USA, 2008.

[MK09] S. Mithun and L. Kosseim. Summarizing Blog Entries versus News Texts. In *Proceedings of Events in Emerging Text Types (eETTS). A Workshop of Recent Advances in Natural Language Processing RANLP 2009*, pages 35–42, Borovets, Bulgaria, September 2009.

[MK10] S. Mithun and L. Kosseim. A Hybrid Approach to Utilize Rhetorical Relations for Blog Summarization. In *Proceedings of TALN*, Montreal, Canada, 2010.

[MK11a] S. Mithun and L. Kosseim. Comparing Approaches to Tag Discourse Relations. In *Proceedings of CICLing*, pages 328–339, Tokyo, Japan, 2011.

[MK11b] S. Mithun and L. Kosseim. Discourse Structures to Reduce Discourse Incoherence in Blog Summarization. In *Proceedings of Recent Advances*

in Natural Language Processing (RANLP), pages 479–486, Hissar, Bulgaria, 2011.

[MKH+02] K. McKeown, J. Klavans, V. Hatzivassiloglou, R. Barzilay, and E. Eskin. Towards Multidocument Summarization by Reformulation: Progress and prospects. In *Proceedings of AAAI/IAAI*, pages 27–36, Edmonton, Canada, 2002.

[MKP12] S. Mithun, L. Kosseim, and P. Perera. Discrepancy Between Automatic and Manual Evaluation of Summaries. In *NAACL-HLT 2012 Workshop on Evaluation Metrics and System Comparison for Automatic Summarization (Accepted)*, 2012.

[MOS07] C. Macdonald, I. Ounis, and I. Soboroff. Overview of the TREC 2007 Blog Track. In *Proceedings of the Sixteenth Text REtrieval Conference (TREC 2007)*, Gaithersburg, Maryland, USA, November 2007.

[MR06] A.A. Mohamed and S. Rajasekaran. Query-Based Summarization Based on Document Graphs. In *Proceedings of IEEE International Symposium on Signal Processing and Information Technology*, pages 408–410, Vancouver, Canada, 2006.

[MT88] W.C. Mann and S.A. Thompson. Rhetorical Structure Theory: Toward a Functional Theory of Text Organisation. *Text*, 3(8):234–281, 1988.

[NFK02] J.L. Neto, A.A. Freitas, and C.A.A. Kaestner. Automatic Text Summarization Using a Machine Learning Approach. In *Proceedings of the 16th Brazilian Symposium on Artificial Intelligence: Advances in Artificial Intelligence*, pages 205–215, London, UK, 2002.

[NFP08] V. Nastase, K. Filippova, and S.P. Ponzetto. Generating Update Summaries with Spreading Activation. In *Proceedings of the Text Analysis Conference*, Gaithersburg, USA, 2008.

[ORL02] J.C. Otterbacher, D.R. Radev, and A. Luo. Revisions that Improve Cohesion in Multi-document Summaries: A Preliminary Study. In *Proceedings of the Workshop on Automatic Summarization. A Workshop of ACL-2002*, pages 27–36, Philadelphia, USA, July 2002.

[Par85] C.L. Paris. Description Strategies for Naive and Expert Users. In *Proceedings of the 23rd Annual Meeting on Association for Computational Linguistics*, pages 238–245, Chicago, Illinois, July 1985.

[PB10] M. Potthast and S. Becker. Opinion Summarization of Web Comments. In *Proceedings of the 32nd European Colloquium on IR Research*, pages 668–669, 2010.

[PBMW99] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report 1999-66, Stanford InfoLab, November 1999.

[PL08] B. Pang and L. Lee. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[PLV02] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Philadelphia, USA, July 2002.

[PMD+08] R. Prasad, E. Miltsakaki, N. Dinesh, A. Lee, A. Joshi, L. Robaldo, and B. Webber. The Penn Discourse Treebank 2.0. Annotation Manual.

Technical Report IRCS-08-01, Institute for Research in Cognitive Science, University of Pennsylvania, 2008.

