SYNTACTIC SENTENCE COMPRESSION FOR TEXT SUMMARIZATION

Paththamestrige Perera

A thesis

in

The Department

of

Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements

For the Degree of Master of Computer Science

Concordia University

Montréal, Québec, Canada

October 2013

Concordia University

School of Graduate Studies

This is to certify that the thesis prepared

By: **Paththamestrige Perera**

Entitled: **Syntactic Sentence Compression for Text Summarization**

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

Master of Computer Science

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

Signed by the final examining committee:

______________________________ Chair

Dr. Volker Haarslev

______________________________ Examiner

Dr. Sabine Bergler

______________________________ Examiner

Dr. Olga Ormandjieva

______________________________ Supervisor

Dr. Leila Kosseim

Approved ______________________________

Chair of Department or Graduate Program Director

____________ 20 ______ ______________________________

Dr. Christopher W. Trueman, Dean

Faculty of Engineering and Computer Science

Abstract

Syntactic Sentence Compression for Text Summarization

Paththamestrige Perera

Automatic text summarization is a dynamic area in Natural Language Processing that has gained much attention in the past few decades. As a vast amount of data is accumulating and becoming available online, providing automatic summaries of specific subjects/topics has become an important user requirement. To encourage the growth of this research area, several shared tasks are held annually and different types of benchmarks are made available. Early work on automatic text summarization focused on improving the relevance of the summary content but now the trend is more towards generating more abstractive and coherent summaries. As a result of this, sentence simplification has become a prominent requirement in automatic summarization.

This thesis presents our work on sentence compression using syntactic pruning methods in order to improve automatic text summarization. Sentence compression has several applications in Natural Language Processing such as text simplification, topic and subtitle generation, removal of redundant information and text summarization. Effective sentence compression techniques can contribute to text summarization by simplifying texts, avoiding redundant and irrelevant information and allowing more space for useful information. In our work, we have focused on pruning individual sentences, using their phrase structure grammar representations. We have implemented several types of pruning techniques and the results were evaluated in the context of automatic summarization, using standard evaluation metrics. In addition, we have performed a series of human evaluations and a comparison with other sentence compression techniques used in automatic summarization. Our results show that our syntactic pruning techniques achieve compression rates that are similar to previous work and also with what humans achieve. However, the automatic evaluation using ROUGE shows that any type of sentence compression causes a decrease in content compared to the original summary and extra content addition does not show a significant improvement in ROUGE. The human evaluation shows that our syntactic pruning techniques remove syntactic structures that are similar to what humans remove and inter-annotator content evaluation using ROUGE shows that our techniques perform well compared to other baseline techniques. However, when we evaluate our techniques with a grammar structure based F-measure, the results show that our pruning techniques perform better and seem to approximate human techniques better than baseline techniques.

Acknowledgments

I will forever be grateful to Dr. Leila Kosseim for giving me the invaluable opportunity to complete my Masters under her supervision. From the task of selecting the specific research topic to finishing my final thesis, the enormous guidance she gave me was incomparable. This thesis would have been an impossible dream for me without her supervision.

I would like to thank the members of my thesis defense committee, Dr. Sabine Bergler and Dr. Olga Ormandjieva who read my thesis and gave me valuable comments and feedback.

I'm also grateful to Dr. Shamima Mithun for providing me with initial guidance in my research work and helping me with the BlogSum summarizer system. Her research work became the root for my thesis work as well.

Also, I would like to offer my sincere gratitude to my colleagues in CLaC lab: Félix-Hervé Bachand, Ishrar Hussain, Reda Siblini and also Dritan Harizaj and Zoe Briscoe who contributed to my research work as human annotators. Their efforts are greatly appreciated. They helped to complete my research work successfully.

I would also like to thank David Gurnsey, who helped me immensely in setting up baseline systems for my research evaluations.

Finally, I wish to dedicate this work to my family for giving me their unconditional support and encouragement to pursue my goal.

Contents

List of Figures **x**

List of Tables **xi**

1 Introduction **1**

1.1 Text Summarization . . . 1

1.1.1 Single Document Summarization . . . 1

1.1.2 Multi-Document Summarization . . . 2

1.1.3 General Text Summarization . . . 2

1.1.4 Opinion Text Summarization . . . 2

1.1.5 Extractive Text Summarization . . . 2

1.1.6 Abstractive Text Summarization . . . 3

1.2 Problem Domain and Research Motivation . . . 3

1.3 Introduction to BlogSum . . . 3

1.4 Focus of Our Work . . . 5

1.5 Evaluation of BlogSum . . . 5

1.6 Sentence Compression for Text Summarization . . . 6

1.7 Automatic and Manual Evaluation of Sentence Compression . . . 8

1.8 Contributions . . . 9

1.8.1 Performance and Error Analysis of BlogSum . . . 9

1.8.2 Syntactic Based Sentence Pruning . . . 10

1.8.3 Analysis of Human Sentence Compression Techniques . . . 10

1.9 Outline of the Thesis 10

2 Evaluation of BlogSum 12

2.1 Performance Comparison Between BlogSum and MEAD 12

2.2 Sentence Organization in BlogSum 17

2.3 Analysis of BlogSum Errors 20

2.3.1 Evaluation Methodology 20

2.3.2 Content Selection Errors 23

Topic Irrelevancy 23

Query Irrelevancy 23

Miscellaneous Errors 24

2.3.3 Sentence Organization Errors 25

Dangling Anaphora 26

Sentence Aggregation Mismatches 26

Insertion of Cue Phrases 28

2.3.4 Summary of the Manual Evaluation Results 28

2.3.5 Readability Measures of BlogSum Summaries 29

2.4 Chapter Summary 31

3 Literature Review 33

3.1 Previous Work 33

3.2 Machine Learning and Classifier Based Techniques 34

3.2.1 Cut and Paste Text Summarization 34

3.2.2 Summarization Beyond Sentence Extraction 35

3.2.3 Decision-Based Model for Sentence Compression 36

3.2.4 Probabilistic Sentence Reduction 38

3.2.5 Integer Linear Programming Approach 38

3.2.6 Maximum Entropy Based Approach 39

3.3 Keyword Based Techniques 40

3.4 Syntax Based Techniques 41

3.5 Chapter Summary 43

4 Syntactic Sentence Compression 44

4.1 Syntax-Driven Sentence Pruning 46

4.1.1 Pruning Relative Clauses 47

4.1.2 Pruning Adjective Phrases 49

4.1.3 Pruning Adverbial Phrases 51

4.1.4 Pruning Conjoined Verb Phrases 51

4.1.5 Pruning Appositive Phrases 54

4.1.6 Pruning Prepositional Phrases 54

4.2 Syntax and Relevancy Based Sentence Pruning 59

4.3 Relevancy-Driven Sentence Pruning 61

4.4 Chapter Summary 63

5 Automatic Evaluation 66

5.1 Evaluation of Compression Rates 66

5.1.1 Syntax-Driven Pruning 67

5.1.2 Syntax and Relevancy Based Pruning 68

5.1.3 Relevancy-Driven Pruning 68

5.2 Evaluation of Readability Measures 69

5.3 Evaluation of Content 71

5.3.1 Syntax-Driven Pruning 71

5.3.2 Syntax and Relevancy Based Pruning 72

5.3.3 Relevancy-Driven Pruning 72

5.3.4 Discussion 73

5.4 Chapter Summary 74

6 Human Evaluation 77

6.1 Human Gold Standard 77

6.2 Baseline Compression Techniques 78

6.2.1 Baseline 1: Random Word Removal 78

6.2.2 Baseline 2: Keyword/Phrase Based Approach 78

6.2.3 Baseline 3: Machine Learning and Classifier Based Approach 82

6.3 Human Compression Rate 83

6.4 Human Pruning of Syntactic Structures 85

6.4.1 Comparison with Our Heuristics 91

6.4.2 Pruning Noun and Verb Phrases 93

6.5 Content Evaluation 94

6.6 Inter-Annotator Content Evaluation 97

6.7 Content Evaluation Based on Grammatical Relations 101

6.7.1 Evaluation Methodology 103

6.7.2 Inter-Annotator Content Evaluation Using Dependency Grammar Structures 105

6.8 Chapter Summary 107

7 Conclusion and Future work 111

7.1 Main Findings of Our Work 113

7.2 Future Work 115

Appendix A Penn Treebank Tagset 127

Appendix B Keyword List 130

Appendix C Sample Compressed Summaries 132

List of Figures

1 Types of Errors Identified in BlogSum and Frequency (Over all Errors). . . 22

2 Phrase Structure for Sentence (28). 45

3 Phrase Structure for Sentence (30). 48

4 Phrase Structure for Sentence (33). 50

5 Phrase Structure for Sentence (36). 52

6 Phrase Structure for Sentence (37). 53

7 Phrase Structure for Sentence (39). 55

8 Phrase Structure for Sentence (49). 59

9 Phrase Structure for Sentence (50). 62

10 Phrase Structure for Sentence (49). 64

11 Proportions of Syntactic Structures Removed by Human Annotators. 87

12 Proportions of Syntactic Structures Removed by Our Syntactic Pruning Techniques. 87

13 Inter-Annotator Content Evaluation Results of ROUGE-2 Scores. 98

14 Inter-Annotator Content Evaluation Results of ROUGE-SU4 Scores. 98

15 Inter-Annotator Content Evaluation Results of Dependency Structure based F-Measure. 108

List of Tables

1 Automatic Evaluation Using ROUGE on the TAC 2008 Dataset. 16

2 Automatic Evaluation Using ROUGE on the DUC 2007 Dataset. 17

3 Automatic Evaluation of BlogSum's Post-Schema Heuristics Using ROUGE on the TAC 2008 Dataset. 19

4 Automatic Evaluation of BlogSum's Post-Schema Heuristics Using ROUGE on the DUC 2007 Dataset. 19

5 Manual Evaluation of Content Selection Errors on the TAC 2008 Dataset. . 29

6 Manual Evaluation of Sentence Organization Errors on the TAC 2008 Dataset. 29

7 Readability Measures of BlogSum Summaries on the TAC 2008 Dataset. . . 30

8 Readability Measures of BlogSum Summaries on the DUC 2007 Dataset. . . 31

9 Sentence Compression Rates of Syntax-Driven Pruning. 67

10 Sentence Compression Rates of Syntax and Relevancy Based Pruning . . . 68

11 Sentence Compression Rates of Relevancy-Driven Syntactic Pruning. 69

12 Syntactic Phrase Structures Removed by Relevancy-Driven Pruning. 69

13 Readability Measures of Different Techniques on the DUC 2007 Dataset. . . 70

14 Readability Measures of Different Techniques on the TAC 2008 Dataset. . . 70

15 Content Evaluation of Compressed Summaries with Syntax-Driven Pruning (Without Filling). 72

16 Content Evaluation of Compressed Summaries with Syntax-Driven Pruning (With Filling). 73

17 Content Evaluation of Compressed Summaries with Syntax with Relevancy Based Pruning (Without Filling). 74

18 Content Evaluation of Compressed Summaries with Syntax with Relevancy Based Pruning (With Filling). 75

19 Content Evaluation of Compressed Summaries with Relevancy-Driven Syntactic Pruning (With and Without Filling). 76

20 Sentence Compression Rates of Annotations 84

21 Sentence Compression Rates of Different Techniques 85

22 Syntactic Structures Removed by Human Annotators. 88

23 Proportion of Syntactic Structures Removed by Human Annotators. 89

24 Proportion of Syntactic Structures Removed by Syntactic Pruning Techniques. 90

25 Sentence Compression by Human Annotators Based on Syntactic Structures. 91

26 Sentence Compression by Human Annotators Based on Syntactic Structures. 92

27 Content Evaluation of Human Annotations. 94

28 Content Evaluation of Compressed Summaries. 95

29 Inter-Annotator Content Evaluation of Compressed Summaries Using ROUGE. 97

30 Content Evaluation of Compressed Summaries Against Human Annotations Using ROUGE F-Measure. 99

31 Dependency Structures for Sentences (64), (64c1) and (64c2). 104

32 Inter-Annotator Content Evaluation of Compressed Summaries Using Dependency Structure Based F-Measure. 106

33 Content Evaluation of Compressed Summaries Against Human Annotations Using Dependency Structure Based F-Measure. 107

Chapter 1

Introduction

1.1 Text Summarization

Text summarization has attracted the interest of the Natural Language Processing research community for the past half century [Luhn, 1958]. Many techniques have been explored and evaluated to improve automatic text summarization. In particular, system evaluation competitions like TREC [Harman, 1992], TAC [Dang and Owczarzak, 2008] and DUC [Harman, 2001] have been held to provide data and benchmark evaluations for automatic text summarization thus providing baselines to compare existing methods in text summarization. A summary can be defined 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" [Radev et al., 2002]. Text summarization tasks can be further categorized into sub tasks based on the number of source documents to be summarized: single document and multi-document summarization, the type of text to be summarized: general text summarization and opinionated text summarization and the methods used in summary generation: extractive summarization and abstractive summarization. Let us give an introduction to these sub-categories:

1.1.1 Single Document Summarization

Single document summarization is the task of providing a shorter summary of a single document which should be considerably shorter than the original document [Mani et al., 1999, Mani and Maybury, 2001, Wan and Yang, 2007]. Early research work in text summarization has mainly focused on single document summarization. The first DUC summarization track, DUC 2001 [Harman, 2001], for example, consisted of a single document summarization task.

1.1.2 Multi-Document Summarization

Multi-document summarization is the task of providing a single summary of two or more documents [Erkan and Radev, 2004, Radev et al., 2004]. Usually multi-document summaries are targeted on a particular given topic and many summarization tracks are designed to summarize a cluster of documents that are relevant to a particular subject/topic. In the DUC summarization conference, the task of multi-document summarization was introduced in the 2002 [Hahn and Harman, 2002] and most of the summarization tasks are now focused on multi-document summarization as opposed to single document summarization.

1.1.3 General Text Summarization

General text summarization is the task of summarizing texts which comes from sources such as news, scientific and technical articles, books etc. [Hovy and Lin, 1998]. Earlier summarization tracks consisted of document collections created from these sources and the summarization tracks were focused on extracting topic or query relevant information to create related summaries.

1.1.4 Opinion Text Summarization

With the growth of texts available on the world wide web, now there is a large availability of opinionated text. Opinionated text can be defined as texts that express a particular opinion about a topic/subject by an individual person or a group of people [Liu et al., 2005]. Opinionated text summarization was introduced in the TAC 2008 [Dang and Owczarzak, 2008] summarization track (described in Section 2.1) and many techniques have been developed to summarize opinionated text.

1.1.5 Extractive Text Summarization

Extractive summarization can be defined as the task of extracting relevant sentences based on a scoring mechanism and rank them to select the most relevant sentences to create a summary. In extractive summarization, the sentences are extracted often according to a given subject, topic or query and are ordered to generate a meaningful summary [Gupta and Lehal, 2010]. In this task, in most cases, sentences are not processed and they appear as they did in the original texts

1.1.6 Abstractive Text Summarization

In abstractive summarization, the meaning of sentences is somehow extracted and combined to generate the final summary [Radev and McKeown, 1998, Ganesan et al., 2010]. These summaries, thus, may not contain the original sentences. Abstractive summarization is still at the research level as most of the current techniques used in abstractive summarization need to be improved to produce a fluent, well formulated summary.

1.2 Problem Domain and Research Motivation

BlogSum is an automatic summarization system that was developed at the CLAC research lab [Mithun, 2010]. BlogSum was developed mainly to summarize opinionated texts by identifying stated opinions in texts and organizing the selected sentences using discourse schemata. The focus of our work was first to perform a thorough evaluation of the performance of BlogSum using various parametric measures to identify in which areas BlogSum can be improved. The section below will describe our evaluations and findings and finally, present our motivation on applying sentence compression to extractive summarization.

1.3 Introduction to BlogSum

The project BlogSum was initiated and developed to improve opinionated summary generation. Opinionated text summarization is more challenging compared to other types of text summarization tasks such as news data summarization [Mithun, 2010, Ganesan et al., 2010]. Unlike summarization of general text corpora, opinionated summarization should be focused on identifying the nature of the opinions stated in the texts. In opinionated summarization, for a given topic and a query which inquires about a specific opinion, an opinionated summary needs to be formulated by the automatic summarization system. As an example, consider the following query and the two extracted sentences.

Query: *Why do people like Picasa?*

(1) *Picasa is another Google product, that is almost enough to make it great and i really like Picasa as an image organizer application.*

(2) *But this 'Hello' software from Picasa is worse than useless.*

Here, the query asks for positive opinions about the product *"Picasa"*. Though the given two sentences are appropriate in terms of topic relevance, they represent contrasting opinions about the product *Picasa*. So the first sentence is appropriate to answer the given query as it represents a positive opinion about the product, but the second will not be as it represents a negative opinion.

The above example illustrates an important reason why generic summarization systems would not perform well on opinionated queries as their extraction process mainly focuses on query relevance and topic relevance of sentences. BlogSum takes into account the opinion expressed in a particular sentence and uses a weighting mechanism to rank the relevant sentences to be included in a summary. It improves summary relevance by identifying the stated opinions of sentences and selecting relevant extractive sentences to match the given query. Also to improve coherence, BlogSum identifies rhetorical relations between sentences and orders them to generate more coherent summaries. BlogSum uses sentence polarity, rhetorical relations, and discourse schemata to achieve this task.

With these approaches, BlogSum has shown to improve the extractive summary generation task for opinionated texts. The system has been tested for its performance using various text corpora and summarization tasks such as the Text Analysis Conference (TAC) 2008 [Dang and Owczarzak, 2008] data for summary contents, the Document Understanding Conference (DUC) 2007 [Copeck et al., 2007] and the OpinRank Review Dataset [Ganesan and Zhai, 2012].

1.4 Focus of Our Work

The focus of our work was first to conduct a series of experiments to evaluate the performance and the quality of the summaries generated by BlogSum. For these evaluations, we used two types of criteria: automated and manual evaluations. Based on the results, we identified key areas where the BlogSum system could be improved. In particular, BlogSum's extracted sentences are syntactically complex. In order to improve text coherence further, we felt that compressing these sentences was a necessity. Hence, the rest of our work focused on sentence compression.

1.5 Evaluation of BlogSum

In the research area of automatic summarization, opinionated text summarization is becoming more important than ever since a vast amount of information containing opinionated text is becoming more and more available. The sources of opinionated texts include the Blogosphere, media sharing sites and product/service reviews. In the first series of experiments, we performed a comparison of BlogSum with a well known summarization system, MEAD [Radev et al., 2004] and showed that BlogSum could perform far better in opinionated text summarization compared to the generic text summarizer MEAD. Also it could perform just as well as MEAD on query based general text summarization.

In the second series of evaluations, we performed a human evaluation, identifying the most frequent types of errors and their frequencies in the summaries generated using BlogSum. With that, we identified that not only BlogSum can be improved in content selection but also on sentence organization as well. Specifically, it was apparent that more comprehensive methods are needed to improve sentence aggregation and introduction of cue phrases. Additionally, as we saw, many text corpora available now are created from news articles or from blogs which contain more complex sentences. So in order to implement better sentence organization schemata, there is a necessity to simplify complex sentences before they are aggregated as we will see in Chapter 2. This would improve the summary readability and hence improve the quality of extractive summary generation systems. The rest of our work therefore focused on sentence compression for text summarization.

1.6 Sentence Compression for Text Summarization

Sentence compression has several practical applications in natural language processing tasks such as text simplification [Chandrasekar et al., 1996], headline generation [Dorr et al., 2003] and text summarization [Knight and Marcu, 2002]. The goal of automatic text summarization is to produce a shorter version of information from a text repository and produce a meaningful summary. The sentence extraction process is generally based on assigning a score to sentences according to a given topic/query similarity [Murray et al., 2008] or some other parametric score calculated to determine how important is a given sentence to produce a summary. In extractive summarization, a particular sentence is not modified when included in the summary. Summaries are usually generated considering a word or sentence limit and within these limits, the challenge is to extract and include as much relevant information as possible. Moreover, since the sentences are not processed or modified, they may consist of partially irrelevant information. Here, what partially irrelevant information means is that a sentence could contain words or phrases which may be irrelevant or may not contribute to the targeted summary. The relevance is mostly determined by the topic and the query given for the targeted summary. As an example, consider the following topic, query and the sentence (3), extracted as a relevant sentence.

Topic: *Southern Poverty Law Center*

Query: *Describe the activities of Morris Dees and the Southern Poverty Law Center*

(3) *Since co-founding the Southern Poverty Law Center in 1971, Dees has wielded the civil lawsuit like a buck knife, carving financial assets out of hate group leaders who inspire followers to beat, burn and kill.*

As we see, the candidate sentence has 35 words (without punctuations) and may be considered too long to fit into a summary. However, it also contains several phrase structures that are less relevant considering the topic and the query given. So a few possible short forms or compressed forms of the above sentence can be :

(3c1) *Since co-founding the Southern Poverty Law Center ~~in 1971~~, Dees has wielded the civil lawsuit like a buck knife, carving financial assets out of hate group leaders who inspire followers to beat, burn and kill.*

(3c2) *~~Since co-founding the Southern Poverty Law Center in 1971,~~ Dees has wielded the civil lawsuit like a buck knife, carving financial assets out of hate group leaders who inspire followers to beat, burn and kill.*

(3c3) *~~Since co-founding the Southern Poverty Law Center in 1971,~~ Dees has wielded the civil lawsuit like a buck knife, carving financial assets out of hate group leaders ~~who inspire followers to beat, burn and kill~~.*

In the given simplified sentences, we have tried to remove particular phrases which we considered not relevant while retaining the main content. The result is a shorter simplified sentence which may contribute to a better summary. Overall, sentence compression can contribute to text summarization for the following reasons:

1. Simplify complex sentences, hence improve readability.
2. Remove redundant and irrelevant information within sentences.
3. Preserve space for more useful information for length-limit summaries.

We have developed three techniques for sentence compression: one based solely on syntactic structures, one based on syntactic structures and a relevance measure and the third one based solely on the relevance measure. Chapter 4 will present the three techniques we have developed to compress sentences.

1.7 Automatic and Manual Evaluation of Sentence Compression

Sentence compression techniques are used in several areas of NLP including automatic text summarization. Previous work has proposed sentence compression as a part of their automatic text summarization methods. However, very few research work has performed a thorough extrinsic evaluation of sentence compression as a part of automatic summarization. Several automatic evaluation techniques have been introduced in the past years for text summarization, but the predominant one is called Recall Oriented Gisting Evaluation (ROUGE), which evaluates summary content against gold standard summaries. In our work, we have evaluated our sentence compression techniques extrinsically using two measures: ROUGE and compression rate. We have evaluated the compression rates we could achieve using each technique and have evaluated how content was affected when sentences were simplified and more sentences were added to the summary. Our evaluations and findings are presented in Chapter 5.

In most previous work, sentence compression was evaluated by a human evaluation. Here, the compressed sentences were either compared against human compressed sentences or automatically compressed sentences were given to human judges to rate them against baseline compression techniques [Knight and Marcu, 2002, Le Nguyen et al., 2004]. In our research work, we were also motivated to evaluate our sentence compression techniques against human judgment. In this work, we have given a set of summaries to human annotators and asked them to compress these summaries to be used as gold standard summaries for our evaluations. We have analyzed these human compressed summaries and evaluated our techniques against human compressed summaries to see to what extent our techniques are similar to human compressions. In addition, we have compared three baseline techniques alongside our techniques and presented our findings and conclusions in Chapter 6.

Our results show that our syntactic pruning techniques achieve compression rates that are similar to previous work and also with what humans achieve. However, the automatic evaluation using ROUGE shows that any type of sentence compression results in a decrease in content compared to the original summary and extra content addition does not show a significant improvement in ROUGE. The human evaluation shows that our syntactic pruning techniques remove structures that are similar to what humans remove and inter-annotator content evaluation using ROUGE shows that our techniques perform well compared to other baseline techniques. However, when we evaluate our techniques with a grammar structure based F-measure, the results show that our pruning techniques perform better and seem to approximate human techniques better than baseline techniques.

1.8 Contributions

The following sections describe our contributions to the fields of automatic summarization and sentence compression.

1.8.1 Performance and Error Analysis of BlogSum

We have performed a performance comparison of BlogSum and MEAD based on content evaluation. With this evaluation, we show that generalized summarization techniques are not adequate for the task of opinionated summarization. The result of this work was published in [Mithun et al., 2012]. In addition, we have performed an error analysis of automatic summarization using the BlogSum summarizer. In this analysis, we categorized the errors we have found in BlogSum summaries. With this analysis we have shown that existing automatic evaluation metrics such as ROUGE cannot evaluate abstractive summarization techniques like sentence aggregation and sentence planning. Also we have found that sentence aggregation techniques are not effective in extractive summarization when applying these techniques on complex sentences.

1.8.2 Syntactic Based Sentence Pruning

We have experimented on sentence simplification through syntactic based sentence pruning as a part of automatic summarization. In this work, we have introduced three different techniques: syntax-driven, syntax with relevancy based and relevancy-driven sentence pruning. We have defined syntactic pruning heuristics, identifying sentence structures that can be removed to simplify complex sentences. In addition, we have used the given topic/query in automatic summarization to introduce a relevance filter to tone down our syntactic pruning heuristics to preserve important content in topic and query driven summarization. Lastly, we have used the relevance as our main objective and remove syntactic structures to simplify sentences in automatic summarization. We have performed a thorough evaluation of our approaches and conclude that the automatic evaluation metric ROUGE may not be a suitable evaluation method to evaluate the effectiveness of sentence simplification with respect to automatic summarization. This work was presented in [Perera and Kosseim, 2013].

1.8.3 Analysis of Human Sentence Compression Techniques

We have performed a manual evaluation of human compressed summaries. Here, we have asked five human annotators to simplify sentences that belong to a set of automatically generated summaries, with respect to the given topic and query pairs. We have analyzed the summaries and have shown that human give priority in preserving grammaticality in sentence compression and also humans tend to remove syntactic structures to achieve sentence compression through word deletion. Also we have shown that our syntactic pruning techniques approximate well what human annotators do and we have evaluated inter-annotator content evaluations using ROUGE and a dependency structure overlapping F-measure.

1.9 Outline of the Thesis

The remainder of the thesis is organized as follows: In Chapter 2, we present our evaluation of the BlogSum summarizer and conclusions on which areas it can be improved. In Chapter 3, we introduce sentence simplification through sentence compression and emphasize the previous work that has been done on sentence compression. Next, in Chapter 4, we discuss our approach to sentence compression. In Chapter 5, we present our automatic evaluation performed on our sentence compression techniques. Chapter 6 is dedicated to our work on evaluating our sentence compression techniques with a series of human evaluations and comparing our results with other sentence compression techniques. And finally, Chapter 7 will present a summary of our work and directions for future work.

Chapter 2

Evaluation of BlogSum

The goal of our evaluation of BlogSum was to analyze its performance in order to identify its strengths and weaknesses. This evaluation was two-fold: first, we performed a comparison with a well established summarizer (described in Section 2.1) and a manual analysis of its errors (see Section 2.3).

2.1 Performance Comparison Between BlogSum and MEAD

Today, automatic text summarization has reached a point where several systems are now publicly available. The MEAD project [Radev et al., 2004] is a multi-document summarizer that is now available and used in research work as well as in commercial work. It was initiated at the University of Michigan in the year 2000 and participated successfully in several summarization benchmarks. MEAD was developed with a set of features that users can combine to create different types of summaries. Some of these features are position-based scores, centroid based scores, *tf*idf* and query-based algorithms. In the first part of our experiments, we wanted to evaluate how well BlogSum can perform compared to MEAD in opinionated text summarization and to what extent BlogSum is reliable in generic text summarization compared to MEAD. Also in these experiments, we have performed ablation tests (i.e. enable/disable various features in both BlogSum and MEAD) in order to measure the contribution of those features. For our experiments we have chosen two datasets from two different genres: the DUC-2007 [Copeck et al., 2007] and TAC-2008 [Dang and Owczarzak, 2008] datasets.