[PNMS05] R.J. Passonneau, A. Nenkova, K. McKeown, and S. Sigelman. Applying the Pyramid Method in DUC 2005. In *Proceedings of the Document Understanding Conference (DUC) Workshop at NAACL-HLT 2005*, Vancouver, Canada, October 2005.

[PZG10] M.J. Paul, C. Zhai, and R. Girju. Summarizing Contrastive Viewpoints in Opinionated Text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 65–75, 2010.

[RABG+04] D. Radev, T. Allison, S. Blair-Goldensohn, J. Blitzer, A. Celebi, S. Dimitrov, E. Drabek, A. Hakim, W. Lam, D. Liu, J. Otterbacher, H. Qi, H. Saggion, S. Teufel, M. Topper, A. Winkel, and Z. Zhang. MEAD -A Platform for Multidocument Multilingual Text Summarization. In *Proceedings of the the 4th International Conference on Language Resources and Evaluation*, pages 1–4, Lisbon, Portugal, 2004.

[Rad04] D.R. Radev. Lexrank: Graph-based Lexical Centrality as Salience in Text Summarization. *Artificial Intelligence Research*, 22:457–479, 2004.

[RHM02] D.R. Radev, E. Hovy, and K. McKeown. Introduction to the Special Issue on Summarization. *Computational Linguistics*, 28(4):399–408, 2002.

[RK08] M. Razmara and L. Kosseim. Concordia University at the TAC-2008 QA Track. In *Proceedings of the Text Analysis Conference*, Gaithersburg, USA, 2008.

[RM98] D.R. Radev and K. McKeown. Generating Natural Language Summaries from Multiple On-Line Sources. *Computational Linguistics*, 24(3):469–500, 1998.

[SB01] J.D. Schlesinger and D.J. Baker. Using Document Features and Statistical Modeling to Improve Query-based Summarization. In *Proceedings of Workshop on Document Understanding Conferences*, New Orleans, LA, 2001.

[SB09] C. Sauper and R. Barzilay. Automatically Generating Wikipedia Articles: A Structure-Aware Approach. In *Proceedings of the Joint Conference of ACL and AFNLP*, pages 208–216, Suntec, Singapore, 2009.

[Sek08] Y. Seki. Summarization Focusing on Polarity or Opinion Fragments in Blogs. In *Proceedings of the Text Analysis Conference*, Gaithersburg, USA, 2008.

[SK08] F. Schilder and R. Kondadadi. FastSum: Fast and Accurate Query-based Multi-document Summarization. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 205–208, 2008.

[SM03] R. Soricut and D. Marcu. Sentence Level Discourse Parsing using Syntactic and Lexical Information. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 149–156, Edmonton, Canada, 2003.

[SWWS07] S. Somasundaran, T. Wilson, J. Wiebe, and V. Stoyanov. QA with Attitude: Exploiting Opinion Type Analysis for Improving Question

Answering in On-line Discussions and the News. In *Proceedings of the International Conference on Weblogs and Social Media*, Boulder, Colorado, USA, March 2007.

[Tab06] M. Taboada. Discourse Markers as Signals (or not) of Rhetorical Relations. *Pragmatics*, 38(4):567–592, 2006.

[Tat91] C. Tattersall. Exploiting Text Generation Techniques in the Provision of Help. In *Proceedings of the Seventh IEEE Conference on Artificial Intelligence Applications*, volume 1, pages 443–449, Miami Beach, FL, USA, 1991.

[VPK+08] V. Varma, P. Pingali, R. Katragadda, S. Krishna, S. Ganesh, K. Sarvabhotla, H. Garapati, H. Gopisetty, V. B. Reddy, K. Reddy, P. Bysani, and R. Bharadwaj. IIIT Hyderabad at TAC 2008. In *Proceedings of the Text Analysis Conference*, Gaithersburg, USA, 2008.

[Wil83] W. Williams. *Composition and Rhetoric*. Heath, Boston, USA, 1983.