Document Understanding Conference (DUC) 2007 The Document Understanding Conference was series of conferences, held from 2001 to 2007, with the motive of furthering the progress of research in automatic text summarization. The conference series was ran by the National Institute of Standards and Technology (NIST) and each year they have provided a text collection and a track in automatic summarization. The DUC 2007 dataset [Copeck et al., 2007] was created from the AQUAINT corpus [Graff, 2002], which consists of newswire articles from the Associated Press, the New York Times (1998-2000)[1] and the Xinhua News Agency (1996-2000)[2]. The dataset contains 1126 documents and 45 topic/query narrative pairs for automatic summarization task.

Text Analysis Conference (TAC) 2008 The Text Analysis Conference is an ongoing series of conferences, which provide infrastructure for large-scale evaluations of Natural Language Processing technology. TAC provides several different tracks on different applications of NLP and one of them is automatic text summarization. TAC 2008 [Dang and Owczarzak, 2008] provided a track for opinionated text summarization using a dataset consisting of texts extracted from blogs. This dataset was created using the TREC Blogs06 Collection [Macdonald and Ounis, 2006] and consists of 1893 documents, 50 topics and 75 queries.

Using these two datasets, we have generated two sets of summaries and evaluated these using the automatic evaluation metric ROUGE [Lin, 2004].

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) ROUGE [Lin, 2004] is an automatic evaluation metric designed to evaluate automatically generated summaries

[1] http://www.nytimes.com/pages/aponline/news/index.html

[2] http://www.chinavitae.com/institution/PRS/7050.001

based on their content. The methodology of the ROUGE evaluation is to calculate the percentage of n-gram co-occurrences between an automatically generated summary and a gold standard summary. A gold standard summary is the summary against which automatic summaries will be compared in order to evaluate and rank automatic summary generation techniques. In most of the summarization tracks and ROUGE evaluations, human written summaries have been used as the gold standard summaries. The ROUGE measure was introduced in the Document Understanding Conference (DUC) 2004 as a part of their automatic evaluation metrics. The ROUGE evaluation package comes with several types of n-gram matching criteria; but in automatic summarization, the most frequently used models are ROUGE-2 and ROUGE-SU4. ROUGE-2 measures bi-gram co-occurrences between the model summary and the automatic summary while ROUGE-SU4 measures the skip-bigram co-occurrences with maximum gap length of 4. These measures calculate the precision, recall and F-score per each summary based on these models. As an example, let us take the following pairs of gold standard sentences/summary and extracted summary sentences.

> Gold standard summary: *"In 1998 the US Department of Interior announced it would remove 2,500 wolves in Minnesota, Michigan and Wisconsin from the Endangered Species Act.*
> *Smaller populations in Montana, Wyoming and Idaho would be reclassified from endangered to threatened."*

> Machine extracted summary: *"The small population of Mexican gray wolves recently introduced to parts of New Mexico and Arizona – only 22 wolves – would remain endangered because they continue to be under the threat of extinction, officials said. Jamie Rappaport Clark, director of the U.S. Fish and Wildlife Service, said the recovery of the gray wolf – also known as the timber wolf in some parts of the country – was an endangered species success story."*

One of the common configurations of ROUGE is to stem the words [Porter, 1997] prior to the evaluation and then count the n-grams and n-gram overlapping statistics. In this example, we will calculate ROUGE-2 score for the machine extracted summary. ROUGE-2 calculates the bi-gram co-occurrences where a bi-gram is defined as a sequence of two adjacent elements in a string of tokens, in this case words. Below are a few sample bi-grams from the above machine extracted summary.

Bi-grams: "*The small*", "*small population*", "*population of*", "*of Mexican*", "*Mexican gray*",

According to the ROUGE metric, for the gold standard sentences and machine extracted sentences above, we have:

Total bi-gram count for the gold standard summary: 37

Total bi-gram count for the extracted summary: 70

Total co-occurring bi-grams: 2

So with these counts, for ROUGE-2, the precision, recall and F-measures are calculated as follows:

$$Precision = \frac{2}{70} = 0.0286$$

$$Recall = \frac{2}{37} = 0.0540$$

$$F - Measure = \frac{2 \times 0.0286 \times 0.0540}{0.0286 + 0.0540} = 0.0374$$

The MEAD system uses several features in summary generation [Otterbacher et al., 2003]. They are:

- *Centroid*: The centroid score quantifies the extent to which the sentence contains lexical items that are key to the overall cluster of documents.

- *QueryTitleCosine*: The cosine of the vectors representing the title portion of the query for the cluster and the sentence.

- *QueryNarrativeCosine*: The cosine of the vectors representing the narrative portion of the query for the cluster and the sentence.

For content evaluation, we used BlogSum generated summaries and two configuration of the MEAD system: MEAD with Query (i.e. MEAD uses the given query to compute sentence relevance), MEAD without Query (i.e. MEAD does not use the given query to compute sentence relevance). Other than the query features (QueryTitleCosine, QueryNarrativeCosine features), we used other general features (Centroid feature etc.) in MEAD to generate the summaries. On the other hand, BlogSum uses topic/query relevance as a main feature in content selection. So in order to be fair to MEAD, we evaluated it with and without this feature. We calculated ROUGE F-scores for the generated summaries of these three systems and Tables 1 and 2 show the obtained results: Table 1 shows that BlogSum outperforms the MEAD system in opinionated text summarization. The following results can be explained due to the fact that the MEAD summarizer does not incorporate features to identify opinionated text and to generate relevant summaries in a query based summarization task. On the other hand, BlogSum takes the opinionated nature of text into account and filters out particular sentences according to the given query. So BlogSum outperforms MEAD with respect to opinionated text summarization according to the ROUGE scores. We have also used the DUC 2007 dataset to generate summaries for our next evaluation. We have generated summaries according to DUC 2007 summarization track using BlogSum and MEAD and calculated ROUGE scores for a comparison. Table 2 shows the obtained results. According to Table 2, MEAD performs marginally better than BlogSum in generalized query based summarization. With these results we can conclude that BlogSum is

System	**ROUGE-2**	**ROUGE-SU4**
BlogSum	0.076	0.110
MEAD With Query	0.053	0.081
MEAD Without Query	0.039	0.060

Table 1: Automatic Evaluation Using ROUGE on the TAC 2008 Dataset.

System	**ROUGE-2**	**ROUGE-SU4**
BlogSum	0.086	0.140
MEAD With Query	0.088	0.140
MEAD Without Query	0.080	0.130

Table 2: Automatic Evaluation Using ROUGE on the DUC 2007 Dataset.

more effective in opinionated summary generation and performs just as well as in generalized query based summarization compared to the MEAD summarizer.

2.2 Sentence Organization in BlogSum

Once the overall performance in terms of content selection of BlogSum was evaluated, we wanted to perform ablation tests to evaluate the contribution of specific features. In order to improve extractive summarization, BlogSum performs sentence reordering and sentence aggregation based on several post-schema heuristics. These sentence aggregation techniques include introducing cue phrases to connect sentences and improve the readability of the summaries. BlogSum uses four post-schema heuristics to determine whether sentence aggregation is possible or not. These heuristics can be categorized as follows:

(1) Sentences with neutral polarity are filtered out or taken as individual sentences.

(2) Sentences with similar polarities are aggregated using conjunctions (ex. *and*). For example, consider the two sentences below:

(4) *Picasa is another Google product, that is almost enough to make it great.*

(5) *I really like Picasa as an image organizer application.*

If these two sentences should appear consecutively in the summary according to the schemata; since they both have a positive polarity, they are aggregated as follows.

(6) *Picasa is another Google product, that is almost enough to make it great* ***and*** *I really like Picasa as an image organizer application.*

(3) Sentences with opposing polarities are aggregated using appropriate conjunctions (ex. *But*).

(4) When a sentence aggregation is performed, the next following sentence is further combined adding appropriate cue phrases (ex. *Moreover*).
For example, consider following three sentences:

(7) *Picasa is another Google product, that is almost enough to make it great.*

(8) *I really like Picasa as an image organizer application.*

(9) *one thing that I really like in Picasa is the ability to watch offline folders.*

Consider the given order of sentences after they were ordered by the discourse schema. In this scenario, all three sentences have a positive polarity hence the first two sentences will be aggregated with the conjunction *and* and the third sentence will be connected with the cue phrase *moreover*. So the result will be:

(10) *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.*

These post-schema heuristics are simple and based only on the polarity of each sentence. All sentence aggregations are performed after the sentences are fitted into a specific schema (for more details see [Mithun, 2010]).
Apart from applying sentence aggregation, BlogSum also reorders sentences within schemata to improve coherence of the summaries. Here, BlogSum uses two heuristics to improve coherence within a discourse schema: The similarity between sentences, the relative distance between two sentences in the original text, if they were extracted from the same document. These heuristics are used in order to group similar sentences together. So with these heuristics, the sentences are re-ordered inside a discourse schema to make the summaries

more coherent.

As the next step in our evaluation of BlogSum, we have performed an ablation test of these post-schema heuristics. We calculated ROUGE-2 and ROUGE-SU4 measures on BlogSum summaries which have been generated with these sentence organization features and summaries generated without these sentence organization features. The results we obtained are shown in Tables 3 and 4. For this, we have used three configurations of the BlogSum summarizer:

(1) BlogSum complete : BlogSum with all four post-schema heuristics.

(2) BlogSum without aggregation : BlogSum without any of the four post-schema heuristics. Here, we disabled all sentence aggregation heuristics.

(3) BlogSum without reordering and aggregation : Here, in addition to disabling the post-schema heuristics, we also disabled the sentence reordering feature within a discourse schema. By disabling both reordering and aggregation features, we wanted to evaluate the BlogSum only based on content selection and discourse schema features.

System	**ROUGE-2**	**ROUGE-SU4**
BlogSum: Complete	0.076	0.110
BlogSum: No Aggregation	0.076	0.110
BlogSum: No Aggregation & No Reordering	0.076	0.110

Table 3: Automatic Evaluation of BlogSum's Post-Schema Heuristics Using ROUGE on the TAC 2008 Dataset.

As Table 3 and 4 show, sentence organization in BlogSum has not significantly improved the content of the summaries as measured by ROUGE. However the post-schema heuristics are meant to improve summary readability which is not measured by ROUGE. After evaluating

System	**ROUGE-2**	**ROUGE-SU4**
BlogSum: Complete	0.086	0.140
BlogSum: No Aggregation	0.081	0.130
BlogSum: No Aggregation & No Reordering	0.081	0.130

Table 4: Automatic Evaluation of BlogSum's Post-Schema Heuristics Using ROUGE on the DUC 2007 Dataset.

summaries with the automatic measures, we therefore proceeded to perform a manual analysis of BlogSum's errors.

2.3 Analysis of BlogSum Errors

Our next analysis consisted of identifying and quantifying the different types of errors in BlogSum summaries. In order to do this, we performed a manual analysis of BlogSum summaries. This implied reading through each summary and identifying and categorizing each error in addition to counting them and calculating the percentages of errors.

2.3.1 Evaluation Methodology

First we have selected an opinionated text corpus and generated summaries using BlogSum. For this task we used the TAC 2008 dataset. We have generated a set of 75 summaries based on the topics and the queries given at the TAC 2008 summarization track. We used 50 different topics and 75 different queries to generate these summaries. As the next step, each summary was read by a human annotator (myself) and various types of errors were identified and counted.

Each summary was considered and each sentence in that summary was judged considering different types of errors. There were 752 sentences in all 75 summaries. Some sample sentences and their annotated errors are shown below.

Example 1:

Topic: *Yojimbo*

Query: Describe the reasons given by bloggers for their positive opinions of Yojimbo.

(11) Extracted sentence: *Fight: Yojimbo vs Stickybrain.*

Errors:

(a.) Query irrelevancy

(b.) Fragmented sentence

Example 2:

Topic: *Frank Gehry*

Query: What complaints are made concerning his structures?

(12) Extracted sentence: *Frank is wearing a Roots Team Canada jacket to support Wanye Gretzky against the gambling allegations* ***and*** *diller, gehry: both idiots.*

Errors:

(a.) Query irrelevancy

(b.) Fragmented sentence

(c.) Sentence aggregation mismatch

As shown in the examples above, a single sentence can contain several types of errors. Once all the error categories were identified and counted, we computed the following two types of figures for each error type.

$$Error_{\ e}\ [total\ sentences]\ = \frac{Frequency\ of\ error\ e}{Total\ number\ of\ sentences\ in\ all\ summaries}$$

$$Error_{\ e}\ [total\ errors]\ = \frac{Frequency\ of\ error\ e}{Total\ number\ of\ all\ errors\ found\ in\ all\ summaries}$$

For example, the error of type Fragmented Sentence appeared 53 times, the total number of sentences is 752 and we have counted 788 in total. So for the error Fragmented Sentence, we have:

$$Error_{\ fragmented\ sentence}\ [total\ sentences]\ = \frac{53}{752} = 7.0\%$$

$$Error_{\ fragmented\ sentence}\ [total\ errors]\ = \frac{53}{788} = 6.7\%$$

The result of this experiment is the error classification shown in Figure 1. As Figure 1 shows, we categorized the errors into two broad categories: content selection errors and sentence organization errors (which can only be measured manually). Let us describe the error types in detail.

2.3.2 Content Selection Errors

We have categorized the errors that occur when selecting relevant sentences in automatic summary generation as content selection errors. This category of error accounts for half of the errors we have identified and can be further divided into three sub categories: topic irrelevancy, query irrelevancy and other errors.

Topic Irrelevancy

Topic irrelevancy is defined as the error of selecting irrelevant sentences with respect to the given summary topic. One cause for this error is polysemy: when a topic contains a keyword that may refer to several meanings. The following extracted sentence shows an example of topic irrelevancy.

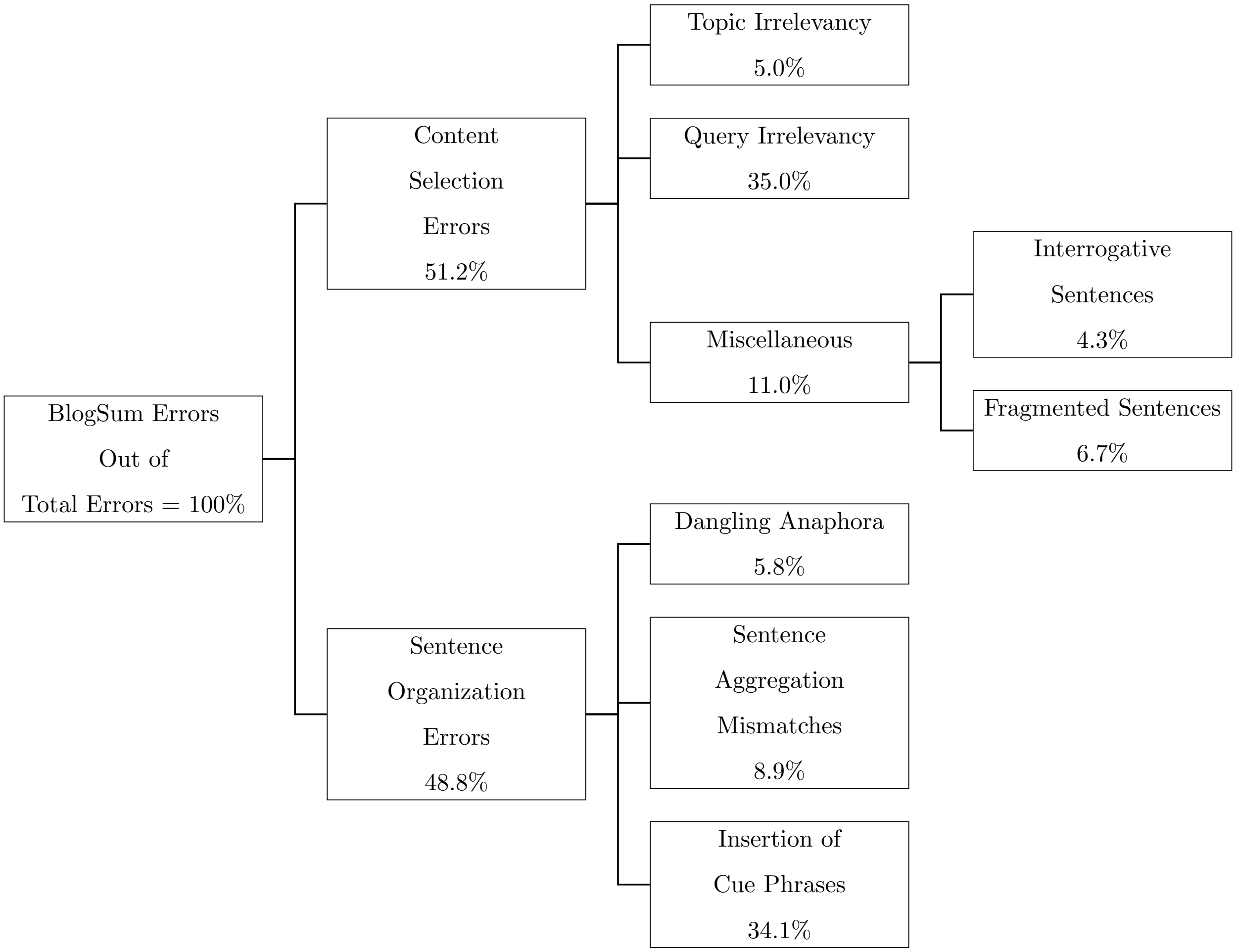

Figure 1: Types of Errors Identified in BlogSum and Frequency (Over all Errors).

Topic: *Edward Norton*

(13) Extracted sentence: *You will find Norton Internet security 2004 Keygen right here Norton Internet security 2004 Keygen and if you have Windows XP, there is almost no reason I can think of for you to have Norton Firewall or Internet Security or any firewall software from any other software manufacturers.*

In this example, the term *"Norton"* is polysemic. In the topic, it refers to a person named *"Edward Norton"* but in the text, it refers to a product name of a popular Internet security ware *"Norton Anti-virus"*. As Figure 1 shows, topic irrelevancy account for 5% of all errors (over all errors).

Query Irrelevancy

Query irrelevancy can be defined as the error of selecting irrelevant sentences with respect to the given query. In BlogSum, this type of error mainly occurs for two reasons. One is selecting irrelevant sentence in terms of content with respect to the given query and the second is selecting sentences that contains an opposing opinion compared to the given query. The following queries and extracted sentences show examples of these two types of query irrelevancy errors.

Query: *Why do people like Starbucks better than Dunkin Donuts?*

(14) Extracted sentence: *No Starbucks isn't at all like McDonalds*

Here, the query asks for a comparison between *"Starbucks"* and *"Dunkin Donuts"*, but the extracted sentence talks about *"Starbucks"* and *"McDonalds"*. So this is a scenario where a partial query match occurred and selected an irrelevant sentence for the automatic summary generation.

Query: *Why don't people like eating at Chipotle?*

(15) Extracted sentence: *One of the reason I like Chipotle is their willingness to use organic, free-range meat and Chipotle also deserves credit for the quality of meat it uses.*

The second example shows the case of not correctly identifying the polarity of the stated opinion of the sentence. The query asks for a negative opinion about the restaurant chain *"Chipotle"* but the extracted sentence contains a positive opinion about the restaurant *"Chipotle"*. Overall, query irrelevancy accounts for 35% of all BlogSum errors.

Miscellaneous Errors

There were other types of errors that occurred infrequently in BlogSum generated summaries and we categorized them all as miscellaneous errors. These errors include mainly the extraction of sentence fragments and interrogative sentences. The following examples show different types of miscellaneous errors.

Error type : Extraction of interrogative sentences

Topic: *Subway*

(16) Extracted sentence: *what's wrong with Subway in Manhattan?*

Here, the summary includes an interrogative type of sentence with relevance to the given topic. This type of sentence will not provide a clear opinion and because of that, they rarely contribute to a meaningful summary. The following example shows a case of including a sentence fragment into a summary. Here, again the sentence is relevant with respect to the given topic but it does not contribute to a meaningful summary.

Error type : Fragmented sentences

Topic: *Yojimbo*

(17) Extracted sentence: *Fight: Yojimbo vs Stickybrain.*

2.3.3 Sentence Organization Errors

The second main type of errors (see Figure 1) are sentence organization errors. These errors reduce the readability and the coherence of extractive summaries and are difficult to measure using automatic metrics such as ROUGE. The categorization of these types of errors is more subjective as it depends on human judgment of the quality of a summary. This category of errors can be further divided into the following three sub-categories of errors.

1. Dangling Anaphora.
2. Sentence Aggregation Mismatches.
3. Insertion of Cue Phrases.

As discussed earlier, there is no effective automated evaluation to identify and count the above mentioned errors so they are much harder to detect.

Dangling Anaphora

Dangling anaphora is defined as failure to extract the anaphoric antecedent from sentences or failure to preserve relevant anaphoric relations while performing sentence organization within a summary. The following example shows an anaphoric error within an extractive summary.

(18) Extracted sentence: ***He*** *then sat there and tried to argue that Wikipedia was a completely valid source and for these topics, Wikipedia had slightly more errors than Britannica.*

In this summary, the antecedent of the pronoun *"he"* was not extracted. In order to generate a meaningful summary, the antecedent of *"he"* needs to be resolved somehow (for example, by extracting the preceding sentence or a few preceding sentences before the following sentence) [Paice and Husk, 1987, Hirst, 1981]. As shown in Figure 1, dangling anaphora errors occur about 6% of the time in BlogSum.

Sentence Aggregation Mismatches

Sentence aggregation errors are defined as errors or mistakes which appear as a result of aggregating sentences. One of the causes for these types of errors is aggregating sentences that should not have been aggregated or sentences that were aggregated improperly by BlogSum. Another cause of error is failing to aggregate sentences which should have been considered as better candidates for aggregation by BlogSum. This cause of error is much harder to identify so it was not considered in our analysis. The following examples illustrate sentence aggregation mismatches.

Example 1:

> (19) Extracted Sentence: *Saudi Arabia has agreed to allow women to attend a football match against Sweden, reversing an earlier decision.*
>
> (20) Extracted Sentence: *If Denmark is producing the butter that is partly responsible for the health problems of many Saudis (whose own dietary choices are the true culprit), then the same Saudis are turning for Danish medicines for treatment.*
>
> (21) Aggregated sentence: *Saudi Arabia has agreed to allow women to attend a football match against Sweden, reversing an earlier decision* ***and*** *if Denmark is producing the butter that is partly responsible for the health problems of many Saudis (whose own dietary choices are the true culprit), then the same Saudis are turning for Danish medicines for treatment.*

Here, the conjunction *and* is used to join the two extracted sentences, sentences (19) and (20), and the resulting sentence is shown in (21). However, the resulting sentence is not a well written sentence. The two clauses share the same polarity but due to the topic dissimilarity and compounded sentence length, the above aggregation certainly does not improve the summary readability. In the following example:

Example 2:

(22) Extracted Sentence: *Wikipedia is a great starting point and often points to primary sources.*

(23) Extracted Sentence: *Wikipedia isn't even a very reputable source!*

(24) Aggregated sentence: *Wikipedia is a great starting point and often points to primary sources* ***and*** *Wikipedia isn't even a very reputable source!*

The conjunction *and* is used to aggregate the above two sentences, sentences (22) and (23), which results in sentence (24). The two sentences have different polarity of opinions and should not have been aggregated or should have been aggregated with a conjunction that marks a contrast, such as *but*. As shown in Figure 1, overall, this type of error accounts for 9% of the errors.

Insertion of Cue Phrases

As described in Section 2.2, insertion of cue phrases is one of the post-schema heuristics used in BlogSum system after organizing sentences and applying discourse schemata. These sentence aggregations were judged based on its effectiveness for the summary. The following is an example summary where the cue phrase *"Moreover"* was added by the system.

Aggregated Sentences:

Picasa is amazing and is Picasa 2 better than iPhoto? ***Moreover****, in Picasa it works exceptionally. 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. Indeed, Gav said to me, "If you like Picasa, youll love the Mac".*

Here, the system has added the cue phrase *"moreover"* two times based on the post-schema heuristics. The cue phrase *"moreover"* appears within two consecutive sentences and it shows that applying cue phrases just after a sentence aggregation could lead to repetition of these phrases inappropriately and it does not contribute much to a fluent summary.

2.3.4 Summary of the Manual Evaluation Results

Tables 5 and 6 show the results of the manual evaluation. From these results we can see that the content selection mechanism in BlogSum still can be improved (51.2% of errors are of this type). Other than that, BlogSum has performed well in terms of topic relevancy (only 5% of errors) and has a small percentage of other content selection errors which occur in the sentence selection process. Considering the errors in sentence organization shown in Table 6, the numbers seem to indicate that there are two particular areas where BlogSum can be improved. Those are sentence aggregation (which account for 9.0% of errors) and cue phrase insertion (36% of errors). It must be noted that these types of errors cannot be quantified by existing automatic evaluation metrics and were evaluated purely based on human judgment. So from these results, we can conclude that there is much room to improve the readability of an automatically generated summary by extending the capabilities of BlogSum in sentence organization.

Content Selection Errors	**Error/Total Sentences%**	**Error/Total Errors%**
Topic Irrelevancy	5.2%	5.0%
Query Irrelevancy	37.0%	35.1%
Fragmented Sentences	7.0%	6.7%
Interrogative Sentences	4.5%	4.3%
Total	**53.7%**	**51.2%**

Table 5: Manual Evaluation of Content Selection Errors on the TAC 2008 Dataset.

2.3.5 Readability Measures of BlogSum Summaries

From the previous evaluations we have concluded that the performance of BlogSum and especially its sentence organization can be much improved. Our results lead us to conclude that BlogSum could be improved at the level of its sentence aggregation techniques and cue phrase insertion. One of the challenges we see in sentence aggregation is grouping or clustering relevant sentences based on their content and stated opinion. However, one

Sentence Organization Errors	**Error/total sentences%**	**Error/total Errors%**
Dangling Anaphora	6.1%	5.8%
Sentence Aggregation Mismatches	9.3%	8.9%
Cue Phrase Insertion	35.8%	34.1%
Total	**51.2%**	**48.9%**

Table 6: Manual Evaluation of Sentence Organization Errors on the TAC 2008 Dataset.

important parameter that BlogSum has not handled is the complexity of the sentences. The complexity of a sentence can be defined as,

> *"A complex sentence is a sentence with one independent clause, which represents the main content of the sentence and at least one or more dependent clauses which present the subordinate information."* [Baskervill and Sewell, 1986]

Because it contains subordinate clauses, a complex sentence can become long and difficult to comprehend. So aggregating sentences without taking the complexity of sentences into account would decrease the readability of the resulting sentences and would not contribute to an effective sentence organization. So with this idea in mind, we have computed some standard readability measures [Štajner et al., 2012] for the summaries generated by BlogSum with and without the sentence organization post-schema heuristics. Tables 7 and 8 show the results. The readability measures that we computed are standard measurements that indicate the comprehension level of a given text. The Flesch-Kincaid Reading Ease test [Štajner et al., 2012] gives a value between 0 to 100 and indicates the easiness in comprehension; the higher the score is, the lower the comprehension level required by an audience. The other tests (Flesch Kincaid Grade Level, Gunning Fog Score, Coleman Liau Index, SMOG Index and Automated Readability Index [Štajner et al., 2012]) give an approximation of the grade level that is required to understand a particular text. So in these

	BlogSum: Complete	**BlogSum: No Aggregation**
Flesch-Kincaid Reading Ease	55.1	60.4
Flesch Kincaid Grade Level	10.6	8.9
Gunning Fog Score	12.4	10.5
Coleman Liau Index	11.7	11.6
SMOG Index	9.6	8.4
Automated Readability Index	10.7	8.6

Table 7: Readability Measures of BlogSum Summaries on the TAC 2008 Dataset.

	BlogSum: Complete	**BlogSum: No Aggregation**
Flesch-Kincaid Reading Ease	45.8	50.2
Flesch Kincaid Grade Level	12.0	10.2
Gunning Fog Score	13.3	11.3
Coleman Liau Index	13.9	14.0
SMOG Index	10.7	9.4
Automated Readability Index	12.7	10.4

Table 8: Readability Measures of BlogSum Summaries on the DUC 2007 Dataset.

measurements, the lower the value, the easier is the text to be understood by an audience. The results of Table 7 and 8 show that BlogSum sentence organization techniques have, in general, caused the summaries to decrease in comprehension level. We believe that the reason for these results is that the complexity of the individual sentences are not taken into account when they are aggregated and the resulting sentences tend to become longer and more complex. These results show that in order to apply sentence organization, there needs to be a mechanism to simplify sentences first.

2.4 Chapter Summary

In this chapter, we have presented our evaluation of the BlogSum system [Mithun, 2010]. We have performed a content evaluation of BlogSum against the MEAD summarizer using different configurations. The results of the automatic evaluation using ROUGE showed that BlogSum can perform better in opinionated text summarization, compared to MEAD and also perform just as well as MEAD in generalized summarization. Then we performed a manual evaluation of BlogSum summaries which lead us to identify BlogSum errors, mainly content selection errors and sentence organization type errors. After estimating the frequency of these, we also evaluated the complexity of the summaries before and after applying sentence aggregation techniques using automatic readability measures. These experiments have shown that sentence simplification is needed in order to improve BlogSum sentence organization techniques. The next chapter will therefore describe the previous work done in sentence compression in details.

Chapter 3

Literature Review

In our previous chapter, we gave an introduction to text summarization and presented the evaluations we performed on the automatic summarizer BlogSum. Our evaluations have lead us to conclude that BlogSum can be improved in its sentence organization methods. Also we showed that in order to improve sentence organization, there is a necessity to simplify the sentences before performing the sentence aggregation task. So this became our motivation to apply sentence compression as an improvement to automatic summarization.

3.1 Previous Work

Early work on automatic text summarization was mainly focused on content selection methods to generate summaries. The summaries generated by selecting important content out of documents are known as extractive summaries and much research has been done to improve extractive methods. When extractive methods have been improved to a certain level, other concepts were slowly introduced to the area of automatic summary generation. One of them is text compression as a method to improve the quality of the automatically generated summaries. Sentence compression can improve a summary in three different ways. First, text compression, or specifically sentence pruning, can lead to simplification of the content of a summary, which is a requirement in summarization. Second, sentence pruning can help to reduce redundant and irrelevant information in summaries. And lastly, text compression produces more space to include useful information for length-limit summaries. In the following sections, we will review the main approaches used in sentence compression for automatic summarization.

Previous work on sentence compression can be categorized into three main classes: machine learning and classifier based approaches (e.g. [Knight and Marcu, 2002]), keyword and phrase structure based sentence trimming (e.g. [Conroy et al., 2006, Pingali et al., 2007]) and syntax based sentence pruning (e.g. [Jing, 2000, Gagnon and Da Sylva, 2006]). Machine learning and classifier based techniques rely on an annotated corpus and almost all the evaluations done using a set of sentences paired with human annotations and evaluations.

3.2 Machine Learning and Classifier Based Techniques

3.2.1 Cut and Paste Text Summarization

[Jing and McKeown, 2000] have presented one of the early approaches on sentence compression using machine learning and classifier based techniques. This research work was focused on removing inessential phrases in extractive summaries based on an analysis of human written abstracts. For this experimental work, the authors have used human-written abstracts, which were collected from the free daily news service and communications related headlines, provided by the Benton Foundation[1]. This text corpus consisted of news reports on telecommunication related issues and other topics. In their work, the authors have used several techniques to improve phrase removal thus resulting in simplified texts. The authors have used a syntactic parser to generate sentence graphs and mark important words in order to preserve the grammaticality of sentences. As the next step, they used this model and generated different graphs for sentences extracted from a text corpus and the corresponding human written sentences. With that, they identified different types of phrases which are present in the original text but not in human written simplified sentences. These phrases are tagged and used as a training set to develop a statistical

[1] http://www.benton.org

sentence decomposition module based on a Naive Bayes Classifier to decide how likely a phrase can be removed from a sentence. This classifier was tested against a test set of the corpus plus human written summaries. For evaluation, the authors have defined a parameter called success rate, the ratio between the number of occurrences where the module and the human written summaries showed the same decision in removing a sentence phrase and the number of occurrences where both made decisions to remove a phrase structure. In [Jing and McKeown, 2000], the authors reported a 71.8% success rate in decision making by the module but have noted a low success rate in removing adjectives, adverbs or verb phrases.

A sample output of their module is shown below:

> (25) Original sentence: *When it arrives sometime next year in new TV sets, the V-chip will give parents a new and potentially revolutionary device to block out programs they don't want their children to see.*
>
> (26) Automatic Compression: *The V-chip will give parents a new and potentially revolutionary device to block out programs they don't want their children to see.*
>
> (27) Human Compression: *The V-chip will give parents a device to block out programs they don't want their children to see.*

3.2.2 Summarization Beyond Sentence Extraction

A probabilistic approach to sentence compression was proposed by [Knight and Marcu, 2002]. The authors used a noisy channel model as a probabilistic model to effectively compress texts in summarization. In this work, the text compression model was implemented based on the hypothesis that there exists a shorter original sentence and the existing longer sentence was formed by adding optional phrases. They have defined text compression as a task of identifying this original shorter sentence using the noisy channel probabilistic model. Here, the original hypothetical shorter string s is assigned a probability of $P(s)$, which indicates the probability of generating this hypothetical string. If the sentence s is ungrammatical, then $P(s)$ will be very low. They have called this probability the source model. Then the authors have defined the noisy channel model as: given the long string t and every pair of (t, s), a probability $P(t \mid s)$ evaluates the likelihood of arriving at the long string t, when s is expanded. Finally, they have defined the decoder model as: when string t is observed, finding string s such that it maximizes $P(s).P(t \mid s)$. Their model is designed considering two key features: preserving grammaticality and preserving useful information. In order to calculate these probability scores, they used context free grammar parses of the sentences or as they called them "the strings" and the bi-gram scores of the words in these strings. The parse trees were generated using the Collins Parser [Collins, 2003] and these parse trees were used to calculate the probabilistic scores. Then these probability scores are combined with the standard bi-gram scores calculated using the words in the string. Finally these scores are compared to find the most likely short string s for a given long string t. The authors have evaluated their system using the Ziff-Davis corpus[2], a collection of newspaper articles announcing computer products. They have generated compressed strings using their noisy channel model, a baseline algorithm, which compresses strings based on highest bi-gram scores and they also used a corpus of human-written compressed strings. All the compressed sentences were then presented to four human judges who scored them with a value of 1 to 5, considering their grammaticality and importance. The results have showed that the noisy channel model could score similar compression rates (number of words removed) compared to human written compressed texts but the importance and grammaticality scores were slightly lower than in human-written compressed texts.

3.2.3 Decision-Based Model for Sentence Compression

Another statistical approach was described in [Knight and Marcu, 2002] where the authors have used a decision tree based classification to achieve text simplification. In this work, the authors have defined sentence compression as a rewriting task between the sentence t and a shorter sentence s. To achieve this, the authors decomposed the rewriting operations into a sequence of shift-reduce-drop actions that deal with a stack and a sequence of words, from the original long sentence t to the shorter sentence s. The word sequences are

[2] http://www.ziffdavis.com/

labeled with syntactic constituents and the actions named $SHIFT$, $REDUCE$, $DROP$ and $ASSIGNTYPE$ represent operations performed when these word sequences are processed. The authors have used a stack with these operations and for an example, the $SHIFT$ actions transfer a particular word from the input word sequence to the stack, the $REDUCE$ operation pops the top element in the stack and combine it with the next word in the processing sentence and finally, the $DROP$ operation eliminate a certain word from the sequence by removing it and not combining it with the syntactic structure on the top of the stack. The final result of these action sequences and the main sentence word sequence is a reconstructed shorter sentence. In order to build the decision rules, they have used the Ziff-Davis[3] corpus and human simplified texts. Using this corpus, the authors have generated the learning cases by mapping the long sentence to the shorter sentences and the subsequent action sequences. Then the authors have defined operational and syntactic tree specific features which would map learned action sequences. For the evaluation, they used the same criteria as [Knight and Marcu, 2002] and the results showed an improvement in compression rate but performed slightly lower than the human written text simplifications. As an effort to improve the methodology in [Knight and Marcu, 2002], the research work in [Nguyen et al., 2003] was focused on introducing semantic features to improve decision tree based classification. In this research work, the authors used the same concepts as the word sequences and the action operators in [Knight and Marcu, 2002], but in addition, they used Charniak's Parser [McClosky et al., 2006] to generate the syntactic trees for the original sentences and have incorporated semantic information using the WordNet [Miller, 1995] database. With this new information, they have defined extra semantic features which contribute to the decision tree based classification. The semantic information used is general semantic types, such as: $HUMAN$, $THINGS$, $ANIMAL$, $CONCEPT$, $INSTRUCTOR$, $COMPUTER$ etc. For the evaluation, they used the original algorithm in [Knight and Marcu, 2002] as the baseline and used the same corpus and the criteria used in [Knight and Marcu, 2002]. The compression rate they achieved was

[3] http://www.ziffdavis.com/

similar to the work of [Knight and Marcu, 2002] (57.1%) but showed a marginal improvement in Importance score, compared to [Knight and Marcu, 2002]. But still these results were lower than human written gold standard compressions in terms of grammaticality and importance scores.

3.2.4 Probabilistic Sentence Reduction

[Le Nguyen et al., 2004] described another statistical approach used in text compression. Here, the authors argue that most of the earlier text compression methods were focused on finding a local optimum as each sentence or string is compressed independently. So they point out that text compression could be seen as a domain problem of finding a global optimum by considering the compression of the whole text/document. In their approach they applied an SVM (Support Vector Machine) model and reduce the text compression problem to a classification problem. As the first step, they have used an annotated corpus where the original sentences and their corresponding shortened sentences were provided to define a set of features named as operation features. In this case, they looked at syntactic trees of each pairs and defined operational rules to deduce shorted syntactic tree out of original syntactic tree. With that, they identify operational features to use in SVM classification. In addition to the syntactic features, they have also defined a set of features based on a semantic analysis in order to preserve the main content of the given original sentences. Since SVM is a binary classification, they have used a pair wise strategy to evaluate the best text compression. Similarly to other approaches, the authors have used the Ziff-Davis[4] corpus for their evaluation and human judgment. For the comparison of results, they have used the algorithms presented in [Knight and Marcu, 2002] and [Nguyen et al., 2003] as baseline algorithms. The evaluation schema included determining the compression rate, the importance of the data and the grammaticality of the compressed text. Their results have shown a slight improvement in grammaticality and importance scores compared to the baseline algorithms they used but they were lower compared to the scores of the human written abstractions.

[4] http://www.ziffdavis.com/

3.2.5 Integer Linear Programming Approach

A recent effort in sentence compression which highlighted a different approach was presented in [Clarke and Lapata, 2008]. In this paper, the authors have described the use of an integer linear programming (ILP) model to infer globally optimal compressions while adhering to linguistically motivated constraints. In their work, they decomposed the sentence compression task into an ILP problem by defining a language model based on an ILP model. Here, each word introduces a decision variable with a value of 0 or 1, representing its presence in the compressed text. But since a unigram model can contribute to ungrammatical sentences, their ILP model consists mainly on word trigrams and each one is represented by a decision variable. For this ILP model, the authors have introduced several constraints: syntactical word ordering features, significance of content which is measured by cosine similarity, discourse structure and grammaticality based constraints etc. With this model, the objective was to find the global optimal compression for a given text. For training and evaluation, they have gathered and annotated their own corpus using various corpora such as the British National Corpus (BNC)[5], the American News Text corpus [Graff, 1995] and English Broadcast News corpus [Graff et al., 1996] etc. The first part of their evaluation was focused on estimating various parameters for the ILP model using the training set and then the authors have used the rest of the corpus for evaluating of their model. For the evaluation they have used the F-score based automatic evaluation measure, compared n-gram models present in human annotations and have also performed a manual evaluation using human judgment. The authors compared their model with and without the constraint based models and their constraint based model showed better results in automatic and human evaluations.

3.2.6 Maximum Entropy Based Approach

[Galanis and Androutsopoulos, 2010] described another machine learning approach to sentence compression and the authors used two stages of learning and classifying processes

[5] http://www.natcorp.ox.ac.uk/

to find an optimal sentence compression. In their work, they first trained a probabilistic model based on Maximum Entropy (ME) and the goal was to evaluate how likely an edge of a syntactic tree can be removed based on a set of features. The features were defined considering the Part of Speech (POS) tags of the surrounding words of the edge, the head of the edge and the modifier of the edge. The system was trained using a corpus containing original sentences and their human annotated shorter sentences. After training the system, it was used to generate probable candidates for a given original sentence and these candidates were ranked based on the grammaticality (calculated using a language model) and the importance of content (calculated using term frequencies and inverse document frequencies of a corpus). Using these scores and other additional features, they have trained a Support Vector Machines Regression (SVR) model to select the best candidate compression. For the evaluation, the authors have used compressed summaries using a simple algorithm based on Maximum Entropy (ME) classifier and human compressed summaries as baseline summaries. These summaries were judged by human annotators and their results showed that their compression techniques outperformed the baseline algorithm yet underperformed compared to the human annotated compressions.

3.3 Keyword Based Techniques

In keyword based approaches, an effort to identify basic words, phrases or sentence patterns is made. Keyword based approaches do not go into deep language processing techniques but simplify sentences based on identifying common words or phrase patterns and removing them.

A different approach to sentence compression for automatic summarization was described in [Conroy et al., 2006]. This work has taken a more conservative approach to sentence pruning. In their approach, they have used a list of key words or phrases to identify less significant parts of a sentence and removed them to produce a shorter sentence. The keyword list was compiled in an ad hoc fashion and was used in a flexible way to omit some of the terms when needed. Also it was maintained as an expanding list, having the ability to discover more words or phrase patterns and add them to the list. The phrases and words included in this list were mostly adverbs, conjunctions, idioms and phrases like *"As a matter of fact"* and *"At this point"*, that typically appear at the beginning of a sentence. This work has highlighted the use of shallow syntactic parsing of text and direct usage of a list of keywords/phrases for sentence pruning. They have evaluated their overall summarization system CLASSY [Conroy et al., 2005] with DUC 2005 [Dang, 2005], where it showed an improvement in ROUGE scores when using the keyword list. It must be noted that CLASSY participated in several summarization shared-tasks, including DUC 2006 where it scored among the top three based on ROUGE scores.

A similar approach to sentence compression was also described in [Pingali et al., 2007] where the authors have used sentence compression as a part of their text summarization system. They have manually identified word/phrase patterns and used a handcrafted [Pingali et al., 2007] list of these patterns to perform sentence reduction. In their paper, the authors also mentioned that these patterns were selected based on fact that they do not carry useful information and are easy to be removed without losing relevant content. They have also stated that they did not focus much on losing the grammaticality but only in keeping important content while compressing the sentences. Their overall summarization system was the best scoring system based on ROUGE evaluations at the DUC 2007 summarization task but they have not stated whether they evaluated their sentence compression extrinsically.

3.4 Syntax Based Techniques

Finally, the syntax based sentence pruning approaches focus on simplifying sentences based on their syntactic trees. Here, the main idea is to remove sub-syntactic structures to achieve sentence compression.

In contrast to the statistical model approaches we have described above, a different approach on sentence compression was presented in [Gagnon and Da Sylva, 2006] that uses syntactic level sentence pruning based on grammar rules. The grammar rules were identified after analyzing sentences and phrase structure trees. These rules were used to map suitable sub-trees to be pruned while traversing through a complete parsing of a sentence. Here, the authors have mainly focused on simplifying French written texts. For this research work, they have used complete parsed dependency grammar structures and several defined grammar rules to remove syntactic sub-structures. First they compressed the text by only keeping obligatory complements and phrasal specifiers such as determiners. This has resulted a large compression of texts but the compressed texts tended to be ungrammatical. They then used more specific grammar structures to be pruned, including prepositional complements of the verbs, subordinate clauses, noun appositions and interpolated clauses. This lead them to achieve a lower compression compared to the first method yet created more grammatically coherent compressed text. For the testing and evaluation, the authors used text corpora created from 10 different genres. They have evaluated the text compression according to the compression rate, comparing the sentence compression achieved from the grammar rules and human written compressed text. They have demonstrated a compression rate of 74% while retaining grammaticality or readability of text by more than 64%.

A more recent effort based on syntactic pruning was presented in [Zajic et al., 2007]. In this research work, the authors also used the syntactic structures of sentences and applied linguistically motivated filtering to generate compressed sentences. They had used this technique before to generate headlines for single documents and they were motivated to extend the same procedure and rules for multi-document summarization. As the first step of formulating their grammar rules, they have performed a corpus analysis using subset of documents from the TIPSTER [Harman and Liberman, 1993] corpus and their human written summaries. Using these summaries, they have identified and counted syntactic patterns which were absent from human-written summaries compared to original documents. Some of these grammatical structures were mentioned as preposed adjuncts, conjoined clauses, conjoined verb phrases and relative clauses etc. With the statistical data they gathered, they defined a trimming algorithm consists of operations to remove sub-trees from the syntactic structure of a sentence while traversing through a complete parsed tree. They have evaluated their sentence compression technique with the DUC 2003 summarization task and it has shown an improvement in ROUGE scores compared to the uncompressed length-limit summaries. A more recent research work published in [Jaoua et al., 2012] discusses a syntax level pruning based on grammar phrase structures as a part of their automatic summarization system. They have implemented a compression module which selects adverbial modifiers and relative clauses as suitable candidate for removing to achieve sentence compression. Their evaluations were performed using the DUC 2007 summarization track and they have reported an improvement in ROUGE scores after adding the compression module to their summarization system.

3.5 Chapter Summary

In this chapter we have presented an overview of previous work on sentence compression. The machine learning and keyword based techniques described in Sections 3.2 and 3.3 focus on simplifying sentences based on general features. However, our objective is applying sentence compression for automatic summarization where the relevancy of the content depends on the given topic and the query. In addition, machine learning and classifier techniques rely on a training corpus that is not available for our specific task. Therefore in order to implement sentence compression as a part of BlogSum summarizer, we decided to explore syntactic level based approaches. Our goal was to implement a syntactic based sentence compression technique (see Section 3.4) which focused on preserving grammaticality and relevant content. The next chapter will describe our approach in detail.

Chapter 4

Syntactic Sentence Compression

In the previous chapter, we have elaborated on the previous work that has been done on sentence compression. We have categorized the various approaches as machine learning and classifier based techniques, keyword based techniques and syntax based sentence pruning techniques. For our goal of sentence compression, we decided to explore syntax-based methods and perform an evaluation of our techniques. As described in Section 3.4, previous work on syntax based sentence pruning has focused on identifying specific types of phrase structures and removing them to simplify sentences. The reason for predefining which syntactic structures to be removed is to ensure grammaticality of the sentences. We were influenced by this technique but our goal was not only to preserve grammaticality, but also to preserve relevant content as well while pruning syntactic structures. In extractive summarization, for a certain summary, the given topic and the query determine the relevant content. We decided to take this information into account when pruning syntactic structures in our techniques. Overall, the approaches we have developed are based on the use of:

1. Complete syntactically parsed sentence tree structures.

2. Predefined sentence pruning heuristics.

3. A relevancy score based filtering.

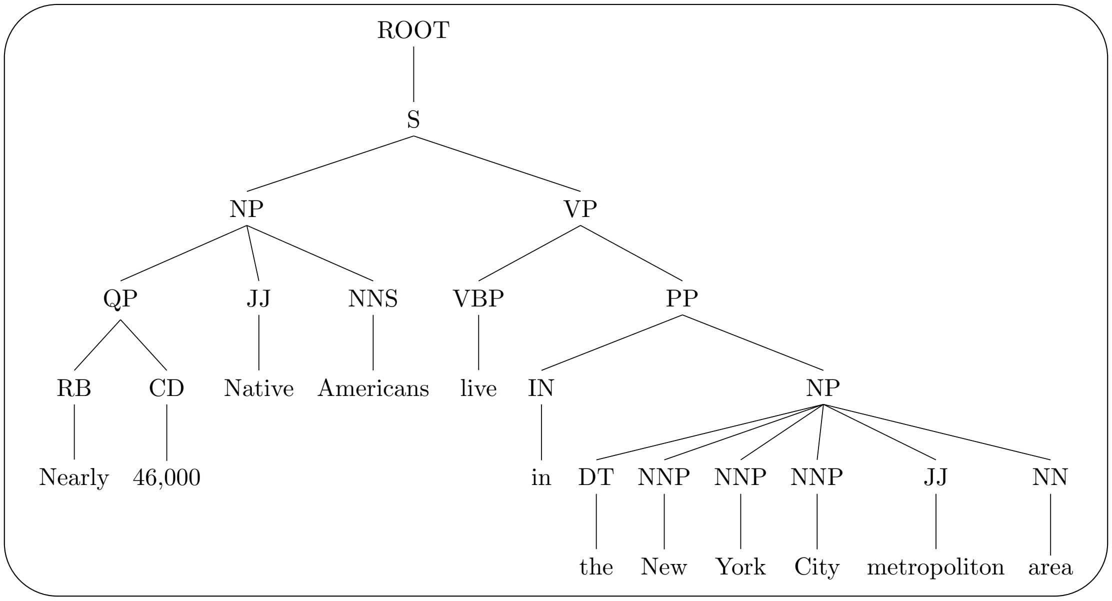

Figure 2: Phrase Structure for Sentence (28).

The following will describe these in more detail.

Our method is based on a complete syntactic parse of the sentences. In order to generate syntactic trees, we used the Stanford Parser [Marneffe and Manning, 2008], a statistical natural language parser that generates grammatical structures of sentences. The Stanford Parser is as a probabilistic parser that uses the knowledge of the language learned from already hand-parsed sentences to produce the most likely parse tree for a given new sentence. We used the Stanford Parser to generate complete parse of the sentences in a summary and used these syntactic structures to identify specific sub structures to be pruned. A sample syntactic tree generated for sentence (28) is shown in Figure 2.

(28) *Nearly 46,000 Native Americans live in the New York City metropolitan area.*

As shown in the Figure 2, the sentence is mainly constructed by a Noun Phrase (NP) followed by a Verb Phrase (VP). The NP is itself made of a Quantifier Phrase (QP) *"Nearly 46,000"*, an Adjective (JJ) *"Native"* and the head Plural Noun (NNS) *"Americans"*. The verb phrase consists of the present verb (VBP) *"live"* and the Prepositional Phrase (PP) *"in the New York City metropolitan area"*. The Stanford Parser uses the Penn Treebank [Marcus et al., 1993] constituents and 109 tags which are inventoried in Appendix A.

In order to play with the influence of content versus syntax only, we have implemented three different sentence compression techniques. They are,

1. Syntax-Driven Sentence Pruning
   Removal of syntactic sub-structures based on syntax-driven heuristics (described in Section 4.1). This approach prunes sentences based only on syntactic features.

2. Syntax and Relevancy Based Sentence Pruning
   This is a toned-down approach of syntax-driven pruning using a relevancy filtering threshold (described in Section 4.2).

3. Relevancy-Driven Sentence Pruning
   This approach removes embedding syntactic sub-structures based on a relevancy threshold and content filtering (described in Section 4.3).

The following sections describe these approaches in details.

4.1 Syntax-Driven Sentence Pruning

In our first technique, our goal was to focus more on preserving sentence grammaticality by removing specific syntactic sub-structures. Also as we described in the previous section, we assume that most of our selected syntactic structures carry secondary information and hence we expected that using these grammatical structures would not remove relevant content significantly. In order to prune syntactic sub-structures, we have defined syntax-based pruning heuristics based on grammar rules. The analysis we did on summaries generated using BlogSum (see Section 1.3) and also previous work on syntax based sentence compression such as [Gagnon and Da Sylva, 2006, Zajic et al., 2007] helped us to identify the syntactic structures to be removed. The heuristics we applied were:

1. Pruning Relative Clauses
2. Pruning Adjective Phrases
3. Pruning Adverbial Phrases
4. Pruning Conjoined Verb Phrases
5. Pruning Appositive Phrases
6. Pruning Prepositional Phrases

Let us describe these grammar based syntactic pruning heuristics now.

4.1.1 Pruning Relative Clauses

A relative clause is a post modifier clause that modifies a noun or a noun phrase and is connected to the noun by a relative pronoun (*which*, *that*, *who*, *whom*, *whose* etc.), a relative adverb (*where*, *when*, *why* etc.), or as a zero relative clause [Dietrich, 2007]. Relative clauses are often used in English to provide complementary information. They are considered as subordinate clauses in general and contribute to the complexity of a sentence. As an example, consider the following sentence that consists of a relative clause attached to its head noun *“Animal”*.

> (29) *Animals, whom we have made our slaves, are not considered our equal.*

An important characteristic of relative clauses is that they can easily be removed from a sentence without losing the grammaticality of the sentence. Considering this fact and their usage in providing additional information, we decided to implement a syntactic heuristic that prunes sentences by removing relative clauses that are connected to noun phrases. As an example, consider the following sentence (30) and its parse tree shown in Figure 3.

> (30) *“It’s over”, said Tom Browning, an attorney for Newt Gingrich, who was not present at Thursday’s hearing.*

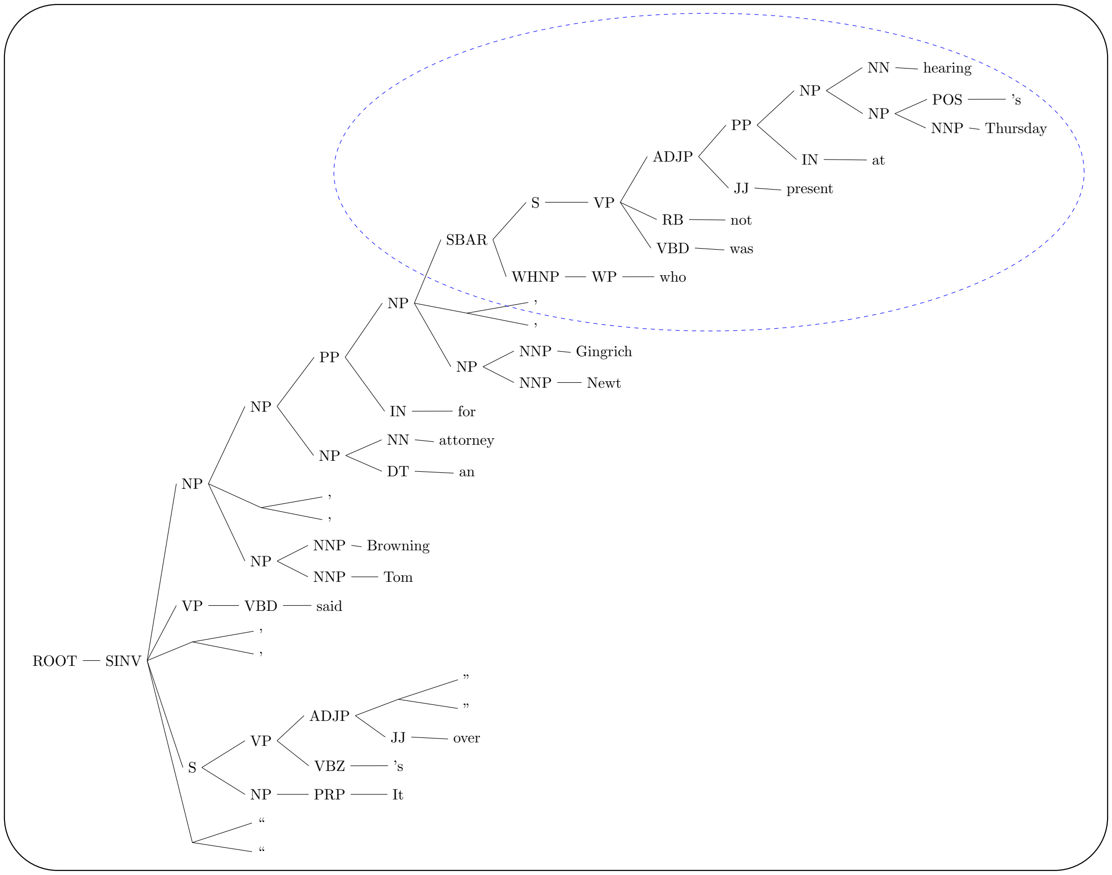

Figure 3: Phrase Structure for Sentence (30).