[WL11] D. Wang and Y. Liu. Pilot Study of Opinion Summarization in Conversations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 331–339, Oregon, USA, 2011.

[XEN08] Z. Xie, B.D. Eugenio, and P.C. Nelson. From Extracting to Abstracting: Generating Quasi-abstractive Summaries. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008.

[YCWK03] H. Yang, T.S. Chua, S. Wang, and C.K. Koh. Structured use of External Knowledge for Event-based Open Domain Question Answering. In

Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 33–40, Toronto, Canada, 2003.

[YYL+07] J.C. Ying, S.J. Yen, Y.S. Lee, Y.C. Wu, and J.C. Yang. Language Model Passage Retrieval for Question-Oriented Multi Document Summarization. In *Proceedings of the 7th Document Understanding Conference*, pages 72–77, Rochester, New York, USA, 2007.

[ZF11] N.A.H.B. Zahri and F. Fukumoto. Multi-document Summarization Using Link Analysis Based on Rhetorical Relations between Sentences. In *Proceedings of CICLing*, pages 328–338, Tokyo, Japan, 2011.

Appendix A

Sample Summaries Generated by BlogSum

Figure 50: BlogSum-generated Sample Summary 1

Topic: Carmax

Question: What motivated positive opinions of CARMAX from car buyers?

Summary

Sometimes I wonder why all businesses can't be like Carmax. we bought it at Carmax, and I continue to have nothing bad to say about that company. Moreover, Carmax did split the bill (which made me happy). Not sure if you have a Carmax near you, but I've had 2 good buying experiences from them. At Carmax, the price is the price and when you want a car you go get one. have to say that carmax rocks.

Figure 51: BlogSum-generated Sample Summary 2

Topic: Jiffy Lube

Question: What reasons are given for liking the services provided by Jiffy Lube?

Summary

Fortunately, I was right in front of a Jiffy Lube. I know it's fine cause Jiffy Lube sent me a little card in the mail and I have about a month before I need an oil change. When we first got here my bf took his car to jiffy lube cuz we didn't have a jack and its too low to crawl under. When I worked for my last auto shop, our best oil change was 49.99 to the same exact one for jiffy lube was damn near 100 bucks. Well, I suppose it is a little bit of a PITA to figure out what to do with the spent oil, but after some digging, I found out that every Jiffy Lube will take used oil for free! My bet is that Dallas will probably just go back to Jiffy Lube, since I'm not capable of a simple oil change. I had the biotches from Jiffy Lube strip out my oil drain plug! Mom got stuck in the snow, and the jiffy lube people pushed her out.

Figure 52: BlogSum-generated Sample Summary 3

Topic: Jiffy Lube

Question: What reasons are given for not liking the services provided by Jiffy Lube?

Summary

When we first got here my bf took his car to jiffy lube cuz we didnt have a jack and its too low to crawl under. i've had sunglasses stole years ago from a Jiffy Lube - bastards. It seems that the kids at Jiffy Lube screwed the 17mm oil drainage bolt in so hard that it would take a super human to unscrew it, or leverage. Apparently the connectors are tapered in one direction, and when the boys at Jiffy Lube put my new battery in a few months ago, they put them on upside down.

Figure 53: BlogSum-generated Sample Summary 4

Topic: UN Commission on Human Rights

Question: What steps are being suggested to correct this problem?

Summary

I'd be far more interested in the building of a new organization to do the job of the Human Rights Commission than our "withdrawal" in protest. It is quite a bold statement and includes some very substantive proposals to change not only the Security Council, but the Commission on Human Rights as well. The sole purpose of this group is to protect human rights, and the draft says voting members are to "take into consideration" the human rights record of the countries when voting? To ensure this new body would be no facsimile of its predecessor, the legislation prohibits membership to countries that violate human rights or are subject to specific human rights resolutions. An important one dealt with human rights abuses in Darfur. As for "human rights", they simply don't fit in this system. On such countries there are no 'agreed measures currently in place...for human rights violations.'". If Mr. Bolton succeeds, he will have created a rights watchdog that is more respectable and more effective than the current commission. In the past two years, Israel has had 101 human rights resolutions passed against it. Yea, the UN was so supportive that it kicked the U.S. off the Human Rights Commission for the 2002 season, in favor of such lowlifes as Algeria, Burundi, Indonesia, Libya, Saudi Arabia, Sudan, Syria and *Vietnam.* One improvement that could be made to this section of the bill would be to require that member states participating in the human rights bodies at the U.N. be democracies.