So according to the heuristic, in Figure 3: the sub tree structure SBAR[1], which represents a relative clause is taken as a candidate sub-tree to be pruned. The resulting shortened sentence is shown below.

(30c) *"It's over", said Tom Browning, an attorney for Newt Gingrich* ~~*, who was not present at Thursday's hearing.*~~

4.1.2 Pruning Adjective Phrases

Another type of phrase structure that we identified as a suitable candidate to be pruned are adjective phrases. An adjective phrase is word, phrase, or a sentence element that enhances limits or qualifies the meaning of a noun phrase. The following sentences are examples that contain adjective phrases.

(31) *The little bird flew gracefully.*

(32) *The SPLC represented the predominantly black Macedonia Baptist Church in Clarendon.*

Adjective phrases contribute to longer sentences. When considered as complementary phrases, they can be removed from a sentence without hurting the content or the grammaticality of the sentence. We took this property of adjective phrases into account and defined another heuristic to remove adjective phrases for your sentence compression schema. As an example, consider the following sentence:

(33) *An editorial accompanying the obesity issue of JAMA calls for developing a comprehensive national strategy to prevent obesity.*

The grammatical phrase structure for the above sentence (33) is shown in Figure 4. According to the pruning adjective phrase heuristic we have defined, the phrases *"obesity"* and *"comprehensive"* are identified as suitable candidates to be pruned from the original sentence. The resulting sentence would be as below:

[1] According to the Penn Treebank labels, $SBAR$ labels clauses introduced by a subordinated conjunction.

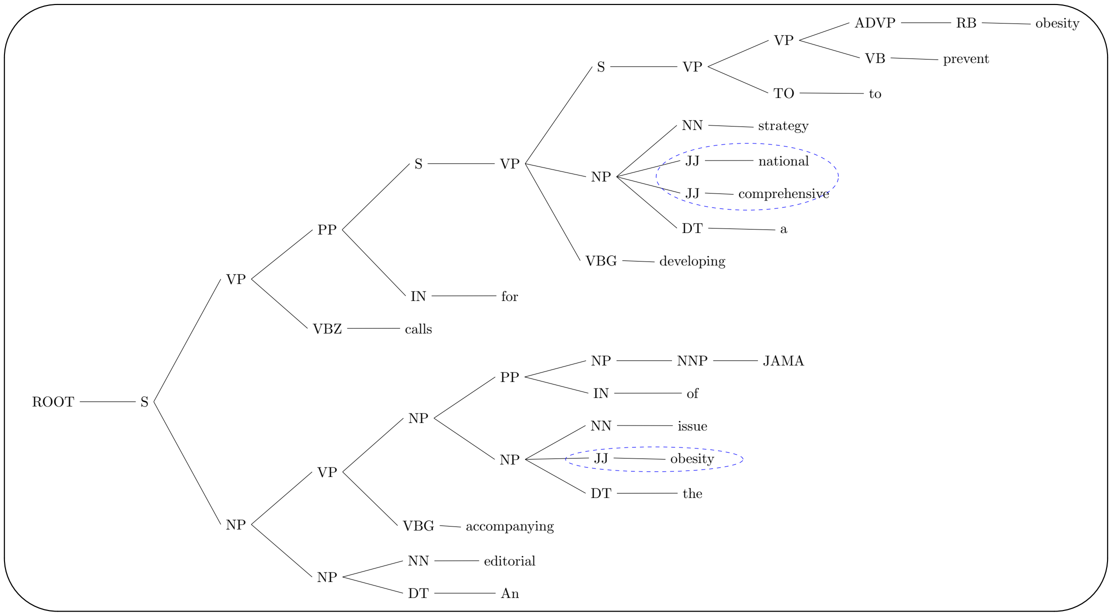

Figure 4: Phrase Structure for Sentence (33).

(33c) *An editorial accompanying the ~~obesity~~ issue of JAMA calls for developing a ~~comprehensive national~~ strategy to prevent obesity.*

Pruning all adjective phrases may be too harsh of a heuristic because an adjective phrase may contain information that is necessary to understand a longer context. As an example, consider the following sentence:

(34) *The Southern Poverty Law Center, which was founded in the 1970s to battle bias, won fights against the Ku Klux Klan and white supremacist groups.*

Here, the phrase *"white supremacist groups"* consists of an adjective and a noun phrase. As a whole, it represents a different meaning than the sum of the meanings of individual words. In linguistics, this phenomenon is called non-compositionality. In our example, if we prune the adjective *"white"*, the sentence would lose its significant meaning. In order to avoid this, we have toned down our adjective phrase pruning by filtering out non-compositional phrases. To do so, we have implemented a dictionary approach to identify these phrases. We used WordNet [Miller, 1995], a lexical database for English, widely used in NLP applications. In our pruning mechanism, when we find an adjective phrase to be pruned, we use the WordNet database to identify whether it represents a collocation and use this information to decide whether to prune the adjective phrase or not. For that, we query the WordNet database for the particular phrase. If the database does not contain the phrase, we remove the adjective phrase from the sentence, assuming it is not a strong collocative phrase.

4.1.3 Pruning Adverbial Phrases

Adverbial phrases modify verb phrases, an adjective or another adverb. An adverbial phrase is a word, phrase or a sentence element that enhances limits or qualifies the meaning of the modifying phrase. The following sentence shows an example of the usage of an adverbial phrase.

> (35) *The SPLC previously recorded a 20-percent increase in hate groups from 1996 to 1997.*

Having similar properties as adjective phrases, we decided to define another heuristic on pruning adverbial phrases. As an example, consider the following sentence and its grammatical phrase structure:

> (36) *So surely there will be a large number of people who only know us for Yojimbo.*

The grammatical phrase structure for the above sentence (36) is shown in Figure 5. According to the pruning adverbial phrase heuristic, the phrases *"surely"* and *"only"* will be considered as candidates to be pruned from the original grammatical phrase structure and will result the following sentence.

> (36c) *So ~~surely~~ there will be a large number of people who ~~only~~ know us for Yojimbo.*

4.1.4 Pruning Conjoined Verb Phrases

Conjunctions have several uses in English, one of which is to combine two or more propositions together to form a longer descriptive sentence that presents common or opposing

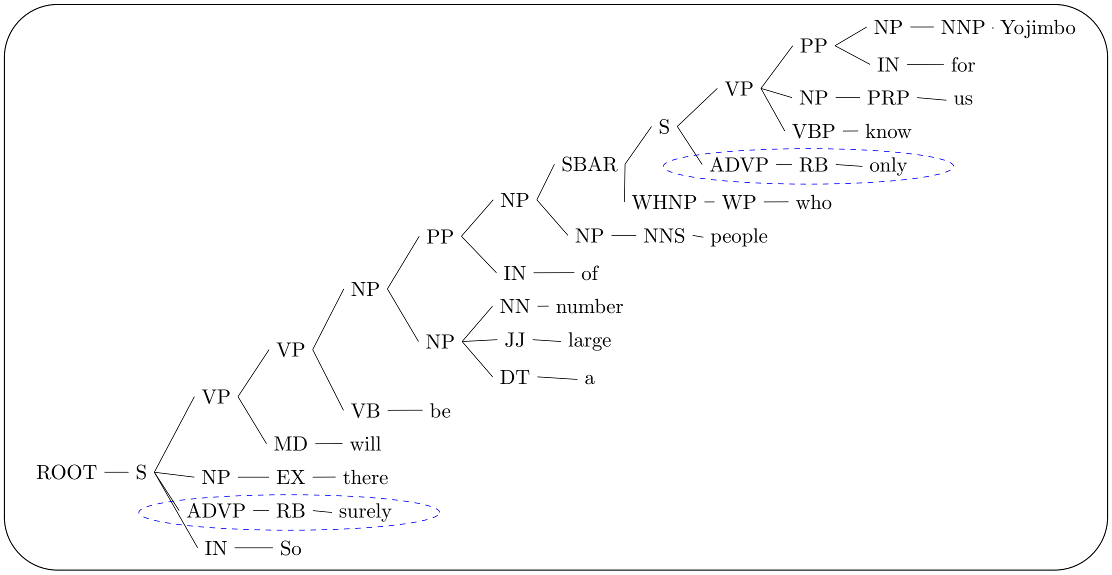

Figure 5: Phrase Structure for Sentence (36).

subject matters. Sometimes, conjunctions are used to attach relative information to the main content of a sentence, especially a relatively shorter phrase at the end of a sentence, connecting them with a conjunction. Given this, we decided to introduce another heuristic to prune additional or relatively less important conjoined verb phrases. For an example, consider the following sentence:

> (37) *The Southern Poverty Law Center has accumulated enough wealth in recent years to embark on a major construction project and to have assets totaling around $100 million.*

The grammatical phrase structure for the above sentence (37) is shown in Figure 6. The conjunction *"and"* connects the verb phrases and the second verb phrase provides additional information to the main subject of the sentence. So the conjunction based pruning heuristic we defined would remove this conjoined verb phrase as follows:

> (37c) *The Southern Poverty Law Center has accumulated enough wealth in recent years to embark on a major construction project ~~and to have assets totaling around~~*

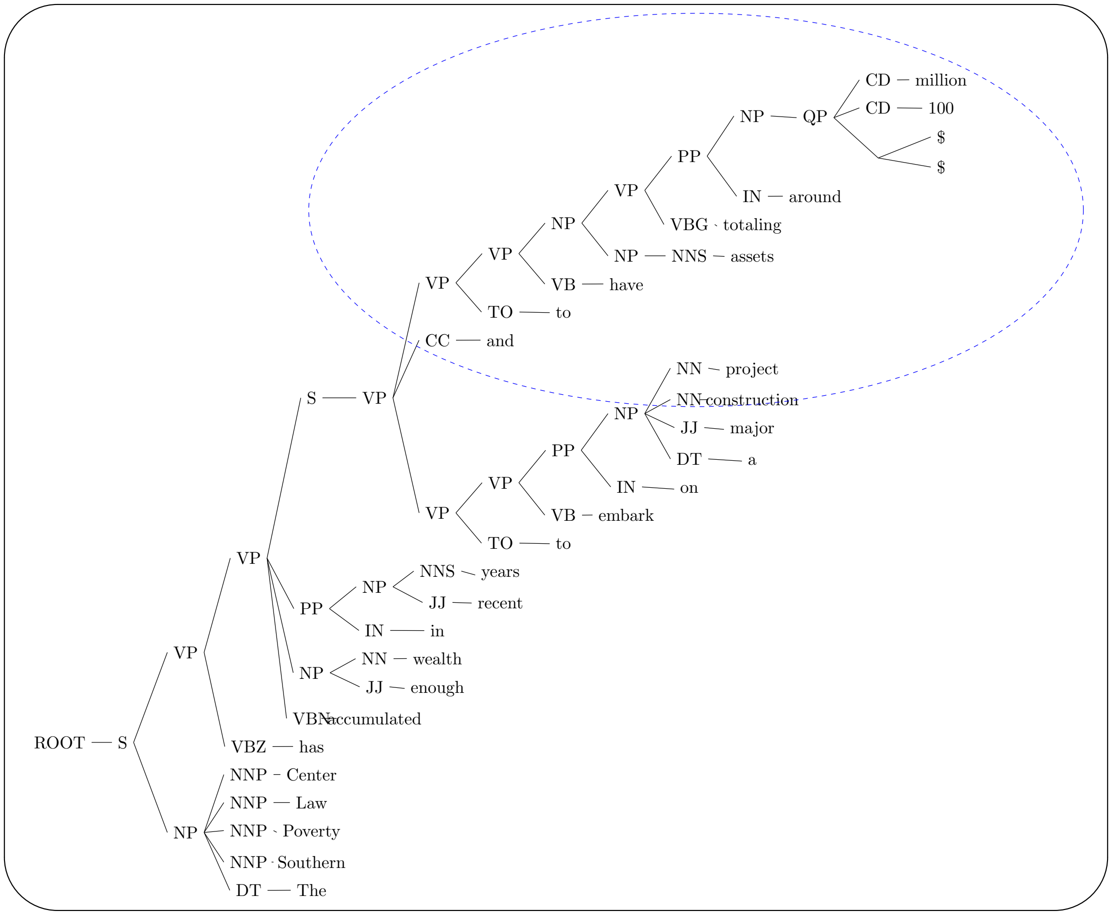

Figure 6: Phrase Structure for Sentence (37).

~~*$100 million*~~.

4.1.5 Pruning Appositive Phrases

Appositive phrases are used in English to provide complementary information about a noun or a pronoun. More formally, an appositive phrase is a word or a phrase that further explains, quantifies or modifies the preceding noun or pronoun. For an example of an appositive phrase, consider the following sentence:

> (38) *Earth, the only planet in our galaxy known to support life, is sometimes called the third rock from the sun.*

Appositive phrases are always in a parenthetical situation (i.e. between commas or parenthesis). Being modifiers to noun phrases, appositive phrases are another type of grammatical structure that can be removed from a sentence without hurting its grammaticality. In light of this, we have defined an appositive phrase based heuristic in our pruning schema. As an example, consider the following sentence:

> (39) *The notice was the first indication that the lawsuit, brought by the Southern Poverty Law Center, may drive the group out of Idaho.*

The grammatical phrase structure for the above sentence (39) is shown in Figure 7. In the given example, the appositive phrase *"brought by the Southern Poverty Law Center"* modifies the noun phrase, *"the lawsuit"* and we can remove this appositive phrase from the sentence without hurting its grammaticality. The resulting sentence would be as follows:

> (39c) *The notice was the first indication that the lawsuit ~~, brought by the Southern Poverty Law Center,~~ may drive the group out of Idaho.*

4.1.6 Pruning Prepositional Phrases

Prepositional Phrases (PP) are one of the most common grammatical structures in English. A prepositional phrase is a phrase that consists of a preposition as the first word and end with a noun, pronoun, gerund or clause and is used as a modifier for nouns, verbs or

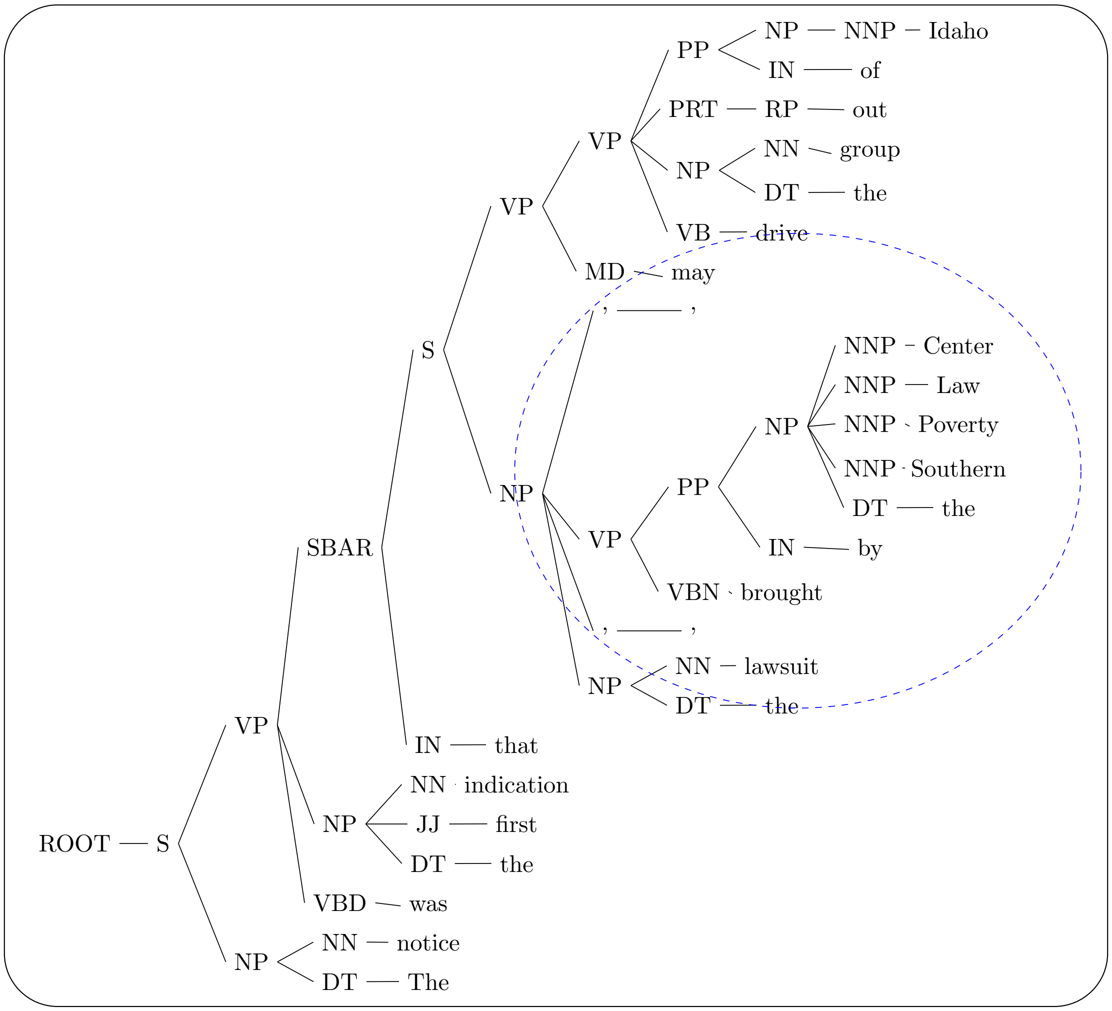

Figure 7: Phrase Structure for Sentence (39).

complete clauses.

The following sentences show examples of these three types of prepositional phrases.

(40) *The hills across the valley of the Ebro were long and white.*

(41) *The salesperson skimmed over the product's real cost.*

(42) *After graduating from City College, Professor Baker's studies were continued at State University*

Sentence (40) contains the prepositional phrase *"across the vally of the Ebro"* that modifies the noun *"hills"*. In sentence (41), the prepositional phrase *"over the product's real cost"* acts as an adverbial phrase, modifying the verb *"skimmed"*. In the last example, sentence (42) contains the prepositional phrase *"After graduating from the City college"* that serves as an introductory modifier for the entire clause.

In sentence pruning, removing prepositional phrases is possible as sometimes they provide secondary information. However, removing all prepositional phrases may hurt the grammaticality or the meaning significantly in some cases. As an example, consider the following sentences:

(43) *The decorator has painted along the trim.*

(44) *The farmer was in the field.*

Here, in Sentences (43) and (44), the prepositional phrases *"along the trim"* and *"in the field"* can be considered as direct objects of the verbs. They provide necessary complements to the verbs, hence removing these prepositional phrases would hurt the main content and the grammaticality of these sentences. In light of this, we decided to remove only specific types of prepositional phrases.

Removing Noun Modifying Prepositional Phrases In most of the cases, when prepositional phrases are used as noun modifiers, they can be pruned without hurting the grammaticality of the sentence. For an example, consider the following sentence:

(45) *The interview was broadcast from a public elementary school in East Harlem.*

The prepositional phrase *"in the East Harlem"* is attached to the noun, *"school"*. In this case, it acts as a noun modifier and can be removed just like we removed adjective phrases (see Section 4.1.2). This results the following sentence:

(45c) *The interview was broadcast from a public elementary school ~~in East Harlem~~.*

Removing Verb Modifying Prepositional Phrases As shown in sentences (43) and (44), removing prepositional phrases attached to verbs can hurt the grammaticality or the meaning of a sentence. Because of this, we decided to remove verb modifying prepositional phrases with caution. In English, a prepositional phrase attached to a verb can act as a verb complement or a verb adjunct [Merlo and Ferrer, 2006]. An adjunct 'modifies' the meaning of its head and is considered optional, while a complement 'completes' the meaning of its head and is considered as obligatory [Dowty, 2000]. A verb complement prepositional phrases are often attached to a transitive verb and they cannot be removed without hurting to the main content of a sentence (see Sentences (43) and (44)). On the other hand, when a prepositional phrase acts as an adjunct, it is often attached to an intransitive verb and it can be removed easily most of the time. As an example, consider the following sentences:

(46) *The San Francisco-based Spinelli Coffee Co. was purchased in July by Tully's Coffee Corp.*

(47) *In July, the San Francisco-based Spinelli Coffee Co. was purchased by Tully's Coffee Corp.*

Sentence (46) and (47) differs only by the location of the prepositional phrase *"in July"* within the sentence. The prepositional phrase *"in July"* modifies the verb *"purchased"* and in English, these verb modifying prepositional phrases are considered as adverbial modifiers. When a prepositional phrase is attached to a verb and acts as an adjunct, it can appear before or after the verb it modifies as in sentences (46) and (47). Taking this into account, we decided to prune prepositional phrases attached to VPs only when they appear before the verbs they modify. So for the sentence (46), the resulting sentence would be:

> (46c) *~~In July,~~ the San Francisco-based Spinelli Coffee Co. was purchased by Tully’s Coffee Corp.*

On the other hand, the prepositional phrase *“by Tully’s Coffee Corp.”* also modifies the verb *“purchased”* but because it is placed after the verb, it may very well be an obligatory complement which should not be removed.

Removing Introductory Clauses When prepositional phrases appear as introductory clauses, they often modify an entire clause and just like relative clauses and appositive clauses, they can be easily removed without hurting the grammaticality and the main content of the sentence. For example, consider the following sentence:

> (48) *In a lawsuit that goes to trial Monday, Attorney Dees of SPLC is representing a mother and son who were attacked by security guards for the white supremacist group.*

The underlined prepositional phrase acts as an introductory clause to the main clause and can easily be removed. The resulting sentence will be as follows:

> (48c) *~~In a lawsuit that goes to trial Monday,~~ Attorney Dees of SPLC is representing a mother and son who were attacked by security guards for the white supremacist group.*

These cases described above allow us to preserve grammaticality while removing secondary information carried by the prepositional phrases. These three heuristics are used in conjunction and can be applied to a single sentence. For an example in Sentence (49):

> (49) *As a result, they possibly wouldn’t be able to do transactions in Euro from the day it takes effect.*

The phrase structure for the above sentence is given in Figure 10. In the above sentence, there are three prepositional phrases: the phrase, *“As a result”* acts as an adverbial modifier

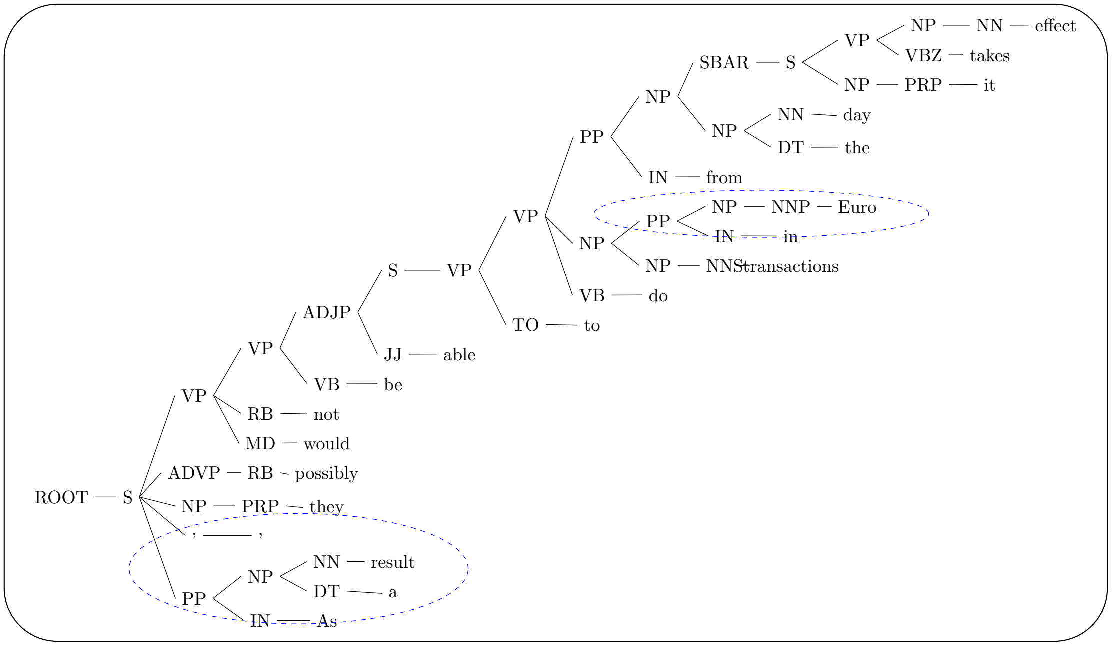

Figure 8: Phrase Structure for Sentence (49).

and appears at the beginning of the sentence, the phrase *"in Euro"* modifies the noun *"transactions"* and the prepositional phrase *"from the day it takes effect"* modifies the verb *"do"*. So according to our heuristics of removing prepositional phrases, the resulting sentence would be as follows:

> (49c) *~~As a result~~, they possibly wouldn't be able to do transactions ~~in Euro~~ from the day it takes effect.*

4.2 Syntax and Relevancy Based Sentence Pruning

In the previous section, we introduced the grammar based heuristics we used for sentence pruning. These heuristics were defined to focus more on preserving grammaticality of the sentence while removing syntactic sub-structures. The purpose of these heuristics is to remove structures that contain secondary information and removing them should not affect the main content of a sentence. But sometimes, these structures may consist of useful information as well. Since our goal is to apply sentence compression effectively on automatic text summarization, a generalized sentence compression approach may not be effective as the summary sentences may relate to the given topic and query. In light of this, we decided to introduce a topic and query based similarity score to tone down our grammar based heuristics, with the objective of preserving topic relevance as well.

In our second technique, our goal was to tone down our first technique, syntax-driven sentence pruning, by associating a relevancy score to the candidate structures that the pruning heuristics would consider removing as a filtering method. Here, our goal is to not only focus on preserving sentence grammaticality but also preserving relevant content as well. The relevancy score will therefore be helpful to avoid removing topic and query related important content.

In order to achieve this, we calculate the $tf.idf$ value for each term in a document cluster. Using this $tf.idf$ value, we calculate a combined $tf.idf$ value for each syntactic candidate structure. This is similar to the regular term $tf.idf$ but is calculated on a per syntactic structure basis [Nguyen and Leveling, 2013]. For each candidate syntactic structure to be pruned, we calculate the cosine similarity of the topic and the query using the precalculated $tf.idf$ values. At each compression, we define a particular threshold value and remove syntactic structure only if the calculated score for the candidate structure is lower than the threshold. So for a particular candidate structure:

$$\begin{aligned} similarity\ score \ =\ & cosine\ similarity(topic(tf.idf), syntactic\ sub\ structure(tf.idf)) \\ & + \\ & cosine\ similarity(query(tf.idf), syntactic\ sub\ structure(tf.idf)) \end{aligned}$$

As an example, consider the following topic, query and sentence:

> Topic: *Israel / Mossad "The Cyprus Affair"*
>
> Query: *Two alleged Israeli Mossad agents were arrested in Cyprus. Determine why they were arrested, who they were, how the situation was resolved and what repercussions there were.*
>
> (50) *Cypriot police officials, who spoke on condition of anonymity, told Haaretz the two men were believed to be spying on behalf of Turkey, a military ally of Israel.*

Figure 9 shows the similarity scores calculated for each of the three syntactic structure that the syntactic pruning heuristics (see Section 4.1) consider removing. Indeed, the pruning heuristics selected the following structures as candidates to be pruned. They are: Adjective (JJ), Relative Clause ($SBAR$) and Prepositional Phrase (PP). The JJ covers the word *"Cypriot"* while the $SBAR$ covers the phrase *"who spoke on condition of anonymity"* that both have a relevancy score of 0.0, calculated against the given topic/query. However, the PP attached to the NP covers the phrase *"of Turkey, a military ally of Israel"* which has a relevancy score of 0.11. If we set the relevancy threshold of $t > 0$, we will remove the sub-structures, JJ and $SBAR$ but not the PP.

4.3 Relevancy-Driven Sentence Pruning

As our last technique, we wanted to focus more on relevant content, identified by our relevancy score and drive the syntactic pruning based on relevancy as opposed to grammatical considerations. Here, our focus was to preserve relevant content hence we have calculated a relevancy score for each syntactic structure and removed it if the relevancy score was lower than a given threshold. In this technique, we decided not to remove any noun or verb phrase structures to simply make sure we do not significantly affect the grammaticality or semantic content of a sentence. To illustrate our technique, consider the following topic, query and the sentence:

> Topic: *Basque separatism*

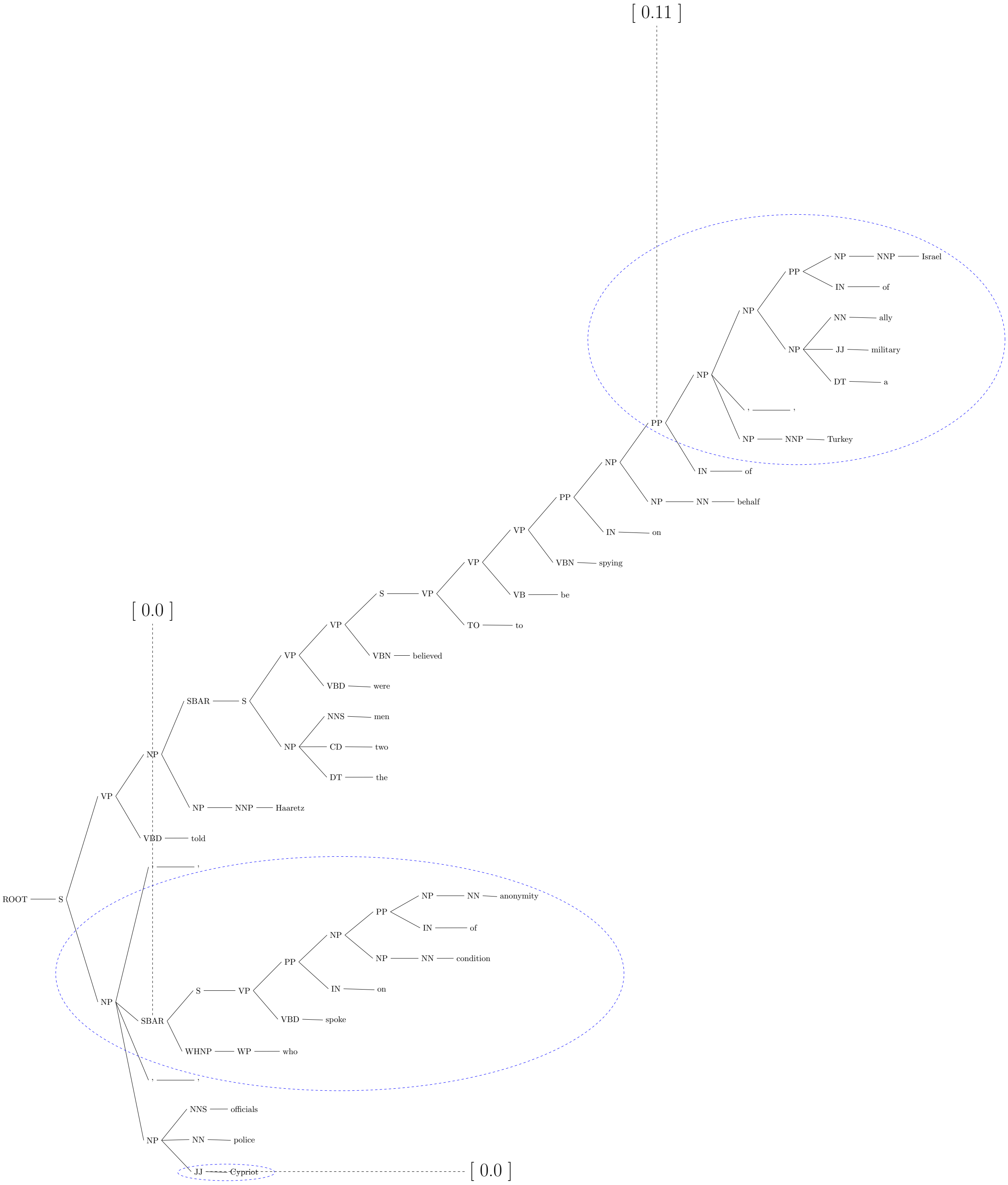

Figure 9: Phrase Structure for Sentence (50).

Query: *Describe developments in the Basque separatist movement 1996-2000*

(51) *After that incident, Herri Batasuna, the political party linked to the armed separatist group ETA, said for the first time that the party opposed street violence as a way to further the Basque separatist cause.*

In this example, the following relevancy scores are calculated relevant to the given topic/query.

PP : "After that incident" : 0.0

PP : "to the armed separatist group ETA" : 0.136

PP : "for the first time" : 0.0

SBAR : "that the party opposed street violence as a way to further the Basque separatist cause" : 0.117

PP : "as a way" : 0.0

If we use a relevancy score threshold of $t = 0$, the resulting sentence will be:

(51c) ~~*After that incident,*~~ *Herri Batasuna, the political party linked to the armed separatist group ETA, said* ~~*for the first time*~~ *that the party opposed street violence* ~~*as a way*~~ *to further the Basque separatist cause.*

4.4 Chapter Summary

In this chapter, we have described our syntactic based sentence pruning techniques. We have implemented three techniques: syntax-driven pruning, syntax with relevancy based pruning and relevancy-driven syntactic pruning. In syntax-driven pruning, we defined linguistically influenced syntactic pruning heuristics. The goal of the syntax-driven technique was to preserve sentence grammaticality while removing syntactic structures that we assumed carrying secondary information. In the second approach, syntax with relevancy based pruning, we toned-down our syntax-driven technique with a relevancy score that we

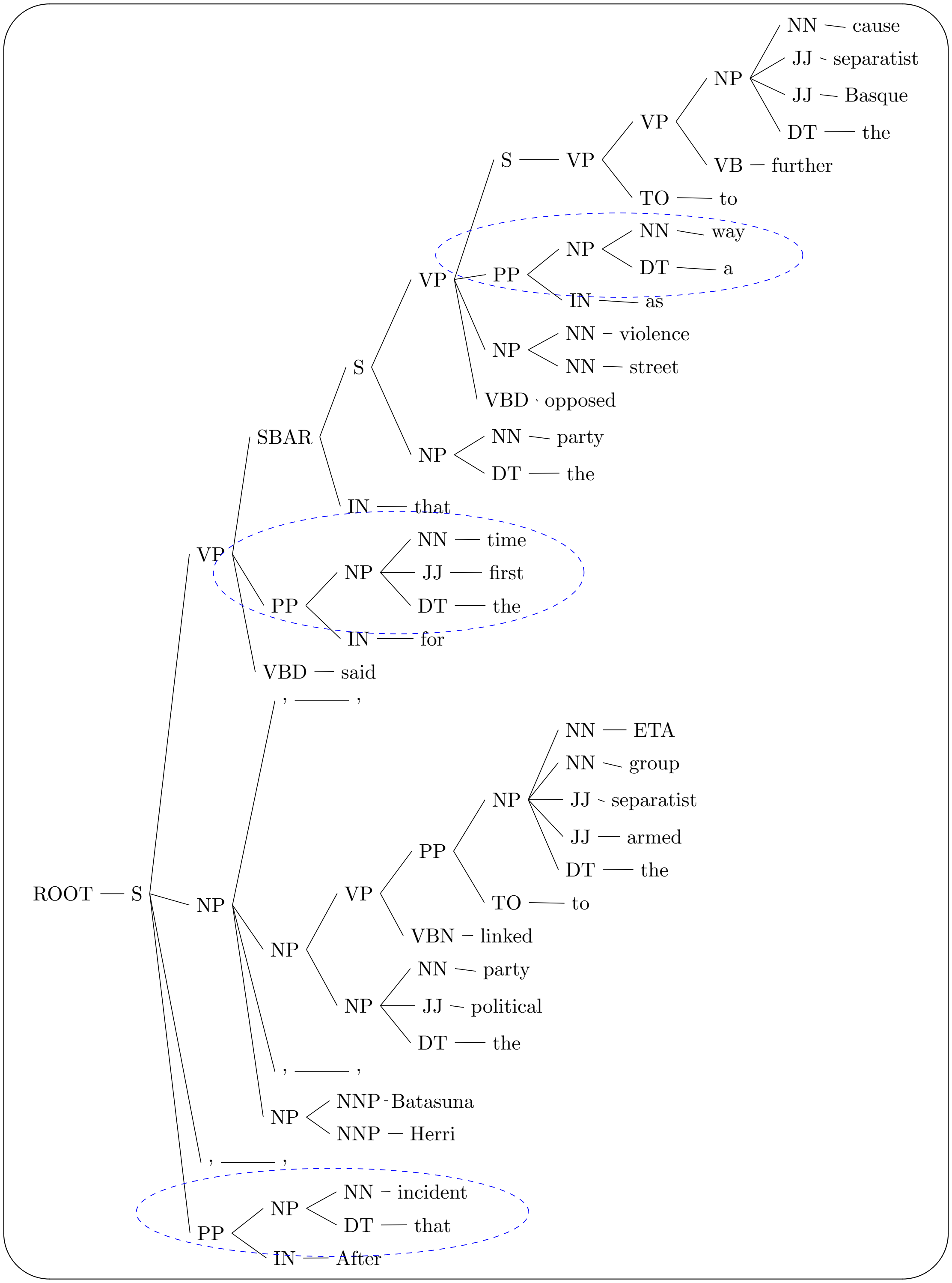

Figure 10: Phrase Structure for Sentence (49).

calculated (as opposed to assume non-relevance). In the third approach, we relied on this relevancy threshold to remove different types of embedding syntactic structures and we were less focused on preserving the grammaticality of the sentences.

Once we developed these techniques, our next goal was to apply these techniques in an automatic text summarization task and evaluate these techniques extrinsically. We have performed both an automatic and a manual evaluation of our techniques. Our next chapter presents the automatic evaluation, while Chapter 6 will present the manual evaluation.

Chapter 5

Automatic Evaluation

To evaluate our pruning techniques extrinsically for the purpose of summary generation, we have used the same standard text corpora we used in our evaluations of the BlogSum summarizer in Chapter 2: the Text Analysis Conference (TAC) 2008, which provides a text corpus created from blogs and the Document Understanding Conference (DUC) 2007 which provides a text corpus of news articles. To ensure that our results were not tailored to one specific summarizer, we used the two summarizer systems we mentioned earlier in Chapter 2: BlogSum [Mithun, 2010], an automatic summarizer based on discourse relations and MEAD [Radev et al., 2004], a generic automatic summarization system. To evaluate each pruning technique, first we generated summaries without any compression. Then we compressed these summaries using the three techniques described in Chapter 4 and compared the results based on three metrics: compression rates, readability measures for complexity of texts and the ROUGE scores for content evaluation.

5.1 Evaluation of Compression Rates

To measure the compression rate of each technique, we first created summaries using BlogSum and MEAD, setting a limit of 250 words per summary, then applied each sentence pruning technique to generate different sets of summaries. Here, we calculated the compression rate as:

$$Compression\ Rate = \frac{No.\ Words\ in\ Compressed\ Text}{No.\ Words\ in\ Original\ Text}$$

In [Knight and Marcu, 2002], the authors have used this ratio to calculate the compression rate and in [Gagnon and Da Sylva, 2006], the authors have used the same measurement but called it the "Reduction Rate". In order to do a standard comparison with the previous work, we calculated the same measure and refer to it as the compression rate as [Knight and Marcu, 2002] do.

5.1.1 Syntax-Driven Pruning

Table 9 shows the compression rates achieved by each heuristic for both summarizers and both datasets. As Table 9 shows, with both datasets, apart from the combined approach, the highest sentence compression was achieved by preposition based pruning (PP pruning); while the lowest compression was observed with relative clause (RC), adverbial phrases (Adv) and conjoined verb phrases (CC-VP) pruning. This is not surprising as PPs are a priori more frequent than the other syntactic constructions. Also not surprisingly, the combined approach which applies all pruning heuristics achieved the highest compression rate in both datasets reaching about 55% to 65% compression rates.

	BlogSum				**MEAD**			
	TAC 2008		**DUC 2007**		**TAC 2008**		**DUC 2007**	
	No of. Words	**Compression Rate**	**No of. Words**	**Compression Rate**	**No of. Words**	**Compression Rate**	**No of. Wordss**	**Compression Rate**
Original	11139	100.0%	10545	100.0%	11341	100.0%	10995	100.0%
Adv Pruning	10710	96.1%	10305	97.7%	11126	98.1%	10729	97.6%
RC Pruning	10653	95.6%	10159	96.3%	10907	96.2%	10437	94.9%
CC-VP Pruning	10744	96.4%	10191	96.6%	11125	98.1%	10532	95.8%
AP Pruning	10787	96.8%	9983	94.7%	11124	98.1%	10293	93.6%
Adj Pruning	10335	92.8%	9786	92.8%	10860	95.7%	10214	92.9%
PP Pruning	9090	81.6%	8252	78.2%	10194	89.8%	8158	74.2%
Combined	7072	63.5%	6473	61.4%	9155	80.7%	6220	56.6%

Table 9: Sentence Compression Rates of Syntax-Driven Pruning.

5.1.2 Syntax and Relevancy Based Pruning

Table 10 shows the compression rate achieved by each heuristic using the syntax and relevancy based pruning. As the results show, with both datasets, the compression effect of each heuristic has been toned down, but the relative ranking of the heuristics are the same. This seems to imply that each type of syntactic phrase is as likely to contain irrelevant information; and one particular construction should not be privileged for pruning purposes. Overall, when all pruning heuristics are combined, the relevancy factor reduces the pruning by about 15 to 20% (from 55-65% to 75-85%)[1].

	BlogSum				**MEAD**			
	TAC 2008		**DUC 2007**		**TAC 2008**		**DUC 2007**	
	No of. Words	**Compression Rate**	**No of. Words**	**Compression Rate**	**No of. Words**	**Compression Rate**	**No of. Wordss**	**Compression Rate**
Original	11272	100.0%	10545	100.0%	11759	100.0%	10995	100.0%
Adv Pruning	10758	96.6%	10318	97.8%	11147	98.3%	10812	98.3%
RC Pruning	10926	98.1%	10442	99.0%	11125	98.1%	10757	97.8%
CC-VP Pruning	10961	98.4%	10414	98.7%	11249	99.2%	10777	98.0%
AP Pruning	10946	98.3%	10378	98.4%	11189	98.6%	10752	97.8%
Adj Pruning	10518	94.4%	9974	94.6%	10925	96.3%	10355	94.2%
PP Pruning	10018	89.9%	9645	91.5%	10644	93.8%	9542	86.8%
Combined	8593	77.1%	8495	80.5%	9913	87.4%	8268	75.2%

Table 10: Sentence Compression Rates of Syntax and Relevancy Based Pruning

5.1.3 Relevancy-Driven Pruning

Table 11 shows the results of the compression rate achieved by relevancy-driven syntactic pruning (see Section 4.3). The relevancy-driven syntactic pruning has achieved a higher compression rate than syntax and relevancy based pruning. Table 12 shows the types of syntactic structures that were removed by the relevancy-driven pruning and their relative frequencies. As the result shows, the most frequent syntactic structures removed were PPs and the least were adverbial phrases (Adv). This result correlates with our two previous

[1] The reduction rate is of course proportional to the relevancy threshold used (see Section 4.2). In this experiment, we set the threshold to be the most conservative (t = 0), hence keeping everything that has any relevance to the topic/query.

	BlogSum				**MEAD**			
	TAC 2008		**DUC 2007**		**TAC 2008**		**DUC 2007**	
	No Of. Words	**Compression Rate**	**No Of. Words**	**Compression Rate**	**No Of. Words**	**Compression Rate**	**No Of. Words**	**Compression Rate**
Original	11272	100.0%	10545	100.0%	11759	100.0%	10995	100.0%
Relevancy-Driven	7457	66.1%	7879	74.0%	7122	60.6%	6801	69.0%

Table 11: Sentence Compression Rates of Relevancy-Driven Syntactic Pruning.

pruning techniques as we achieved similar individual compression rates for these phrase structures.

	BlogSum				**MEAD**			
	TAC 2008		**DUC 2007**		**TAC 2008**		**DUC 2007**	
	No of. Phrases	**Relative Frequency**	**No of. Phrases**	**Relative Frequency**	**No of. Phrases**	**Relative Frequency**	**No of. Phrases**	**Relative Frequency**
PP Pruning	395	50.5%	402	62.4%	177	42.3%	408	63.6%
Other	189	24.1%	136	29.3%	157	31.6%	149	30.1%
RC Pruning	94	12.0%	56	8.7%	44	10.5%	59	9.2%
Adj Pruning	75	9%	35	5.4%	26	6.2%	20	3.1%
Adv Pruning	29	3.7%	15	2.4%	14	3.3%	5	1.0%
Total	782	100%	644	100%	418	100%	641	100%

Table 12: Syntactic Phrase Structures Removed by Relevancy-Driven Pruning.

5.2 Evaluation of Readability Measures

Next we evaluated the compressed summaries using readability measures as we did in Section 2.3.5. Here, we wanted to evaluate how our sentence compression techniques affect these measures. Similarly to Section 2.3.5, we calculated six measures: Flesch-Kincaid Reading Ease, Flesch Kincaid Grade Level, Gunning Fog Score, Coleman Liau Index, SMOG Index and Automated Readability Index. Tables 13 and 14 show the results. According to the results we obtained, the readability measurements seem to agree with

	BlogSum				**MEAD**			
	Original	**Syntax Driven**	**Syntax With Relevancy**	**Relevancy Driven**	**Original**	**Syntax Driven**	**Syntax With Relevancy**	**Relevancy Driven**
Flesch-Kincaid Reading Ease	48.0	58.3	53.4	49.3	45.5	51.2	51.2	56.1
Flesch Kincaid Grade Level	11.0	8.0	9.5	9.8	12.9	11.0	11.0	9.2
Gunning Fog Score	12.2	9.2	10.7	11.0	14.0	12.1	12.1	10.4
Coleman Liau Index	14.0	13.9	13.9	15.0	12.8	13.0	13.0	13.1
SMOG Index	9.9	7.7	8.8	10.2	9.0	9.8	9.8	8.7
Automated Readability Index	11.2	8.0	9.8	10.2	13.5	11.4	11.4	9.4

Table 13: Readability Measures of Different Techniques on the DUC 2007 Dataset.

	BlogSum				**MEAD**			
	Original	**Syntax Driven**	**Syntax With Relevancy**	**Relevancy Driven**	**Original**	**Syntax Driven**	**Syntax With Relevancy**	**Relevancy Driven**
Flesch-Kincaid Reading Ease	60.0	69.6	66.4	64.5	38.5	42.3	41.9	39.0
Flesch Kincaid Grade Level	9.0	6.2	7.2	7.1	15.3	13.9	14.1	14.9
Gunning Fog Score	10.6	7.8	8.6	8.2	14.9	13.2	13.5	13.7
Coleman Liau Index	11.6	11.1	11.3	12.1	12.2	12.3	12.2	11.6
SMOG Index	8.4	6.4	7.1	7.0	11.7	11.0	11.1	10.7
Automated Readability Index	8.8	5.5	6.7	6.6	16.0	14.2	14.5	14.8

Table 14: Readability Measures of Different Techniques on the TAC 2008 Dataset.

the assumption that our three techniques can simplify the summary content, compared to the original texts. The Flesch-Kincaid Reading Ease shows an increase with both datasets, DUC 2007 (Original: 48.00 and 58.28, 53.41, 49.29 with syntax-driven, syntax with relevancy and relevancy driven) and TAC 2008 (Original: 60.04 and 69.65, 66.41, 64.49 with syntax-driven, syntax with relevancy and relevancy driven) on BlogSum summaries. With the other measures, the figures decrease, which implies an improvement in the readability of the text (i.e. in Automated Readability Index, original: 11.22 and 8.04, 9.78, 10.23 with syntax-driven, syntax with relevancy and relevancy driven techniques for the DUC 2007 summaries, created using BlogSum).

5.3 Evaluation of Content

Compression rate and readability measures are interesting, but not at the cost of pruning useful information. In order to measure the effect of the pruning strategies on the content of summaries, we ran the same experiments again but this time we calculated the F-measures of the ROUGE scores (R-2 and R-SU4). In principle, pruning sentences should shorten summaries thus allowing us to fill the summary with new relevant sentences and hence improve its overall content. In order to evaluate the effect of sentence compression on this, we first created summaries with a word limit of 250 and then created two summaries: summaries without filling and summaries with filling.

Summaries Without Filling Here, first we created summary sets of 250 word limit per summary, then compressed them with our sentence pruning techniques and evaluated them using ROUGE. The final summaries we evaluated therefore contained less than 250 words but since we believed to remove secondary information, we expected to see a negligible effect on ROUGE scores.

Summaries With Filling As opposed to the summaries without filling, here we first compressed the summaries and then filled the resulting summaries with extra sentences in order to reach the 250 word limit again. In principle, these summaries contain more content compared to original 250 word limit summaries. As a result, we expected to see a content improvement which would be reflected in ROUGE scores.

5.3.1 Syntax-Driven Pruning

Tables 15 and 16 show the results obtained with and without content filling respectively. Table 15 shows a drop in ROUGE score for both summarization systems and both datasets. This goes against our hypothesis that by default specific syntactic constructions can be removed without losing much content. In addition, when filling the summary with extra sentences, ROUGE scores do seem to improve (as shown in Table 16); however Pearson's χ^2 and t-tests show that this difference is not statistically significant. What is more surprising is that this phenomenon is true not only for the combined heuristics, but also for each individual pruning heuristic as well.

	BlogSum				**MEAD**			
	TAC 2008		**DUC 2007**		**TAC 2008**		**DUC 2007**	
	R-2	**R-SU4**	**R-2**	**R-SU4**	**R-2**	**R-SU4**	**R-2**	**R-SU4**
Original	0.074	0.112	0.086	0.140	0.040	0.063	0.083	0.136
Adv Pruning	0.075	0.113	0.087	0.141	0.040	0.063	0.084	0.136
RC Pruning	0.072	0.109	0.085	0.138	0.039	0.062	0.082	0.134
CC-VP Pruning	0.073	0.111	0.086	0.138	0.040	0.063	0.082	0.134
AP Pruning	0.074	0.112	0.085	0.137	0.038	0.062	0.080	0.133
Adj Pruning	0.068	0.108	0.082	0.138	0.038	0.062	0.077	0.133
PP Pruning	0.064	0.097	0.067	0.114	0.034	0.054	0.064	0.109
Combined	0.057	0.091	0.061	0.108	0.032	0.052	0.053	0.097

Table 15: Content Evaluation of Compressed Summaries with Syntax-Driven Pruning (Without Filling).

5.3.2 Syntax and Relevancy Based Pruning

Recall that syntax-driven pruning did not consider the relevancy of the sub-trees when pruning them. When we do take the relevancy to account; surprisingly the ROUGE scores do not improve significantly either. Tables 17 and 18 show the ROUGE scores of the compressed summaries based on syntax and relevancy without content filling and with content filling. Again any semblance of improvement is not statistically significant.

5.3.3 Relevancy-Driven Pruning

Table 19 shows the results of relevancy-driven pruning with and without content filling and compares them to the original summaries. Again the results are surprisingly low. This last approach was also not able to improve ROUGE scores significantly.

	BlogSum				**MEAD**			
	TAC 2008		**DUC 2007**		**TAC 2008**		**DUC 2007**	
	R-2	**R-SU4**	**R-2**	**R-SU4**	**R-2**	**R-SU4**	**R-2**	**R-SU4**
Original	0.074	0.112	0.086	0.140	0.040	0.063	0.083	0.136
Adv Pruning	0.069	0.110	0.089	0.142	0.041	0.065	0.084	0.138
RC Pruning	0.070	0.107	0.087	0.139	0.040	0.064	0.083	0.136
CC-VP Pruning	0.069	0.108	0.088	0.139	0.041	0.064	0.083	0.136
AP Pruning	0.070	0.108	0.087	0.139	0.040	0.064	0.084	0.137
Adj Pruning	0.065	0.105	0.085	0.140	0.039	0.064	0.081	0.137
PP Pruning	0.067	0.103	0.074	0.125	0.036	0.058	0.073	0.121
Combined	0.055	0.094	0.074	0.127	0.035	0.058	0.074	0.128

Table 16: Content Evaluation of Compressed Summaries with Syntax-Driven Pruning (With Filling).

5.3.4 Discussion

The results of the compression rates we obtained were similar to the work of [Zajic et al., 2007, Gagnon and Da Sylva, 2006]. However, we were surprised at the results of the content evaluation and this might explain why, to our knowledge, so little work can be found in the literature on the evaluation of syntactic sentence pruning for summarization. Our pruning heuristics could of course be fine-tuned to be more discriminating. We could, for example, use verb frames or lexico-grammatical rules to prune PPs; but we do not foresee a significant increase in ROUGE scores. The relevance measure that we used (see Section 4.2) could also be experimented with, but again, we do not expect much increase from that end. Using a better performing summarizer might also be a possible avenue of investigation to provide us with better input sentences and better “filling” sentences after compression.

	BlogSum				**MEAD**			
	TAC 2008		**DUC 2007**		**TAC 2008**		**DUC 2007**	
	R-2	**R-SU4**	**R-2**	**R-SU4**	**R-2**	**R-SU4**	**R-2**	**R-SU4**
Original	0.074	0.112	0.086	0.140	0.040	0.063	0.083	0.136
Adv Pruning	0.074	0.113	0.087	0.141	0.040	0.063	0.084	0.137
RC Pruning	0.039	0.063	0.086	0.139	0.039	0.063	0.083	0.136
CC-VP Pruning	0.074	0.111	0.086	0.138	0.040	0.063	0.082	0.135
AP Pruning	0.074	0.112	0.086	0.134	0.039	0.062	0.083	0.135
Adj Pruning	0.070	0.110	0.084	0.140	0.038	0.063	0.080	0.135
PP Pruning	0.070	0.107	0.082	0.133	0.037	0.059	0.077	0.126
Combined	0.067	0.105	0.081	0.134	0.036	0.059	0.073	0.123

Table 17: Content Evaluation of Compressed Summaries with Syntax with Relevancy Based Pruning (Without Filling).

5.4 Chapter Summary

In this chapter, we have evaluated our three syntactic based sentence pruning methods described in Chapter 4 extrinsically for the task of automatic text summarization. These techniques were applied to the sentences extracted by two different summarizers to generate compressed summaries and evaluated on the TAC 2008 and DUC 2007 benchmarks. According to the results, these pruning techniques generate a compression rate between 60% to 88% which is similar to the previous work [Gagnon and Da Sylva, 2006, Zajic et al., 2007]. Also, we performed an automatic evaluation on the complexity of the compressed sentences using readability measures and they showed that the complexity of the sentences has been reduced by our techniques. However, whether or not we use the extra space to include additional sentences, the content evaluation does not show a significant improvement in ROUGE scores.

	BlogSum				**MEAD**			
	TAC 2008		**DUC 2007**		**TAC 2008**		**DUC 2007**	
	R-2	**R-SU4**	**R-2**	**R-SU4**	**R-2**	**R-SU4**	**R-2**	**R-SU4**
Original	0.074	0.112	0.086	0.140	0.040	0.063	0.083	0.136
Adv Pruning	0.070	0.110	0.088	0.141	0.040	0.065	0.085	0.138
RC Pruning	0.066	0.106	0.087	0.139	0.040	0.065	0.083	0.137
CC-VP Pruning	0.070	0.108	0.088	0.139	0.041	0.065	0.083	0.136
AP Pruning	0.070	0.108	0.087	0.134	0.040	0.064	0.083	0.135
Adj Pruning	0.066	0.106	0.086	0.142	0.040	0.065	0.082	0.138
PP Pruning	0.064	0.104	0.087	0.139	0.039	0.062	0.081	0.132
Combined	0.063	0.105	0.086	0.140	0.037	0.062	0.081	0.134

Table 18: Content Evaluation of Compressed Summaries with Syntax with Relevancy Based Pruning (With Filling).