Figure 54: BlogSum-generated Sample Summary 5

Topic: Internet Explorer (IE) Firefox Opera Safari Mozila browser

Question: Compare the performance between Internet Explorer (IE), Firefox, Opera, Safari, and Mozila browser regarding speed.

Summary

Overall, Opera seems to be the fastest browser for windows. Mozilla and Firefox give an overall good performance, but their script, cache handling and image-based page speed still cannot compare with Opera. Firefox is not faster than Internet Explorer, except for scripting, but for standards support, security and features, it is a better choice. However, it is still not as fast as Opera, and Opera also offers a high level of standards support, security and features. On Mac OS X , Opera and Safari are both very fast , with Safari 2 being faster at starting and rendering CSS , but with Opera still being distinguishably faster for rendering tables , scripting and history (especially compared with the much slower Safari 1.2) . Camino is fast to start, but then it joins its sisters Mozilla and Firefox further down the list. Neither Mozilla, Firefox nor IE perform very well on Mac, being generally slower than on other operating systems.

Figure 55: BlogSum-generated Sample Summary 6

Topic: Xbox 360

Question: Why do people like Xbox 360 better than PlayStation 3 (PS3)?

Summary:

Xbox 360 : After all is said and done , the Xbox 360 will still be able to produce very awesome graphics thanks to its 3 Power PCs and its unified memory architecture . Xbox 360 : The Xbox 360 will work with my iPod and with my Sony PSP ! Moreover, Xbox 360 : The case design is customizable which I just love and the size of the Xbox 360 is relatively similar to that of the PS2 (first generation) and the PS3 . Xbox 360 : PS3 will have Bluetooth 2 and instead of using a great wireless standard like Bluetooth 2 with support for real-time protocols , Xbox is using FastRF from what I learned from bhpaddock . Xbox 360 : The controllers will pass my wrath this time , but I think the console is ugly as well . Xbox 360 and PS3 : They will both have a free online service that is awesome and just to be expected at this point .

Figure 56: BlogSum-generated Sample Summary 7

Topic: 2008 Nissan Altima car

Questio: What features do people like about the 2008 Nissan Altima.

Summary

I bought a 2008 Nissan Altima, trusting into Nissan. Overall, the new 2008 Altima is much larger, more stylish and much "different" car. I just hope Nissan would make a diesel version of the Altima in the future. There is no doubt in my mind Nissan makes the best performing engines and nissan did an incredible job with it. I think Nissan has a winner with the Altima, I'd definitely recommend to anyone who is looking for more bang for the buck. The new Altima improvements are noticeable and you would think they would be in a more expensive car and i was extremely impressed from the altima the first time I drove it. Moreover, i encourage anyone that is shopping for a safe and comfortable car to test drive the Altima. I am so glad that I decided to go with the altima rather than the accord and i really like this car. Good sporty car,good or 2 people-4 is pushin it with space. Before I saw the 2008 Altima, my mind was set on a Camry, but after seeing and reading about the Altima, my mind changed. Nissan doesn't advertise this model too often, so when people see they are very surprised. I have owned several Nissans most of them being various models of the Z car. I have always been a fan of Nissans, but this car is "the truth and so after spending a couple hours at the dealership working out the details, I drove out being the proud owner of a 2008 Nissan Altima 2.5 S.

Appendix B

Sample Manual Evaluation for Content and Discourse Coherence

Figure 57: Sample Manual Evaluation for Question Relevance and Discourse Coherence

Task Description:
Please evaluate the summary for "Question Relevance" and "Discourse Coherence" separately on a scale of 1 to 5 where "Question Relevance" and "Discourse Coherence" are defined in the following way:

Question Relevance
The summary should answer the user given question and should be informative.