To investigate further this surprising result, in the next chapter, we will present a manual human evaluation, as [Knight and Marcu, 2002] and [Nguyen et al., 2003] did in their work. The goal of this evaluation is to find out if human assessors agree with ROUGE scores, and thus we need to re-think our syntactic approach or if a human evaluation does consider our condensed summaries to be more informative than the original ones, hence putting aside ROUGE measures for the task (as [Mithun et al., 2012, Dorr et al., 2005, Owczarzak et al., 2012] criticized).

	BlogSum				**MEAD**			
	TAC 2008		**DUC 2007**		**TAC 2008**		**DUC 2007**	
	R-2	**R-SU4**	**R-2**	**R-SU4**	**R-2**	**R-SU4**	**R-2**	**R-SU4**
Original	0.074	0.112	0.088	0.141	0.040	0.063	0.086	0.139
Relevancy-Driven Without Filling	0.065	0.100	0.077	0.125	0.034	0.055	0.066	0.110
Relevancy-Driven With Filling	0.068	0.106	0.083	0.135	0.033	0.060	0.078	0.128

Table 19: Content Evaluation of Compressed Summaries with Relevancy-Driven Syntactic Pruning (With and Without Filling).

Chapter 6

Human Evaluation

Chapter 5 described the evaluation we have performed on our sentence compression techniques using automatic evaluation metrics. We achieved a promising compression rate (60 - 80%) and an increase in readability but our content evaluation using ROUGE did not show any improvement when we compress our summary sentences and added extra content. This surprising result lead us to evaluate if our syntactic pruning techniques were similar to what humans do in sentence compression. Additionally, we wanted to compare our techniques with different types of sentence compression techniques we described in Chapter 3: the keyword based pruning approach and machine learning and classifier based approaches. Finally, we wanted to evaluate how these automatic techniques behaved when compared with human techniques. Also we compared these techniques with a baseline sentence compression of removing random words/phrases.

6.1 Human Gold Standard

To perform the evaluation of human compression techniques, we have provided a set of summaries to five human annotators and asked them to reduce their length while preserving important content. We provided them with a set of summaries created using the DUC 2007 summarization task along with the relevant topic and the query that used to create these summaries. For this task, we have chosen the summaries created by the best performed system [Pingali et al., 2007] (based on ROUGE measures) at the DUC 2007 summarization track. The human annotators were asked to compress these summaries by removing words or phrases from the sentences that they considered as not relevant to the given topic/query. Each sentence was to be considered independent of the others; hence the annotators could not use the context to influence their compression strategies. Human annotators were chosen from a group of undergraduate and graduate students in different science and engineering streams.

6.2 Baseline Compression Techniques

In order to compare our syntactic compression techniques for automatic text summarization, we implemented and used other simple types of sentence compression techniques that we used as baselines. The sections below describe these baseline techniques.

6.2.1 Baseline 1: Random Word Removal

As the first baseline pruning technique, we implemented a technique that randomly removes words and phrases from summaries to reach a particular compression rate. Using this baseline compression, we have created compressed summaries with compression rates of 57%, 70% and 77% to be similar with the compression rates of our automatic sentence pruning techniques: syntax-driven, relevancy-driven and syntax with relevancy based pruning (see Tables 9, 11 and 10).

6.2.2 Baseline 2: Keyword/Phrase Based Approach

Aside from our three compression techniques which are based on syntactic pruning, we were influenced by the work of [Conroy et al., 2006] and implemented and evaluated a keyword based sentence compression technique. For this work, we used the word/phrase patterns described in [Conroy et al., 2006] and [Dunlavy et al., 2003] plus additional patterns that we learned by analyzing the human annotated summaries. The particular keyword/phrase patterns and syntactic patterns we used are described below.

Removing Meta-Data Information This includes removing date information, editor's comments or specific tags which appear specifically at the start or end of a sentence. These specific words/tags were present in the summary sentences due to the generic sentence extraction methods used by the summarizers. In specific corpora, there are specific words/tags which have been added in their documents/articles and sometimes they end up in extractive summaries as well. For an example consider following sentence.

> (52) *(AP) The two public university systems in the nation's most populous state have agreed to require the same high school course work for admission.*

This sentence was taken from DUC 2007 and the tag *"AP"* indicates that the article comes from the Associated Press[1]. So we removed these meta-data tags as a part of sentence compression.

> (52c) ~~*(AP)*~~ *The two public university systems in the nation's most populous state have agreed to require the same high school course work for admission.*

Removing Temporal Words/Phrases In specific corpora, the documents may contain temporal words/phrases that, without proper context, can be considered not useful. These include specific years, months or relative temporal information. For an example, consider the following sentence:

> (53) *They included an incident earlier this year, when five Mossad agents were caught trying to bug a house in Bern, Switzerland.*

The phrase *"earlier this year"* provides relative temporal information specific to the particular year the event occurred. Without the grounding temporal information, this relative temporal information cannot be dereferenced. So in our sentence compression, we decided to remove all temporal words/phrases.

> (53c) *They included an incident* ~~*earlier this year*~~*, when five Mossad agents were caught trying to bug a house in Bern, Switzerland.*

[1] http://www.ap.org/

Removing Attributive Words/Phrases In some documents, especially in news articles or reports, some sentences provide personal comments along with the source of information. As an example, consider the following sentences:

> (54) *German Economics Minister Guenter Rexrodt said today that most German companies are not prepared for the shift to the single European currency, the Euro, which is due to be launched on January 1, 1999.*

In this particular sentence, the phrase *"German Economics Minister Guenter Rexrodt said today that"* is referencing the information provided by him. We can generally remove these attributions to make sentences shorter.

> (54c) ~~*German Economics Minister Guenter Rexrodt said today that*~~ *most German companies are not prepared for the shift to the single European currency, the Euro, which is due to be launched on January 1, 1999.*

Removing Keywords/Keyphrases In this particular technique, we have used a list of specific words and phrases to remove as a way of compressing sentences. This list of words was created using phrases, specific adverbs, adjectives, idioms and conjunctions. The following sentences show some of the phrases we used.

> (55) *As a result, they possibly wouldn't be able to do transactions in Euro from the day it takes effect.*

> (56) *In contrast to Burma, many Chinese reformers welcome Western political and commercial engagement with their government as a spur to further openness and change.*

> (57) *The lawsuit asserts that American Home Products Corp. underreported the instances of pulmonary hypertension*
> *as a result of using fenfluramine, the "fen" part of the drug combination.*

Here, in these three sentences, we have removed *"As a result"*, *"In contrast to Burma"* and *"as a result of using fenfluramine, the "fen" part of the drug combination"*. We used a keyword/phrase list (here, the phrases are: *"as a result"*, *"in contrast"*) and identified the enclosing phrase structures to remove them without losing the grammaticality of the sentence. The resulting sentences are as below:

> (55c) *~~As a result,~~ they possibly wouldn't be able to do transactions in Euro from the day it takes effect.*

> (56c) *~~In contrast to Burma,~~ many Chinese reformers welcome Western political and commercial engagement with their government as a spur to further openness and change.*

> (57c) *The lawsuit asserts that American Home Products Corp. underreported the instances of pulmonary hypertension ~~as a result of using fenfluramine, the "fen" part of the drug combination.~~*

The keyword list contains 169 phrases and can be found in Appendix B. This keyword list was created using the phrases listed in the work of [Conroy et al., 2006, Dunlavy et al., 2003] and through online thesaurus[2].

Removing Specific Clauses In this technique, we have filtered out some specific clauses such as appositive and relative clauses which contain specific keywords/phrases. For this, we considered appositive clauses that contain gerund verbs and relative clauses which start with the words *"which"*, *"whom"*, *"when"* and *"where"*. As an example, consider the following two sentences:

> (58) *Since co-founding the Southern Poverty Law Center in 1971, Dees has wielded the civil lawsuit like a buck knife, carving financial assets out of hate group leaders who inspire followers to beat, burn and kill.*

> (59) *Lawyer Morris Dees, the co-founder of the Southern Poverty Law Center who is representing Victoria Keenan and her son, Jason, introduced letters, photographs and depositions to contradict the men's testimony.*

[2] http://www.http://thesaurus.com/

Here in sentence (58), the clause *"Since co-founding the Southern Poverty Law Center in 1971"* contains the gerund *"co-founding"* and the whole clause becomes a candidate to be pruned to generate a shorter sentence. So the resulting sentence will be:

> (58c) *~~Since co-founding the Southern Poverty Law Center in 1971,~~ Dees has wielded the civil lawsuit like a buck knife, carving financial assets out of hate group leaders who inspire followers to beat, burn and kill.*

On the other hand in sentence (59), the relative clause, *"who is representing Victoria Keenan and her son, Jason"* starts with the word, *"who"* and becomes a candidate to be pruned according to our sentence compression rule. So the resulting sentence would be:

> (59c) *Lawyer Morris Dees, the co-founder of the Southern Poverty Law Center ~~who is representing Victoria Keenan and her son, Jason,~~ introduced letters, photographs and depositions to contradict the men's testimony.*

With these rules, we have implemented our baseline keywords and phrase structure based pruning technique. As we see, these are not syntactic or machine learning and classification based rules but rather phrase patterns learned by human analysis. So these techniques do not contribute much to the field of Natural Language Processing. We have implemented these techniques as simple baselines and evaluated them along with our syntactic driven techniques and human annotations.

6.2.3 Baseline 3: Machine Learning and Classifier Based Approach

As the third baseline technique, we wanted to compare our syntactically driven sentence compression approaches with an existing implementation of machine learning and classifier based sentence compression technique. For this, we chose the sentence compression system presented in [Galanis and Androutsopoulos, 2010]. This recent work on sentence compression was described in section 3.2.6 which uses a Maximum Entropy based classifier to prune sentences and a Support Vector Regression Model to select the best candidates of all reduced sentences. This project was implemented in the Department of Informatics,

Athens University of Economics and Business and the system is publicly available[3] under GNU General Public License (GPL). We have used this system to compare our syntactic pruning techniques as the third baseline technique. The system requires a training set of original sentences paired with their human written compressed sentences and they have used the Edinburgh's Written and Spoken corpus[4] for this task. The system comes with an initial configuration of a training set of 50 documents (963 sentences paired with their compressed forms) from this corpus and for our task, we used the same dataset for the training task. Additionally, they have used a language model created using about 4.5 million sentences taken from TIPSTER corpus [Harman and Liberman, 1993] and we also used a similar language model built using about 5 million sentences, taken from the same corpus. Finally, they have used an "importance list" which contains words/phrases and a significant score, generated using part of the TIPSTER corpus [Harman and Liberman, 1993] to identify more relevant words and phrases. We used the same list they provided in order to compress our summary set.
For this baseline system, we have created two configurations:

(1) Baseline 3: Machine Learning 1: was created from a training set of 963 instances.

(2) Baseline 3: Machine Learning 1 Reduced 7%: was created from a training set of 680 instances and resulted a higher compression rate.

Our objective in having the second configuration was to bring the compression rate of the system closer to the human compression rate.

The sections below present our evaluations and the results.

6.3 Human Compression Rate

As with our previous evaluations (see Chapter 5), we first evaluated the compression rates of human annotations (as in Section 5.1) and compared these results with the compression

[3] http://nlp.cs.aueb.gr/software.html
[4] http://jamesclarke.net/research/resources

rates achieved by our syntactic based pruning techniques. Second, we analyzed the relative frequencies of the syntactic structure removed by annotators and the corresponding compression rate achieved from these different categories. Finally, we evaluated the percentage of each syntactic structures removed over the total syntactic structures found in given summaries for human annotations and automatically compressed summaries. Table 20 shows the compression rate that each annotator has achieved and the average compression rate of all the annotators. According to Table 20, the compression rates achieved by the five annotators ranged between 75% to 82% with an average compression rate of 78.5%. On the other hand, Table 21 compares average compression rate of human annotations with the compression rates we have achieved by the different syntactic pruning techniques described in Chapter 4 and the two baseline techniques, Baseline 2 and 3. According to Table 21, the highest compression rate for the given summary set was achieved by the syntax-driven technique. Next, the relevancy-driven technique and the syntax with relevancy based techniques achieved the next highest compressions. The annotator average (78.5%) seems to be similar to syntax with relevancy based technique (76.6%). Other than that, the lowest compressions were achieved by Baseline 3 (82.1%) and Baseline 2 (89.5%) techniques. The keyword based sentence pruning technique could not achieve a compression rate that was

	No. of Words	**Compression**
Original	6237	100.0%
Annotator 1 (Dritan)	5106	81.9%
Annotator 2 (Zoe)	5052	81.0%
Annotator 3 (Reda)	4897	78.5%
Annotator 4 (Felix)	4889	78.4%
Annotator 5 (Ishrar)	4657	74.7%
Average	4920	78.5%

Table 20: Sentence Compression Rates of Annotations

Different Techniques	**No. of Words**	**Compression**
Original	6237	100.0%
Annotator Average	4920	78.5%
Syntax-Driven	3552	56.9%
Syntactic with Relevancy	4779	76.6%
Relevancy-Driven	4381	70.2%
Baseline 2: Keyword Based Pruning	5579	89.5%
Baseline 3: Machine Learning Technique	5122	82.1%
Baseline 3: Machine Learning Technique Reduced 7%	4772	76.5%

Table 21: Sentence Compression Rates of Different Techniques

similar to human annotations but Baseline 3, the machine learning based sentence compression could achieve a better compression that was more similar to human annotations and also, as we observed, when the training set was reduced by 7% (680 training instances as opposed to 963), it achieved a compression rate of 76.5% that is much closer to average human compression rate.

6.4 Human Pruning of Syntactic Structures

Once we obtained the manually compressed summaries, we were curious to see which types of words and phrase structures the annotators removed. Therefore, we marked all the words and phrases they removed and categorized them according to their grammatical classes. In general, our human annotators tended to remove the following grammatical phrase structures:

(1) Individual Words

(2) Noun and Verb Phrases

(3) Conjoined Clauses

(4) Appositive Phrases

(5) Adverbial Phrases

(6) Adjective Phrases

(7) Relative Clauses

(8) Prepositional Phrases

Our next evaluation focused on calculating the distribution of each type of structure removed by the annotators. First we calculated the ratio of each removed structure out of all grammatical structures that were removed by the annotators. Table 22 shows these results. According to Table 22, the most frequent structures removed by the annotators are prepositional phrases (around 20-26%) and noun phrases (around 20 to 33%). The least frequent structures removed are conjoined clauses (1.2 to 6%) and adverbial phrases (around 5%). Next we calculated the relative frequency of each syntactic structure removed compared to all the grammatical structures in the dataset. For example, out of all PPs, how many were removed and how many were kept in the given summary set. We have calculated these results on the human compressed summaries and also on the syntax driven and syntax with relevancy based pruning techniques. Tables 23 and 24 show these results.

According to Table 23, of all appositive phrases, human annotators removed between 25.3% to 45.5% of the phrases. The least removed seems to be the conjoined clauses where only 4.3% to 12.3% were removed. This information is interesting compared to our automatic techniques. For example, as Table 24 shows, our syntax-driven technique tends to remove most of the adverbial phrases and a very high percentage of prepositional phrases attached to noun phrases[5]. with the syntax with relevancy based technique, the syntactic

[5] In syntax-driven technique, some of the phrases were shown to be not removed completely but in fact they were removed while they were enclosed by other types of phrases those were removed before. For example, when a prepositional phrase encloses an adjective phrase, it considered that only the prepositional phrase was removed but not the adjective phrase

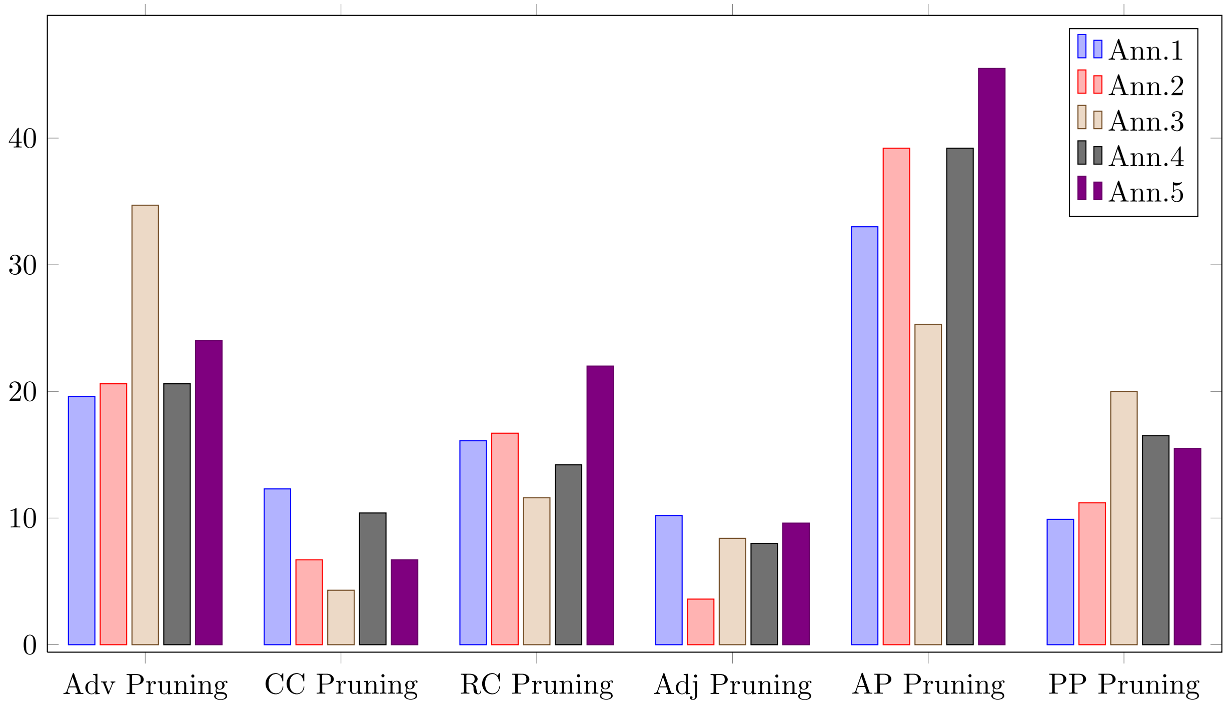

Figure 11: Proportions of Syntactic Structures Removed by Human Annotators.

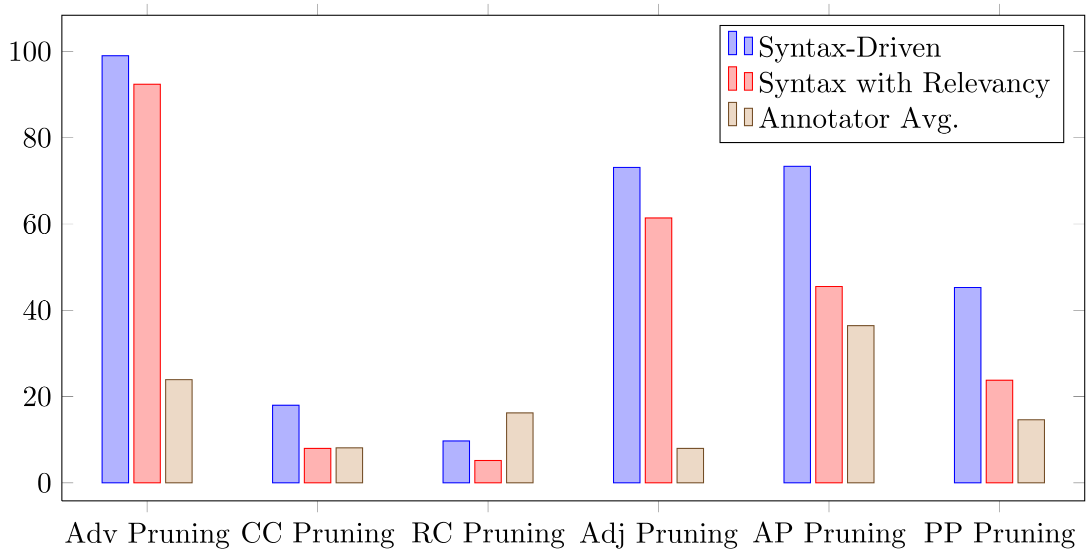

Figure 12: Proportions of Syntactic Structures Removed by Our Syntactic Pruning Techniques.

Syntactic Structures	**Relative Frequency**					
	Ann. 1	**Ann. 2**	**Ann. 3**	**Ann. 4**	**Ann. 5**	**Average**
Individual Words	9.3%	4.4%	11.8%	7.9%	6.3%	7.9%
Adverbial Phrases	5.4%	6.4%	5.5%	4.2%	5.1%	5.3%
Conjoined Clauses	6.0%	3.7%	1.2%	3.7%	2.5%	3.4%
Relative Clauses	7.5%	8.7%	3.1%	4.8%	7.9%	6.4%
Adjective Phrases	15.4%	6.1%	9.5%	8.8%	11.1%	10.2%
Appositive Phrases	7.8%	10.4%	3.5%	6.8%	8.4%	7.4%
Verb Phrases	7.5%	9.4%	8.3%	8.6%	6.5%	8.1%
Noun Phrases	20.0%	25.0%	33.0%	30.0%	27.0%	27.0%
Prepositional Phrases (Total)	20.5%	25.9%	24.0%	25.0%	25.0%	24.1%
PP attached to NP	6.6%	8.4%	10.2%	10.0%	9.3%	8.9%
PP attached to VP	9.3%	11.4%	10.6%	11.4%	11.2%	10.8%
PP attached to Clauses	4.5%	6.1%	3.1%	3.7%	4.4%	4.4%
Total	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Table 22: Syntactic Structures Removed by Human Annotators.

structures with the highest pruning rates are adverbial phrases and adjective phrases. As we can see, our techniques seem to be harsh on removing adverbial and adjective phrases. There is a considerable difference in the number of these types of structures removed by our techniques compared to the number which human annotators had removed. Indeed, our techniques remove almost all adverbial phrases and 61% to 73% of adjective phrases while human annotators were more discriminating and removed only 24% and 8% respectively. Also, it is important to note that these ratios were calculated based on the frequency of independent grammatical structures in the summaries. So in some cases, for longer phrase structures, what human annotators removed can be partially similar to what our techniques pruned yet the result can be different in terms of grammatical structures we counted. For an example, consider the following sentence and its compressed forms:

Syntactic Structures	**Relative Frequency**					
	Ann. 1	**Ann. 2**	**Ann. 3**	**Ann. 4**	**Ann. 5**	**Ann. Avg.**
Adverbial Phrases	19.6%	20.6%	34.7%	20.6%	24.0%	23.9%
Conjoined Clauses	12.3%	6.7%	4.3%	10.4%	6.7%	8.1%
Relative Clauses	16.1%	16.7%	11.6%	14.2%	22.0%	16.2%
Adjective Phrases	10.2%	3.6%	8.4%	8.0%	9.6%	8.0%
Appositive Phrases	33.0%	39.2%	25.3%	39.2%	45.5%	36.4%
Prepositional Phrases (Total)	9.9%	11.2%	20.0%	16.5%	15.5%	14.6%
PP attached to NP	7.6%	12.2%	28.8%	22.0%	19.5%	18.0%
PP attached to VP	8.3%	9.1%	16.3%	13.9%	12.8%	12.8%
PP attached to Clauses	68.2%	81.8%	81.8%	77.3%	86.4%	79.1%

Table 23: Proportion of Syntactic Structures Removed by Human Annotators.

(60) Original Sentence: *The Southern Poverty Law Center won major legal fights against the Ku Klux Klan and other white supremacist groups.*

(60c1) *The Southern Poverty Law Center won major legal fights ~~against the Ku Klux Klan and other white supremacist groups.~~*

(60c2) *The Southern Poverty Law Center won major legal fights against the Ku Klux Klan ~~and other white supremacist groups.~~*

Here, in this example, sentence (60c1) and sentence (60c2) are the results of two human annotators. Both sentences partially agree in terms of the content removed. However, in sentence (60c1), the grammatical structure removed is a *PP* (prepositional phrase) but in sentence (60c2), it is a *CC* (conjoined clause). So because of that, though the content removed agreed partially, the grammatical structures removed can be completely different. Figure 11 shows the overall distribution of the syntactic structures that human annotators removed while Figure 12 shows the distribution of the syntactic structures removed by our syntactic pruning heuristics, compared to the annotator's average. As the Figures clearly

Syntactic Structures	**Relative Frequency**		
	Syntax Driven	**Syntax With Relevancy**	**Annotator Avg.**
Adverbial Phrases	99.0%	92.4%	23.9%
Conjoined Clauses	18.0%	8.0%	8.1%
Relative Clauses	9.7%	5.2%	16.2%
Adjective Phrases	73.1%	61.4%	8.0%
Appositive Phrases	73.4%	45.5%	36.4%
Prepositional Phrases (Total)	43.5%	23.8%	14.6%
PP attached to NP	90.1%	47.5%	18.0%
PP attached to VP	4.7%	3.8%	12.8%
PP attached to Clauses	45.5%	27.3%	79.1%

Table 24: Proportion of Syntactic Structures Removed by Syntactic Pruning Techniques.

show, humans are more subtle in the types of structures that they remove.

Lastly, for the human annotations and our pruning techniques: syntax-driven and syntax with relevancy based pruning, we have calculated the compression rate achieved by removing each syntactic structures. According to Tables 25 and 26, all the annotators have achieved the highest compression rate by removing prepositional phrases and the least were obtained by removing individual words, adverbial phrases and verb phrases. In our syntactic pruning techniques, the highest compression was achieved by removing prepositional phrases and the least compression was achieved by adverbial and relative clauses. Compared with human annotations, our techniques do not remove individual words, verb or noun phrases. In addition, our techniques achieved a higher proportion of compression based on prepositional phrase removals compared to other syntactic structures. From these results, we can see that humans tend to remove the same syntactic structures but the numbers and proportions they contributed to the overall compression seem to be subtle compared to our techniques.

Syntactic Structures	**Relative Frequency**					
	Ann. 1	**Ann. 2**	**Ann. 3**	**Ann. 4**	**Ann. 5**	**Ann. Avg**
Individual Words	2.8%	1.2%	5.3%	2.7%	1.7%	2.7%
Adverbial Phrases	2.3%	2.7%	3.2%	2.1%	2.0%	2.5%
Conjoined Clauses	10.6%	6.5%	2.1%	6.4%	6.5%	6.4%
Relative Clauses	20.4%	18.2%	9.0%	14.6%	22.6%	17.0%
Adjective Phrases	5.3%	1.9%	5.2%	3.3%	3.8%	4.0%
Appositive Phrases	14.7%	21.2%	7.9%	14.5%	15.2%	14.7%
Verb Phrases	2.3%	2.9%	3.8%	3.2%	2.3%	2.9%
Noun Phrases	9.8%	9.8%	24.7%	18.0%	13.5%	15.6%
Prepositional Phrases (Total)	31.7%	35.5%	38.8%	35.1%	32.4%	34.7%
PP attached to NP	9.2%	9.6%	14.6%	12.5%	7.0%	10.6%
PP attached to VP	14.0%	15.5%	16.5%	17.2%	14.1%	15.5%
PP as Clauses	8.5%	10.4%	7.6%	5.4%	11.2%	8.6%
Total	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Table 25: Sentence Compression by Human Annotators Based on Syntactic Structures.

6.4.1 Comparison with Our Heuristics

As shown in the previous section, apart from the individual words, Noun and Verb phrases, the human annotations seem to remove the same syntactic structures as our sentence pruning techniques (but with a more subtle selection). So further we have analyzed the categories which differ from our heuristics.

Pruning Individual Words In the human compressed summaries, a small percentage of individual words are removed. These removed words contribute very little to the overall compression rate. Frequently, these individual words were removed to preserve the grammaticality after removing other syntactic structures. The sentence below shows an example:

Syntactic Structures	**Relative Frequency**		
	Syntax-Driven	**Syntax With Relevancy**	**Ann. Avg**
Individual Words	0.0%	0.0%	2.7%
Adverbial Phrases	4.3%	7.0%	2.5%
Conjoined Clauses	6.7%	5.4%	6.4%
Relative Clauses	4.5%	3.8%	17.0%
Adjective Phrases	15%	22.4%	4.0%
Appositive Phrases	15.7%	17.6%	14.7%
Verb Phrases	0.0%	0.0%	2.9%
Noun Phrases	0.0%	0.0%	15.6%
Prepositional Phrases (Total)	53.7%	43.7%	34.7%
PP attached to NP	47.3%	37.2%	10.6%
PP attached to VP	2.3%	3.1%	15.5%
PP as Clauses	4.2%	3.4%	8.6%
Total	100.0%	100.0%	100.0%

Table 26: Sentence Compression by Human Annotators Based on Syntactic Structures.

(61) *This is Inner City Arts, a nonprofit arts school that is both an enlightened model for arts education and a design landmark where education is embellished by architectural example.*

Here, the words *"that"* and *"both"* were removed by humans as a consequence of removing the relative clause *"where education is embellished by architectural example"*.

Our heuristics do not remove individual words for two main reasons: first, as we mentioned earlier, the contribution of their removal to the overall compression rate is minimal (in the order of 2.7%). Second, their removal is dependent on the removal of other syntactic structures. This in itself is difficult to implement and may result more mistakes than correct removals.

6.4.2 Pruning Noun and Verb Phrases

In human compressed summaries, a negligible number of verb phrases were removed. On the other hand, human annotators have pruned a considerable number of noun phrases. After analyzing these noun phrases that were removed, we identified two main categories of noun phrases.

Pruning Proper and Compound Nouns Given a summary and a topic/query, human annotators seem to prune specific proper and compound nouns based on their level of knowledge. This seems to be subjective for each individual and reflects the annotator's knowledge and perception of the world. As an example, consider the following sentence, pruned by three different annotators:

> (62) Annotator A: *Myanmar's military government has detained another 187 members of pro-democracy leader Aung San Suu Kyi's party, bringing the total to 702 arrested since a crackdown began in May.*
>
> (63) Annotator B: *Myanmar's military government has detained another 187 members of pro-democracy leader Aung San Suu Kyi's party, bringing the total to 702 arrested since a crackdown began in May.*
>
> (64) Annotator C: *Myanmar's military government has detained another 187 members of pro-democracy leader Aung San Suu Kyi's party, bringing the total to 702 arrested since a crackdown began in May.*

Here, Annotator A has only removed the adjectival phrase *"pro-democracy"*; while, Annotator B has gone a bit further and removed *"pro-democracy leader"*. Finally, Annotator C attempted to remove the entire phrase *"pro-democracy leader Aung San"* leaving the remaining phrase, *"Suu Kyi"*. This choice seems to be completely subjective and more influenced by the individuals. This is very hard to implement automatically.

Pruning Temporal Noun Phrases Other than proper and compound noun removal described above, human annotators have removed temporal information in the summaries as well. These temporal expressions can be specific dates, months, years and relative temporal word phrases such as *"yesterday"*, *"today"*, *"last week"* etc. As an example, consider the following sentence:

> (65) *Turkey said Wednesday it wants the EU to make it a formal candidate for membership at its summit later this week.*

Here, the word *"Wednesday"* represents the temporal information about the event and it has been considered as not relevant and removed by annotators. This can be used to modify our heuristics.

6.5 Content Evaluation

Once we have analyzed what type of syntactic structures humans tend to remove, we wanted to evaluate the content of their compressed summaries. To evaluate the content of the pruned summaries, we calculated and compared ROUGE scores (R-2 and SU4) for the original summaries and the five sets of human compressed summaries. Recall from Section 6.1 that the original summaries were created using the output of the best scoring system at DUC-2007. Tables 27 and 28 show the results obtained. Surprisingly, according to Table 27, there is a decrease in ROUGE-2 score between the original summaries and the

	R-2	**R-SU4**
Original	0.127	0.179
Annotator 1 (Mike)	0.119	0.172
Annotator 2 (Zoe)	0.125	0.176
Annotator 3 (Reda)	0.119	0.171
Annotator 4 (Felix)	0.119	0.173
Annotator 5 (Ishrar)	0.118	0.170
Average of Annotators	0.120	0.172

Table 27: Content Evaluation of Human Annotations.

human compressed summaries. On average, annotators have a ROUGE-2 score of 0.120 and ROUGE-SU4 of 0.172 while the original summaries have a ROUGE-2 score of 0.127 and ROUGE-SU4 of 0.179. We have used the one-tailed t-test for each individual annotation and the averaged annotation ROUGE scores to test for significance. The t-test shows that for all the annotators, the difference between ROUGE scores compared to the original summary is statistically significant with a confidence level of 95%. For ROUGE-SU4, all the annotators cause a statistically significant decrease in scores with a confidence level of 95%, except the Annotators 1 and 4. In Table 28, we have compared the average ROUGE scores of human annotators with the ROUGE scores obtained by our pruning techniques and all the baseline techniques (see Section 6.2). Here, we clearly see four clusters of ROUGE-2 scores. The first cluster contains the techniques that scored the best ROUGE-2 scores and evidently the highest of them is the original summaries. The next highest in the same cluster is Baseline 2, the keyword based technique and the last in that cluster is the average ROUGE scores of human compressed summaries. Compared to the original

Different Techniques	**R-2**	**R-SU4**
Original	0.127	0.179
Baseline 2: Keyword Based	0.124	0.176
Average Human Compression	0.120	0.171
Syntax with Relevancy Based	0.110	0.164
Baseline 3: Machine Learning Based	0.110	0.165
Baseline 3: Machine Learning Based 7% Reduced	0.110	0.163
Relevancy-Driven	0.106	0.154
Baseline 1: Random Compression 77%	0.101	0.163
Baseline 1: Random Compression 70%	0.085	0.150
Syntax-Driven	0.084	0.134
Baseline 1: Random Compression 57%	0.072	0.137

Table 28: Content Evaluation of Compressed Summaries.

ROUGE-2 score, the keyword based technique (ROUGE-2: 0.124) does not show a significant decrease in ROUGE score according the one-tailed t-test with a confidence level of 95%. But for the human average score (ROUGE-2: 0.120), the ROUGE-2 score shows a significant decrease compared to the original summaries (with a confidence level of 95%).
In the second cluster, we have three techniques: Syntax with relevancy based pruning, machine learning and classifier based sentence compression technique with two different compression rates (82% and 76.5%). All the techniques in this cluster show a significant decrease in ROUGE-2 scores compared to the original summaries. When compared with average human annotation ROUGE scores, all three techniques seem to have significantly lower ROUGE scores[6]. When compared with each technique in the same cluster (syntax with relevancy (with ROUGE-2: 0.110), machine learning and classifier based techniques with (ROUGE:0.110)), the ROUGE-2 scores are not significantly different from each other.
In the third cluster, we have techniques that scored lower ROUGE scores compared to the second cluster. They are Relevancy-Driven and Baseline 1: Random compression 77%. These two techniques show significantly lower ROUGE-2 scores than the original, average human ROUGE scores and also compared to the Baseline 3: machine learning and classifier based technique. However, when tested for significance in difference using one-tailed t-test between these two techniques, the ROUGE-2 scores are not significantly different from each other with a confidence level of 95%.

In the last cluster, we have the rest of the techniques: Baseline 1: Random Compression 70%, syntax-driven and Baseline 1: Random Compression 57% that scored the lowest ROUGE scores. These three techniques have ROUGE scores those are significantly lower than the original ROUGE score, average human ROUGE score and also lower than relevancy-driven technique's ROUGE score. When tested for significance of the difference compared to the Baseline 1: Random Compression 77%, both techniques seem to have significantly lower ROUGE scores. However, all three techniques have relatively equal scores (Random Compression 70%: 0.085, syntax-driven: 0.084 and Baseline 1: Random Compression 57%: 0.72).

[6] Tested with a one-tailed t-test with a significant level of 95%

The ranking of the techniques is more or less the same when ROUGE-SU4 scores were used for the task.

6.6 Inter-Annotator Content Evaluation

For the next evaluation, we decided to use the human compressed summaries as our gold standard summaries and calculate ROUGE scores for the pruning techniques and the three baseline techniques. Also we have calculated ROUGE scores for human annotators against other annotators. Table 29 shows the numerical results while Figures 13 and 14 show the same results graphically. According to Figure 29, the highest inter-annotator ROUGE-2 scores are achieved by the average annotator's ROUGE scores and keyword based techniques. Then the techniques, syntax with relevancy based pruning, Baseline 3: machine learning and classifier based techniques and relevancy-driven techniques form a second cluster of similar ROUGE scores. These ROUGE scores are significantly lower compared to the highest scored techniques. The third cluster consists of syntax-driven pruning and

Different Techniques	**Ann. 1**		**Ann. 2**		**Ann. 3**		**Ann. 4**		**Ann. 5**	
	R-2	**R-SU4**	**R-2**	**R-SU4**	**R-2**	**R-SU4**	**R-2**	**R-SU4**	**R-2**	**R-SU4**
Syntax-Driven	0.600	0.570	0.578	0.551	0.560	0.537	0.584	0.558	0.592	0.562
Baseline 1: Random Compression 57%	0.390	0.471	0.397	0.470	0.377	0.464	0.379	0.462	0.380	0.459
Relevancy-Driven	0.666	0.622	0.687	0.640	0.642	0.606	0.660	0.616	0.658	0.616
Baseline 1: Random Compression 70%	0.533	0.588	0.536	0.587	0.496	0.555	0.511	0.564	0.503	0.557
Syntactic with Relevancy	0.708	0.688	0.709	0.692	0.676	0.675	0.696	0.678	0.692	0.669
Baseline 1: Random Compression 77%	0.609	0.643	0.601	0.636	0.570	0.612	0.583	0.617	0.575	0.605
Baseline 2: Keyword Based	0.836	0.813	0.838	0.813	0.788	0.758	0.801	0.768	0.801	0.771
Baseline 3: Machine Learning Based	0.706	0.686	0.699	0.681	0.675	0.655	0.686	0.665	0.678	0.660
Baseline 3: Machine Learning Based Reduced 7%	0.673	0.649	0.675	0.653	0.647	0.625	0.661	0.641	0.658	0.634
Annotator 1	*1.000*	*1.000*	0.812	0.788	0.779	0.753	0.801	0.775	0.773	0.744
Annotator 2	0.812	0.789	*1.000*	*1.000*	0.795	0.770	0.818	0.790	0.803	0.774
Annotator 3	0.778	0.753	0.795	0.770	*1.000*	*1.000*	0.788	0.761	0.766	0.740
Annotator 4	0.801	0.774	0.818	0.790	0.788	0.761	*1.000*	*1.000*	0.800	0.772
Annotator 5	0.773	0.774	0.803	0.774	0.766	0.740	0.799	0.772	*1.000*	*1.000*

Table 29: Inter-Annotator Content Evaluation of Compressed Summaries Using ROUGE.

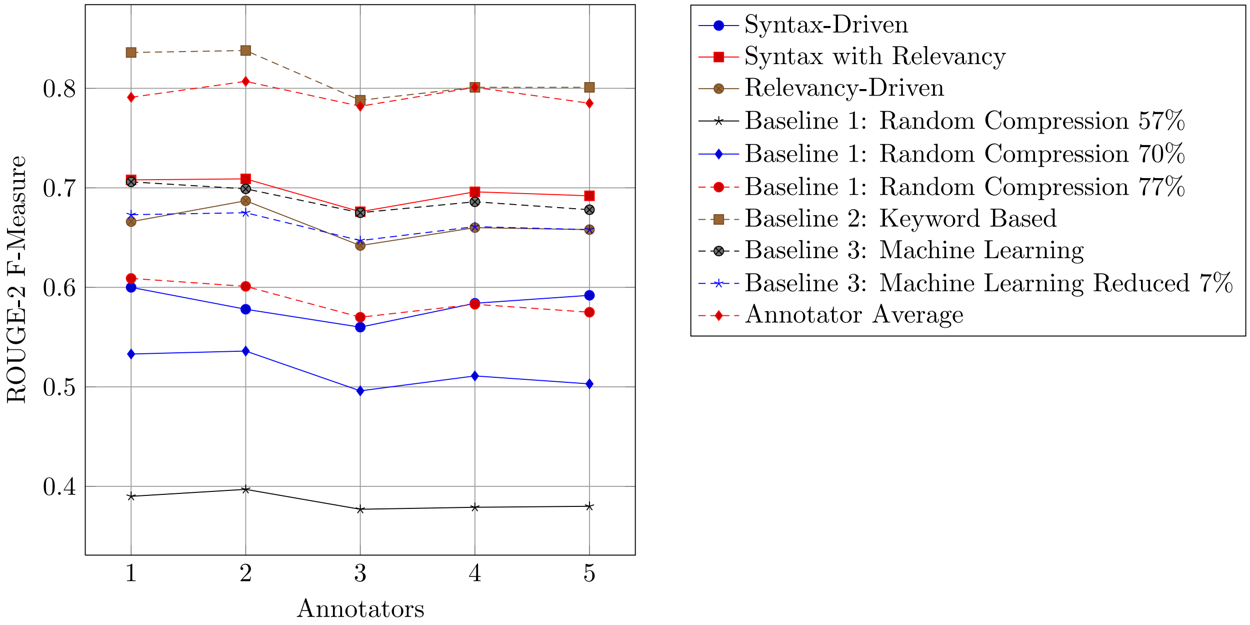

Figure 13: Inter-Annotator Content Evaluation Results of ROUGE-2 Scores.

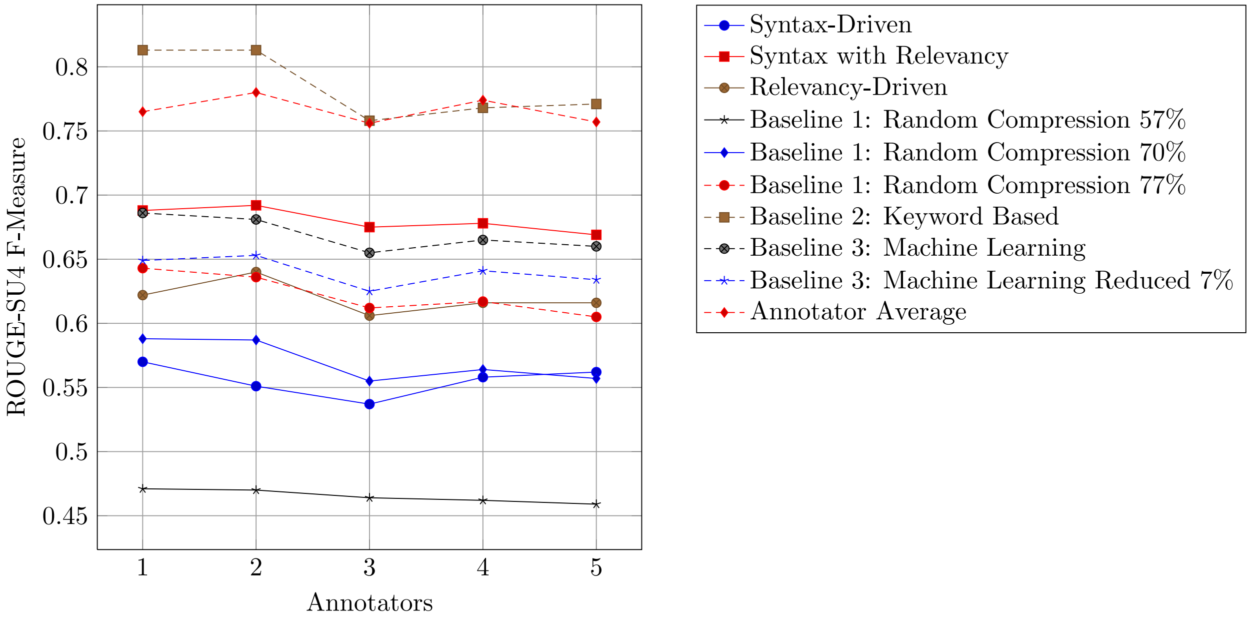

Figure 14: Inter-Annotator Content Evaluation Results of ROUGE-SU4 Scores.

Different Techniques	**R-2**	**R-SU4**
Baseline 2: Keyword Based	0.813	0.785
Annotator 2	0.807	0.780
Annotator 4	0.801	0.774
Annotator 1	0.791	0.765
Annotator 5	0.785	0.757
Annotator 3	0.782	0.756
Syntax with Relevancy	0.696	0.677
Baseline 3: Machine Learning Based	0.689	0.669
Baseline 3: Machine Learning Based Reduced 7%	0.663	0.640
Relevancy-Driven	0.663	0.620
Baseline 1: Random Compression 77%	0.585	0.621
Syntax-Driven	0.583	0.556
Baseline 1: Random Compression 70%	0.516	0.507
Baseline 1: Random Compression 57%	0.385	0.465

Table 30: Content Evaluation of Compressed Summaries Against Human Annotations Using ROUGE F-Measure.

Baseline 1: Random Compression 77%. The last two techniques show significantly lower inter-annotator ROUGE-2 scores and they are: Baseline 1: Random Compression 70% and Baseline 1: Random Compression 57%.

Figure 14 shows similar results based on ROUGE-SU4 scores as well. But compared to ROUGE-2 results, the techniques: Syntax with Relevancy, Relevancy-Driven, Baseline 1: Random Compression 77%, Baseline 3: Machine Learning 1, Baseline 3: Machine Learning and Classifier based techniques all seem to be clustered together based on ROUGE-SU4. Table 30 shows a summary of Table 29 where the average inter-annotator ROUGE scores were calculated against all annotators. Here among all annotators, the highest ROUGE score was achieved by Annotator 2 (ROUGE-2:0.807 and ROUGE-SU4:0.780). Among all Annotators, the lowest ROUGE scores were achieved by Annotator 3 (ROUGE-2: 0.782 and ROUGE-SU4: 0.756). According to the methodology of ROUGE score [Lin, 2004], inter-annotator ROUGE scores show the level of agreement between the gold standard and automatic compressions. Hence in this case, we can predict the level of content agreement between human annotators and the automatic sentence compression techniques using the ROUGE scores we have obtained. Out of all the automatic sentence compression techniques, the keyword based technique has the highest average ROUGE scores (ROUGE-2: 0.813 and ROUGE-SU4: 0.785) and surprisingly, this was slightly higher than average human ROUGE scores as well. The next highest ROUGE scores of an automatic sentence compression technique was shown by the syntax with relevancy based pruning (ROUGE-2: 0.696 and ROUGE-SU4: 0.677). The Baseline 3: Machine Learning Based techniques scored the next highest ROUGE scores (ROUGE-2: 0.689 and ROUGE-SU4: 0.669). The techniques, relevancy-driven technique and the Baseline 3: Machine Learning and classifier based techniques (with a compression rate of 76.5%) scored similar ROUGE scores (ROUGE-2: 0.663 and ROUGE-SU4: 0.624) as the next highest ROUGE scores. Finally, the techniques Baseline 1: Random Compression 77%, syntax-driven, Baseline 1: Random Compression 70% and Baseline 1: Random Compression 57% scored the lowest ROUGE scores respectively.

We can point out interesting results we observed with inter-annotator content evaluation. First, other than Baseline 1: Random Compression 77%, all other random removals (compression rates with 70% and 57%) scored significantly lower inter-annotator ROUGE scores. This was as expected since the random removals fail to contain the notions of grammaticality or content relevancy. Out of all syntactic pruning techniques, the lowest score was obtained by the syntax-driven technique. However, the highest compression rate was also achieved by the syntax-driven technique (see Section 5.3.1). So we can conclude that the syntax-driven technique is harsh in removing syntactic structures compared to what humans do in sentence compression. The relevancy-driven technique has a relatively higher

ROUGE score compared to the syntax-driven technique and the highest ROUGE scores was achieved by the syntax with relevancy based pruning technique.
So with these results, we can conclude that syntax with relevancy based technique can approximate what humans do in sentence compressions better than the other two syntactic based sentence compression techniques. The Baseline 3: machine learning and classifier based techniques also approximate human compression better than the random removals yet it is slightly lower than syntax with relevancy based pruning technique. Though the keyword based techniques surprisingly scored the best ROUGE scores, it still failed to achieve a compression rate that is similar to all other sentence compression techniques.

6.7 Content Evaluation Based on Grammatical Relations

In the previous section, we have calculated inter-annotator content evaluation with the ROUGE metrics (R2 and SU4) using human summaries as the gold standard. This gave us an interesting comparison between all the techniques compared to human summaries. However, when the compression rate is very low, the scores do not seem to reflect the disagreement between human annotations and automatic techniques, especially with the keyword based technique. Another problem with the previous evaluation is that ROUGE scores are based on n-gram models (see Section 2.1). When these n-gram co-occurrences are calculated, the ROUGE metric does not take into account factors like the position of the n-grams in a sentence. For example, consider the following scenario:

> Gold standard compression: *"Despite skepticism about the actual realization of a single European currency as scheduled on January 1, 1999, preparations for the design of the Euro note have already begun.*
> *German Economics Minister Guenter Rexrodt said today that most German companies are not prepared for the shift to the single European currency, the Euro, which is due to be launched ~~on January 1, 1999~~."*
>
> Automatic compression: *Despite skepticism about the actual realization of a single European currency as scheduled ~~on January 1, 1999~~, preparations for the design of*

the Euro note have already begun.

German Economics Minister Guenter Rexrodt said today that most German companies are not prepared for the shift to the single European currency, the Euro, which is due to be launched on January 1, 1999.

In this example, both gold standard compression and automatic compression removed the prepositional phrase *"on January 1, 1999"* but in two different sentences. When ROUGE-1 and ROUGE-2 score were calculated for automatic compression against the gold standard compression, we have obtained following results.

Total bi-gram count for the gold standard summary: 58

Total bi-gram count for the extracted summary: 58

Total co-occurring bi-grams: 58

For ROUGE-1, the precision, recall and F-measures are calculated as follows.

$Precision = \frac{58}{58} = 1.000$

$Recall = \frac{58}{58} = 1.000$

$F - Measure = 1.000$

For ROUGE-2, the precision, recall and F-measures are calculated as follows.

Total bi-gram count for the gold standard summary: 57

Total bi-gram count for the extracted summary: 57

Total co-occurring bi-grams: 55

$Precision = \frac{55}{57} = 0.965$

$Recall = \frac{55}{57} = 0.965$

$F - Measure = 0.965$

From this example, we can see that even though the gold standard and the automatic compression show a disagreement in the sentences compression, the ROUGE score has failed to take this into account and evaluate correctly. So with this in mind, we decided to evaluate our syntactic pruning techniques compared to human annotations, based on a metric that takes grammatical relations into account. This metric was first introduced in [Riezler et al., 2003] for automatic summary evaluation with the argument of improving automatic evaluation techniques while taking semantic information into account. The authors argue that it is easy to enhance automatic summary evaluation when a dependency parser is available and counting co-occurrences of dependency grammar structures between the gold standard summaries and automatic summaries would be more effective than counting n-gram co-occurrences in evaluating summaries. This technique was used by [Clarke and Lapata, 2006] where the authors evaluated sentence compression using dependency grammar structure co-occurrences. Following them, [Filippova and Strube, 2008] have also used the same mechanism to evaluate their sentence compression techniques, comparing their results to the work of [Clarke and Lapata, 2006]. Since we used the Stanford Parser as our dependency grammar parser, we also evaluated our techniques compared to human annotations using this metric. The following sections will describe this evaluation technique and what parameters are calculated using this technique.

6.7.1 Evaluation Methodology

The dependency grammar relation based evaluation of [Riezler et al., 2003] depends on a dependency parser. The basic approach is to calculate the co-occurrence of grammar structures between gold standard sentences and automatically compressed sentences. These co-occurrence counts are then used to calculate Recall, Precision and F-measures to compare the results. As an example, consider the following sentence and its compressed forms, Sentence (64c1) and (64c2):

> (64) *The SPLC previously recorded a 20-percent increase in hate groups from 1996 to 1997.*

(64c1) *The SPLC ~~previously~~ recorded a 20-percent increase in hate groups ~~from 1996 to 1997.~~*

(64c2) *The SPLC ~~previously~~ recorded a ~~20-percent~~ increase in hate groups from 1996 to 1997.*

Table 31 shows the dependency structures present in the original sentence (Sentence (64)), Sentence (64c1), Sentence (64c2) and the co-occurring dependency structures between compressed sentences, Sentence (64c1) and (64c2). According to Table 31, there are 8 co-occurring dependency structures between the compressed Sentences (64c1) and (64c2). Sentence (64c1) has a total of 9 dependency structures and Sentence (64c2) has 12 dependency grammar structures. The authors who proposed this technique have elaborated following ratios as Recall, Precision and F-measure.

Sentences	**Sentence (64)**	**Sentence (64c1)**	**Sentence (64c2)**	**Co-occurrences**
Structures	det(SPLC, The)	det(SPLC, The)	det(SPLC, The)	det(SPLC, The)
	nsubj(recorded, SPLC)	nsubj(recorded, SPLC)	nsubj(recorded, SPLC)	nsubj(recorded, SPLC)
	advmod(recorded, previously)			
	root(ROOT, recorded)	root(ROOT, recorded)	root(ROOT, recorded)	root(ROOT, recorded)
	det(increase, a)	det(increase, a)	det(increase, a)	det(increase, a)
	amod(increase, 20-percent)	amod(increase, 20-percent)		
	dobj(recorded, increase)	dobj(recorded, increase)	dobj(recorded, increase)	dobj(recorded, increase)
	prep(recorded, in)	prep(recorded, in)	prep(recorded, in)	prep(recorded, in)
	nn(groups, hate)	nn(groups, hate)	nn(groups, hate)	nn(groups, hate)
	pobj(in, groups)	pobj(in, groups)	pobj(in, groups)	pobj(in, groups)
	prep(recorded, from)		prep(recorded, from)	
	num(1997, 1996)		num(1997, 1996)	
	dep(1997, to)		dep(1997, to)	
	pobj(from, 1997)		pobj(from, 1997)	
Total Structures	14	9	12	8

Table 31: Dependency Structures for Sentences (64), (64c1) and (64c2).

$$Precision \;= \frac{\sum Matching\ Grammar\ Structures}{\sum No.\ of\ System\ Compressed\ Grammar\ Structures}$$

$$Recall \;= \frac{\sum Matching\ Grammar\ Structures}{\sum No.\ of\ Gold\ Standard\ Compressed\ Grammar\ Structures}$$

$$F - Measure \ = \frac{(2 \ \times Precision \ \times Recall)}{(Precision \ + \ Recall)}$$

So according to these ratios, for our example above, when we consider the Sentence (64c1) as our gold standard sentence,

$$Precision \ [Sentence \ (64c2)] = \frac{8}{12} = 0.667$$

$$Recall \ [Sentence \ (64c2)] = \frac{8}{9} = 0.889$$

$$F - Measure \ [Sentence \ (64c2)] = \frac{2 \ \times 0.667 \ \times 0.889}{0.667 \ + \ 0.889} = 0.762$$

Using this F-measure, we have evaluated our different techniques and baseline techniques taking human annotations as our gold standard summaries. The following sections will describe the results we obtained.

6.7.2 Inter-Annotator Content Evaluation Using Dependency Grammar Structures

In comparison to the evaluation we performed in Section 6.6 using ROUGE, we first calculated the inter-annotator content evaluation using the dependency grammar structure based metric we described earlier. We have calculated the F-measure and the Table 32 shows the results we obtained. Similarly to the evaluation we performed in Section 6.6 using ROUGE, we calculated the average inter-annotator content evaluation using the dependency grammar structure based metric as well. Table 33 shows the F-measure calculated on all techniques over all five annotators. When compared with the ROUGE-2 inter-annotator content evaluation (see Section 6.6), the dependency grammar structure F-measure seem to show some interesting results. First it shows a better content evaluation result between all annotators compared to the results we obtained from ROUGE measures.