Discourse Coherence
The summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to sentence to a coherent body of information about a topic.

Here is the evaluation scale:

1. Very Poor
2. Poor
3. Barely Acceptable
4. Good
5. Very Good

In this task, for each topic, one or two questions are asked and two summaries on a given question are included for evaluation.

Figure 58: Sample Summaries Distributed for Evaluation Generated by BlogSum and OList

Topic: Affinia Chicago hotel

Question: Why do people like the Affinia Chicago hotel?

Summary1:

Loved Affinia. We booked the Affinia for a 2 night stay during Memorial Day Weekend and from the moment we pulled up to the Front Door the Affinia proved itself over and over to be the best hotel I have ever stayed. Moreover, got a great deal for a weeks stay in the Affinia in Chicago on expedia in October and it was a bargain for the hotel and service we got. Thanks in particular to victor who put up with our midnight antics, that is finding our keys etc, the Affinia should open a hotel in Ireland, we need a great hotel like it and location definitely makes this hotel one I will return to without looking at other hotels. Moreover, in this respect, Affinia is better than your average boutique hotel. The Affinia is a nice hotel with comfortable rooms but a terrible lift (elevator) system . When I'm in Chicago again, I'll be back to stay at the Affinia. My Advice: Book a stay at the Affinia if you find yourself heading to Chicago. Great staff. A colleague and I stayed at the Affinia Chicago on business just because of the reduced price during their remodel. I still wouldn't hesitate to stay at the Affinia if they had a good price and price was excellent compared to hotels within blocks of the Affinia.

Question Relevance	**Discourse Coherence**
1. Very Poor	1. Very Poor
2. Poor	2. Poor
3. Barely Acceptable	3. Barely Acceptable
4. Good	4. Good
5. Very Good	5. Very Good

Summary2:

Loved Affinia. We booked the Affinia for a 2 night stay during Memorial Day Weekend. We had a very enjoyable stay at the Affinia. Thanks to the staff at the Affinia, I can give them FIVE STARS. If you are visiting for a conference at McCormick Place the Affinia does not have a shuttle service but don't worry, there are several hotels next to the Affinia that you can catch. Thanks in particular to victor who put up with our midnight antics, that is finding our keys etc, the Affinia should open a hotel in Ireland, we need a great hotel like it! My wife and I booked our stay at the Affinia Chicago through Perfect Escapes. In this respect, Affinia is better than your average boutique hotel. I refused to pay until they checked with the Affinia. From the moment we pulled up to the Front Door the Affinia proved itself over and over to be the best hotel I have ever stayed. If you're a shopper, there is no better location than the Affinia! My only 3 concerns about the Affinia were the following:1. The Affinia is a nice hotel with comfortable rooms but a terrible lift (elevator) system. Got a great deal for a weeks stay in the Affinia in Chicago on expedia in October and it was a bargain for the hotel and service we got. Otherwise, we were quite pleased and I would definitely stay at the Affinia again on future trips to Chicago.

Question Relevance	**Discourse Coherence**
1. Very Poor	1. Very Poor
2. Poor	2. Poor
3. Barely Acceptable	3. Barely Acceptable
4. Good	4. Good
5. Very Good	5. Very Good

Figure 59: Sample Dataset Distributed for Comparison Predicate Identification for the Manual Performance Evaluation

Instructions:

This file contains comparative and non-comparative sentences. In this task, you have to identify whether a sentence is comparative or non-comparative.

Sample:

I bought 2 of this model for christmas presents and neither one of them work ! (Comparative)

One big difference between the 1201 and the 2600 models is that the 2600 model is virtually silent . (Comparative)

Troubleshooting ad-2500 and ad-2600 no picture scrolling b/w . (Non-Comparative)

Repost from january 13 , 2004 with a better fit title. (Non-Comparative)

++

Please annotate the following sentences to indicate if they are comparative or non-comparative.

++

Or does it play audio and video but scrolling in black and white?

Before you try to return the player or waste hours calling apex tech support, or run the player over with your car, try these simple troubleshooting ideas first.

And does n't need to be placed in a cabinet like the 1201 does.

You might think you are saving money by buying an apex but in the long run you will spend more.

No picture :hopefully you still have the remote control.

This is the best dvd player i 've purchased.

I did not want to have high expectations for this apex player because of the price but it is definitely working out much better than what I would expect from an expensive high-end player.

Without a doubt the finest looking apex dvd player that i 've seen.

If you tossed it out the window , you need to fetch it.

Using the remote control , press the i/p button located on the bottom right corner of the remote.

.....

Figure 60: Sample Dataset Distributed for Attributive Predicate Identification for the Manual Performance Evaluation

Instructions:

In this file, the first line contains the topic (or target) of the sentence, the next line contains the sentence. In this task, you have to identify whether the sentence provides some attributes or describes some functionality of the target.

Sample:

Topic: Picasa

Sentence: Picasa has this bizzare setup where instead of editing the image itself, it edits a copy of it someplace else. (Attributive)

Topic: Zillow

Sentence: Zillow priced the land 30K higher than what I thought it was worth... (Attributive)

Topic: NAFTA

Sentence: Green Party: Renegotiate NAFTA and protect Canadian sovereignty. (Non-Attributive)

Topic: Subway

Sentence: Anyway, I ate subway and all was better. (Non-attributive)

++

Please annotate the following sentences to indicate if they are Attributive or non-attributive.

++

Topic: Carmax

Sentence: Carmax did split the bill (which made me happy). $150This morning, I went out side and discovered my passenger window smashed and the radio stolen.

Topic: Starbucks

Sentence: Not having a Dunkin' Donuts in Sinless City I am obviously missing out... but Starbucks are doing a Christmas Open House today where you can turn up for a free coffee.

Topic: Starbucks

Sentence: Starbucks got some of that market, for sure, but I don't think the grab-a-cup-while-you-get-donuts-or-gas coffee market ever goes away.

Topic: Dunkin Donuts

Sentence: Dunkin Donuts Does Coffee from Business Ethics & Social Enterprise.

…….

Figure 61: Sample Dataset Distributed for Topic-Opinion Predicate Identification for the Manual Performance Evaluation

Instructions:

In this file, the first line contains the topic (or target) of the sentence, the next line contains the sentence. In this task, you have to identify whether the topic of the sentence is associated with any of the subjective term of the sentence.

Sample:

In the sample, for topic-opinion sentences, subjective terms are also shown for participants' understanding.

Topic: Problem

Sentence: Well , its main problem is that it's simply too jumbled. (Topic-opinion. Subjective Term: Simply too jumbled)

Topic: Film

Sentence: The sad part is that the arrow and i both dig on flicks like this , so we actually figured most of it out by the half-way point, so all of the strangeness after that did start to make a little bit of sense , but it still didn't the make the film all that more entertaining. (Topic-opinion. Subjective Term: n't more entertaining)

Topic: Flick

Sentence: Oh , and by the way , this is not a horror or teen slasher flick . . . it's (Non topic-opinion)

++

Please annotate the following sentences to indicate if they are topic-opinion or non topic-opinion.

++

Topic: Plot

Sentence: plot : two teen couples go to a church party , drink and then drive .

Topic: Movie, story.

Sentence: It's got a head start in this movie starring jamie lee curtis and another baldwin brother (william this time) in a story regarding a crew of a tugboat that comes across a deserted russian tech ship that has a strangeness to it when they kick the power back on .

Topic: Action

Sentence: Going for the gore and bringing on a few action sequences here and there , virus still feels very empty , like a movie going for all flash and no substance .

Topic: Acting

Sentence: The acting is below average , even from the likes of curtis .

Topic: movie

Sentence: So , if robots and body parts really turn you on , here's your movie .

…..