Techniques	**Ann. 1**	**Ann. 2**	**Ann. 3**	**Ann. 4**	**Ann. 5**
Syntax-Driven	0.682	0.671	0.643	0.660	0.666
Baseline 1: Random Compression 57%	0.283	0.284	0.274	0.278	0.271
Relevancy-Driven	0.714	0.726	0.695	0.692	0.701
Baseline 1: Random Compression 70%	0.406	0.408	0.395	0.401	0.390
Syntactic with Relevancy	0.773	0.778	0.742	0.751	0.750
Baseline 1: Random Compression 77%	0.512	0.507	0.492	0.497	0.488
Baseline 2: Keyword Based	0.772	0.770	0.728	0.733	0.737
Baseline 3: Machine Learning Based	0.740	0.738	0.712	0.714	0.709
Baseline 3: Machine Learning Based Reduced 7%	0.715	0.718	0.688	0.692	0.684
Annotator 1	*1.000*	0.840	0.808	0.827	0.802
Annotator 2	0.840	*1.000*	0.822	0.829	0.827
Annotator 3	0.808	0.840	*1.000*	0.799	0.785
Annotator 4	0.826	0.828	0.804	*1.000*	0.811
Annotator 5	0.801	0.828	0.785	0.812	*1.000*

Table 32: Inter-Annotator Content Evaluation of Compressed Summaries Using Dependency Structure Based F-Measure.

Also, we see that the baseline 1 technique, where we removed words randomly has been penalized by this measure as we expected. When words are randomly removed, it hurts the grammaticality and the content of the summaries. However, since ROUGE is only calculated based on bi-gram co-occurrences, it fails to penalize the baseline 1 compressions. Using this grammar-based metric, we see that all the automatic sentence pruning techniques have performed significantly better than all the random word removal baselines. Not only that, as we assumed with the keyword based technique, the dependency grammar metric has penalized this sentence compression by giving a low precision compared to the gold standard human summaries. So here, the keyword based technique does not seem to approximate human annotations as we saw with ROUGE inter-annotator content evaluation. Finally, out of all three syntactic pruning techniques, as we expected, the syntax with relevancy based pruning has scored the highest F-measure and the relevancy driven and syntax driven techniques follow in order. The baseline 3 compressions, created using machine learning techniques, have scored slightly lower compared to our syntax with relevancy based pruning technique. This is shown graphically in Figure 15.

Different Techniques	**F-Measure**
Annotator 2	0.829
Annotator 1	0.819
Annotator 4	0.817
Annotator 3	0.808
Annotator 5	0.806
Syntactic with Relevancy	0.759
Baseline 2: Keyword Based	0.748
Baseline 3: Machine Learning Based	0.722
Relevancy-Driven	0.706
Baseline 3: Machine Learning Based Reduced 7%	0.699
Syntax-Driven	0.664
Baseline 1: Random Compression 77%	0.499
Baseline 1: Random Compression 70%	0.400
Baseline 1: Random Compression 57%	0.278

Table 33: Content Evaluation of Compressed Summaries Against Human Annotations Using Dependency Structure Based F-Measure.

6.8 Chapter Summary

In this chapter, we have described the evaluation of our approaches compared with human compressed summaries. We have used the DUC 2007 summarization track for this task and a set of 25 summaries with a word limit of 250, created from the best performing

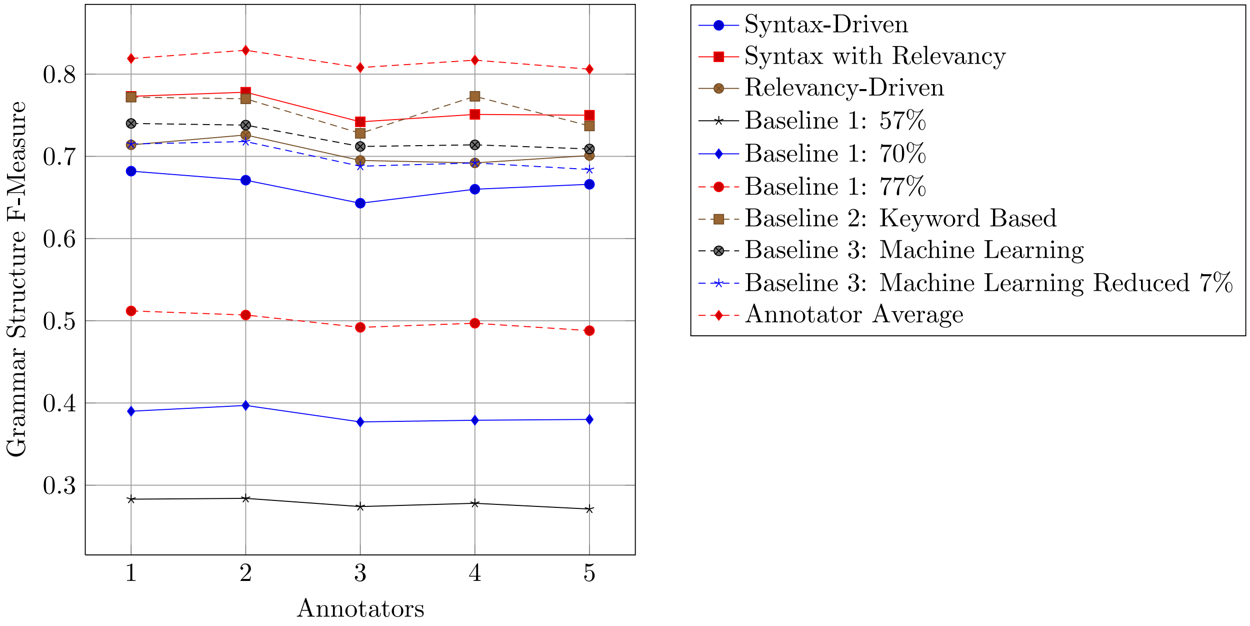

Figure 15: Inter-Annotator Content Evaluation Results of Dependency Structure based F-Measure.

system based on ROUGE scores in that particular track. We have used five sets of human annotations to evaluate our results. Human compressed summaries have obtained an average compression rate of 78.5%. This compression rate is similar to the compression rate of the syntax with relevancy based technique we have implemented. We have also evaluated compression rates for our baseline techniques: baseline 2, keyword based sentence compression and baseline 3, machine learning and classifier based technique. The baseline 3 technique has achieved a compression rate of 76-82% that is similar to human compression rate and our syntax with relevancy based technique. However, baseline 2, the keyword based compression technique achieved a lower compression (89.5%) compared to all the other techniques.

By analyzing the human compressed summaries, we have found that annotators tend to remove syntactic structures more than removing individual words. Also these syntactic structures are similar to the syntactic heuristics we have defined in our syntactic pruning techniques. However, annotators also tend to remove a small number of individual words and verb phrases as a consequence of removing syntactic structures to preserve the grammaticality of sentences. Also, another significant difference is that annotators remove specific types of noun phrases (proper, compound and temporal nouns). The removal of proper noun is rather subjective and rely on the background knowledge of individual annotators. In addition, we also noticed that human annotators removed a considerable number of prepositional phrases attached to verb phrases, something we have done cautiously in our syntax based techniques.

We have performed content evaluations using two metrics: ROUGE [Lin, 2004] and a dependency grammar structure based F-measure [Riezler et al., 2003]. The content evaluation using ROUGE showed that even human compressed summaries tend to lose content and the higher the compression rate is, the greater the decrease in content compared to the original summaries. The evaluations we performed using human compressed summaries as the gold standard showed that there is an overall agreement of 78% in content (with respect to ROUGE-2) between human annotators. The highest agreement in content with the human annotations was obtained by the syntax with relevancy based pruning technique (ROUGE-2: 0.696). For the baseline 1 technique, the inter-annotator ROUGE scores were significantly lower than our syntax with relevancy technique and also lower than baseline 2, the keyword based technique and baseline 3, machine learning and classifier based technique. However, we also saw that the keyword based technique scored significantly better than our techniques when evaluated against human annotations.

In our second series of content evaluation, we calculated a F-measure metric based on dependency grammar structures, introduced by [Riezler et al., 2003, Clarke and Lapata, 2006, Filippova and Strube, 2008] . The results we obtained were interesting as they showed that the grammar structure based metric could discriminate the loss of grammaticality of baseline 1 summaries. In addition, the keyword based technique also seems to show different content evaluation results, when compared with the content evaluation results obtained using ROUGE. The overall results showed that the highest F-measure was achieved by the human annotators with an F-measure of 0.81 and out of all automatic techniques, the syntax with relevancy based sentence compression technique showed the best result with an F-measure of 0.760. Baseline 2, the keyword based technique, achieved an F-measure of 0.750 and baseline 3, the machine learning and classifier based technique, obtained an F-measure of 0.722. These results show that our syntax with relevancy based pruning technique seem to achieve better results compared to existing sentence pruning techniques. Overall, the various evaluations have shown that our syntactic pruning heuristics approximate most human compression techniques with regards to syntactic structure removals. The syntactic structures which humans tend to remove are mostly similar to the syntactic structures we have chosen to remove by our syntactic based pruning techniques. However, surprisingly none of the compression techniques were able to increase ROUGE scores by content filling (see Chapter 5). In addition, these evaluations have highlighted the differences between what our heuristics remove compared to what human remove and gave us hints as to further refinement of our heuristics. For example, the removal of prepositional phrases attached to verb phrases, the removal of adjective phrases and different conjoined clauses removed by human annotators. In addition, other techniques such as classifier based pruning, could be applicable on top of these syntax based pruning techniques as a next step of improvement to approximate human compression techniques.

Chapter 7

Conclusion and Future work

Sentence compression and simplification is a challenging research area in Natural Language Processing. Having several applications, such as text simplification, headline generation and automatic text summarization, over the past decades, many research work has been done to explore different types of techniques to achieve this task. Previous work can be categorized as machine learning and classifier based techniques, keyword based approaches and syntactic based pruning approaches (see Chapter 4). In our work, we have experimented sentence compression based on syntactic pruning techniques to improve automatic text summarization. We have demonstrated the importance of preserving not only sentence grammaticality but also relevant information in the context of applying sentence compression to automatic text summarization. To perform compression while preserving the grammaticality of the sentences, we defined three syntactically driven pruning heuristics:

(1) Syntax-Driven Pruning.

(2) Syntax with Relevancy Based Pruning.

(3) Relevancy-Driven Pruning.

The first technique is based only on grammatical considerations but was too harsh since it assumed non-relevance of specific syntactic structures. To smooth this pruning technique, we explored a method that uses topic/query information given in the task of automatic summarization to detect relevant content and filter them out (see Section 4.2). To our knowledge, this methodology has not been used by any other previous technique and seems to be very effective in filtering out long sentence structures identified as relevant to the given topic/query. As our third approach, we have implemented a technique that is solely based on this relevancy score and filtered out embedding syntactic structures to achieve sentence compression.

As for our evaluation, we have performed a series of automatic evaluations based on compression rate and content. Our compression rates were similar to the previous work we have seen on sentence compression. The content evaluation we performed based on ROUGE measures did not seem to show an improvement in content as we expected even with filling. Because of this discouraging result in content evaluation, we performed an evaluation based on human compressed summaries.

For our human evaluation, we have asked five human annotators to perform sentence compression on a given set of summaries and we evaluated these summaries and compared the results with our techniques. Additionally, three baseline techniques were used to compare with our techniques. We obtained interesting results: We discovered that humans do tend to simplify sentences by removing syntactic structures and through our evaluations, we noted that out of the three techniques we developed, the syntax with relevancy based approach was much similar to human sentence compression techniques. In our evaluation, we calculated compression rate, content evaluation based on ROUGE and inter-annotator content evaluation based on ROUGE in addition to a dependency grammar based F-measure. The results of these evaluations showed that human approaches also caused the ROUGE scores to decrease compared to the original summaries. In addition, our baseline sentence compression techniques also reduce the ROUGE scores significantly. Inter-annotator content evaluation using ROUGE shows that there is an 80% agreement between all the annotators and for our sentence pruning techniques, the syntax-driven technique achieves 58%, the relevancy-driven technique achieves 66% and the syntax with relevancy based technique achieves 70%. For our baseline techniques, the keyword based technique achieves

80%, the machine learning and classifier based technique achieves 69% and all the random word removal techniques achieve 38%, 51% and 59% in descending order of their compression rates. Previous research work on sentence compression has used human judgment [Knight and Marcu, 2002, Le Nguyen et al., 2004, Galanis and Androutsopoulos, 2010] and grammar structure overlapping based F-measure, as mentioned in [Riezler et al., 2003, Clarke and Lapata, 2006, Filippova and Strube, 2008] for content evaluation. As we saw in section 6.7, ROUGE does not take phrase overlapping or the positions of words into account within a summary (see Section 6.7) but merely n-gram overlapping to evaluate content agreement. Since our results show higher inter-annotator content agreement between the keyword based approach even though it failed in achieving a compression rate similar to human compression rate and show a surprisingly higher ROUGE scores for the random word removal techniques (for random compression 77%), we decided to perform an evaluation based on grammar relations as proposed in [Riezler et al., 2003, Clarke and Lapata, 2006, Filippova and Strube, 2008].

The content evaluation based on grammar relations show that our techniques have a better content agreement with human annotations; syntax-driven with 66%, relevancy-driven with 70% and syntax with relevancy based technique with 76%. The average human score shows an agreement of over 81% and our baseline techniques, keyword based showed a 75%, machine learning and classifier based technique showed 72% and the random removals shows 28%, 40% and 50% respectively.

7.1 Main Findings of Our Work

Our work has highlighted several interesting findings. Through the evaluation of the BlogSum system, we have learned that sentence compression techniques are needed in order to improve automatic text summarization beyond extractive summarization. Most of the systems have focused on improving content selection techniques and relatively little work has been done on improving the quality of summaries based on readability and coherence. Our analysis of BlogSum's results showed that if a summary contains complex sentences, applying language generation techniques such as sentence aggregation and cue phrase insertion may actually reduce the readability of the texts. Hence, when proceeding beyond extractive summarization towards abstractive text summarization, sentence compression and simplification is a vital part in automatic text summarization (see Chapter 2).

In sentence compression, three main approaches have been used in earlier work: machine learning and classifier based technique, keyword based techniques and syntactic based pruning techniques (see Chapter 3). These techniques were developed and evaluated on different tasks such as text simplification and headline generation tasks and to our knowledge, very few papers have specifically addressed the evaluation of sentence compression extrinsically for automatic text summarization.

In our work, we have implemented three techniques based on syntactic sentence pruning and these approaches were influenced by the objective of preserving grammaticality and relevant content. Our automatic evaluation of sentence compression rates were similar to previous work on sentence compression but our content evaluation using ROUGE did not seem to show any improvement.

Our human evaluations gave us interesting results. First we discovered that humans tend to remove syntactic phrase structures more than individual words. This was evident with the analysis we performed using human compressions and by looking at the compression rates achieved by removing different types of words/phrase structures. Most of the syntactic structures that humans removed were similar to our syntactic pruning heuristics with the exceptions that humans also remove nouns and noun phrases based on their word knowledge. It was also interesting to note that human compressions also reduced ROUGE scores compared to the original ROUGE; hence letting us ponder about the inappropriateness of the ROUGE measure in evaluating sentence compression.

Out of our three techniques, syntax with relevancy based pruning technique was much similar to human compressions. Also our syntax with relevancy based technique had the highest inter-annotator evaluation results based on the dependency grammar structure overlapping F-measure and performed slightly higher than the machine learning and classifier based techniques and as well as keyword based technique.

When inter-annotator content evaluation was performed using ROUGE, we obtained some interesting results that showed how each technique approximate human compression. However, since the ROUGE measure failed to discriminate different techniques based on the low compression rate compared to human annotations (e.x. for keyword based approach) and the low grammaticality of compressed sentences (i.e. for random words/phrase structure removing techniques), we were again convinced that ROUGE may not be a good measurement to evaluate sentence compression. On the other hand, the grammar relation F-measure, introduced by [Riezler et al., 2003, Clarke and Lapata, 2006, Filippova and Strube, 2008] seems to take these factors into account and using this measure found it to be a better technique in evaluating sentence compression as opposed to the ROUGE metric.

7.2 Future Work

Many challenges are left to be investigated in sentence compression using syntactic pruning techniques to approximate what humans do in sentence compression. The following sections will briefly describe possible improvements and techniques to improve our sentence pruning techniques.

Exploration on Other Syntactic Structures So far, in our techniques, we have pruned six types of syntactic structures (relative clauses, adjective phrases, adverbial phrases, conjoined verb phrases, appositive and prepositional phrases). However, more specific syntactic structures can be considered for pruning as they provide secondary information to the main content in a sentence. Some of these include: gerunds, infinitive markers, interpolated clauses etc. However, most of these structures need to be pruned with caution similarly to the removal of prepositional phrases. Exploring these other structures and adding new syntactic pruning heuristics will be one interesting future work.

A Classification Based Syntactic Pruning In our syntactic pruning heuristics, we identified specific phrase structures to remove and the human evaluations demonstrated that these structures were similar to human sentence pruning as well. However, beyond that, we want to explore if these pruning heuristics can be fine-tuned using simple classifier approaches, especially for long structures such as relative clauses, conjoined clauses, appositive clauses and prepositional phrases. It would be interesting to identify which features humans look for before deciding whether to remove a long phrase structure or not. Some of the features we foresee at the surface level include the length of the phrase, the number of noun phrases enclosed, the head of the phrase (specially for prepositional phrases, the head preposition), the location of the structure within the sentence, enclosing punctuation characters etc. However for this work, we would require a specifically annotated corpus to create a training and testing set to implement our classifiers.

Semantic Topic/Query Expansion In our syntax with relevancy based and relevancy-driven techniques, we used the topic/query relevancy as a measure to decide which structures to prune. This relevancy score was calculated using the cosine similarity of $tf.idf$ values of the sub structures. As future work, we could experiment with this relevancy score by expanding the topic/query pairs within the context of given document cluster for summarization. This approach could be an expansion of queries based on a dictionary approach or some other technique, that gives more power in identifying relevant content or discriminating irrelevant content of structures.

An Integer Linear Programming Approach In our syntactic pruning approaches, we removed all the structures within one sentence if they satisfied our pruning heuristics and relevancy score based filter. However, instead of removing all such phrases, we can see this problem as an integer linear programming problem, similar to the work of [Clarke and Lapata, 2008]. Here, the goal would be to select a subset of syntactic structures to prune, bounded by parameters such as the total relevance score and desired compression rate. This may avoid removing all the structures when there are several structures to be pruned from a single sentence and may optimize our sentence pruning techniques.

Sentence Compression for Abstractive Summarization In abstractive summarization, a few recent work has focused on sentence clustering and generating abstracts out of multiple sentences. In particular, the work of [Ganesan et al., 2010] and [Filippova, 2010] have presented techniques where they create word graphs for multiple sentences, which were clustered based on similarities and generate an abstraction using these word graphs. It would be interesting to apply our sentence pruning techniques to simplify complex sentences as a part of these abstract summary generation techniques. We would be interested to see if the word graph based techniques can benefit from our sentence compression by minimizing the complexity of these graphs before generating abstractions.

Aggregation of Compressed Sentences Finally, recall from Chapter 2, the original goal of the thesis was to investigate sentence compression in order to see if the sentence aggregation heuristics of BlogSum would result in more natural summaries if complex sentences were simplified first. Our research has lead us into a deeper evaluation of sentence compression techniques, but it certainly would be interesting to perform a manual evaluation of the linguistic quality of BlogSum-aggregated sentences with and without our compression techniques.

Bibliography

[Baskervill and Sewell, 1986] Baskervill, W. M. and Sewell, J. W. (1986). *An English Grammar for the Use of High School, Academy, and College Classes*. American Book Company, New York, USA.

[Chandrasekar et al., 1996] Chandrasekar, R., Doran, C., and Srinivas, B. (1996). Motivations and Methods for Text Simplification. In *Proceedings of the* 16th *Conference on Computational Linguistics*, volume 2 of *COLING '96*, pages 1041–1044, Copenhagen, Denmark. Association for Computational Linguistics.

[Clarke and Lapata, 2006] Clarke, J. and Lapata, M. (2006). Models for Sentence Compression: A Comparison Across Domains, Training Requirements and Evaluation Measures. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 377–384, Sydney, Australia.

[Clarke and Lapata, 2008] Clarke, J. and Lapata, M. (2008). Global Inference for Sentence Compression an Integer Linear Programming Approach. *Journal of Artificial Intelligence Research (JAIR)*, 31(1):399–429.

[Collins, 2003] Collins, M. (2003). Head-Driven Statistical Models for Natural Language Parsing. *Computational Linguistics*, 29(4):589–637.

[Conroy et al., 2006] Conroy, J. M., Schlesinger, J. D., O'Leary, D. P., and Goldstein, J. (2006). Back to Basics: CLASSY 2006. In *Proceedings of the HLT-NAACL 2006 Document Understanding Workshop*, New York City.

[Conroy et al., 2005] Conroy, J. M., Stewart, J. G., and Schlesinger, J. D. (2005). CLASSY Query-Based Multi-Document Summarization. In *Proceedings of the Document Understanding Conference Workshop 2005 (DUC 2005)*, Vancouver, British Columbia, Canada. Human Language Technology Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP).

[Copeck et al., 2007] Copeck, T., Inkpen, D., Kazantseva, A., Kennedy, A., Kipp, D., and Szpakowicz, S. (2007). Catch What You Can. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2007)*, Rochester, New York, USA.

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

[Dang, 2005] Dang, H. T. (2005). Overview of DUC 2005. In *Proceedings of the Document Understanding Conference Workshop 2005 (DUC 2005)*, Vancouver, British Columbia, Canada. Human Language Technology Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP).

[Dietrich, 2007] Dietrich, H. (2007). *Relative Clauses with Relative Pronouns*. GRIN Verlag, Munich, Germany.

[Dorr et al., 2003] Dorr, B., Zajic, D., and Schwartz, R. (2003). Hedge Trimmer: A Parse-and-Trim Approach to Headline Generation. In *Proceedings of the HLT-NAACL 03 on Text Summarization Workshop*, volume 5, pages 1–8, Edmonton, Canada.

[Dorr et al., 2005] Dorr, B. J., Monz, C., President, S., Schwartz, R., and Zajic, D. (2005). A Methodology for Extrinsic Evaluation of Text Summarization: Does ROUGE Correlate? In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 1–8, University of Michigan, USA.

[Dowty, 2000] Dowty, D. (2000). The Dual Analysis of Adjuncts/Complements in Categorial Grammar. pages 33–66. Walter de Gruyter, Berlin, Germany.

[Dunlavy et al., 2003] Dunlavy, D. M., Conroy, J. M., Schlesinger, J. D., Goodman, S. A., Okurowski, M. E., O'Leary, D. P., and van Halteren, H. (2003). Performance of a Three-Stage System for Multi-Document Summarization. In *Proceedings of the HLT-NAACL 2003 Document Understanding Workshop*, pages 153–159, Edmonton, Canada.

[Erkan and Radev, 2004] Erkan, G. and Radev, D. R. (2004). LexRank: Graph-Based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research (JAIR)*, 22(1):457–479.

[Filippova, 2010] Filippova, K. (2010). Multi-Sentence Compression: Finding Shortest Paths in Word Graphs. In *Proceedings of the* 23^{rd} *International Conference on Computational Linguistics*, COLING '10, pages 322–330, Beijing, China.

[Filippova and Strube, 2008] Filippova, K. and Strube, M. (2008). Dependency Tree Based Sentence Compression. In *Proceedings of the* 5^{th} *International Natural Language Generation Conference*, INLG '08, pages 25–32, Salt Fork, Ohio, USA.

[Gagnon and Da Sylva, 2006] Gagnon, M. and Da Sylva, L. (2006). Text Compression by Syntactic Pruning. In Ali, M. and Dapoigny, R., editors, *Proceedings of the* 19^{th} *International Conference on Advances in Artificial Intelligence: Canadian Society for Computational Studies of Intelligence*, volume 19, pages 312–323. Quebec City, Canada.

[Galanis and Androutsopoulos, 2010] Galanis, D. and Androutsopoulos, I. (2010). An Extractive Supervised Two-Stage Method for Sentence Compression. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 885–893, Los Angeles, California.

[Ganesan and Zhai, 2012] Ganesan, K. and Zhai, C. (2012). Opinion-Based Entity Ranking. In Clarke, C., Zobel, J., and Mothe, J., editors, *Information Retrieval*, volume 15, pages 116–150. Springer Netherlands.

[Ganesan et al., 2010] Ganesan, K., Zhai, C., and Han, J. (2010). Opinosis: A Graph Based Approach to Abstractive Summarization of Highly Redundant Opinions. In *Proceedings of the* 23^{rd} *International Conference on Computational Linguistics*, COLING '10, pages 340–348, Beijing, China.

[Graff, 1995] Graff, D. (1995). *North American News Text Corpus.* Linguistic Data Consortium, Philadelphia, USA.

[Graff, 2002] Graff, D. (2002). *The AQUAINT Corpus of English News Text.* Linguistic Data Consortium, Philadelphia, USA.

[Graff et al., 1996] Graff, D., Garofolo, J., Fiscus, J., Fisher, W., and Pallett, D. (1996). *English Broadcast News Speech (HUB4).* Linguistic Data Consortium, Philadelphia, USA.

[Gupta and Lehal, 2010] Gupta, V. and Lehal, G. (2010). A Survey of Text Summarization Extractive Techniques. In Mohammed, S., Al-Rousan, M., Li, W., and Al-Dubai, A., editors, *Journal of Emerging Technologies in Web Intelligence*, volume 2, pages 258–268. Academy Publisher Inc. British Virgin Islands, UK.

[Hahn and Harman, 2002] Hahn, U. and Harman, D. (2002). Document Understanding Conference (DUC), Workshop on Text Summarization. In Harman, D. and Hahn, U., editors, *Proceedings of ACL 2002 Workshop on Text Summarization*, DUC 02, Philadelphia, Pennsylvania, USA. In Conjunction with the Association for Computational Linguistic (ACL) 2002.

[Harman, 1992] Harman, D. (1992). Overview of the First Text REtrieval Conference (TREC-1). In *Special Publication 500-207, Online Publication*, TREC 92, pages 1–20, Gaithersburg, Maryland.

[Harman, 2001] Harman, D. (2001). Document Understanding Conference (DUC), Workshop on Text Summarization. DUC 01, New Orleans, Louisiana USA. In Conjunction with the ACM SIGIR Conference 2001.

[Harman and Liberman, 1993] Harman, D. and Liberman, M. (1993). *TIPSTER Complete. Linguistic Data Consortium (LDC)*. Philadelphia.

[Hirst, 1981] Hirst, G. (1981). *Anaphora in Natural Language Understanding: A Survey*, volume 119 of *Lecture Notes in Computer Science*. Springer-Verlag.

[Hovy and Lin, 1998] Hovy, E. and Lin, C.-Y. (1998). Automated Text Summarization and the SUMMARIST System. In *Proceedings of a workshop on held at Baltimore, Maryland: October 13-15, 1998*, TIPSTER '98, pages 197–214, Baltimore, Maryland.

[Jaoua et al., 2012] Jaoua, M., Jaoua, F., Belguith, L. H., and Hamadou, A. B. (2012). Évaluation de l'impact de l'intégration des étapes de filtrage et de compression dans le processus d'automatisation du résumé. In *Résumé automatique de documents*, Volume 15 of Document numérique, pages 67–90.

[Jing, 2000] Jing, H. (2000). Sentence Reduction for Automatic Text Summarization. In *Proceedings of the* 6^{th} *Conference on Applied Natural Language Processing*, ANLC '00, pages 310–315. Association for Computational Linguistics, Seattle, USA.

[Jing and McKeown, 2000] Jing, H. and McKeown, K. R. (2000). Cut and Paste Based Text Summarization. In *Proceedings of the* 1^{st} *North American Chapter of the Association for Computational Linguistics Conference*, Proceedings of NAACL-2000, pages 178–185, Seattle, USA.

[Knight and Marcu, 2002] Knight, K. and Marcu, D. (2002). Summarization Beyond Sentence Extraction: A Probabilistic Approach to Sentence Compression. In *Artificial Intelligence*, volume 139, pages 91–107. Elsevier Science Publishers Ltd., Essex, UK.

[Le Nguyen et al., 2004] Le Nguyen, M., Shimazu, A., Horiguchi, S., Ho, B. T., and Fukushi, M. (2004). Probabilistic Sentence Reduction Using Support Vector Machines. In *Proceedings of the* 20^{th} *International Conference on Computational Linguistics*, Proceedings of COLING'04, pages 743–749, Geneva, Switzerland.

[Lin, 2004] Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In Moens, M.-F. and Szpakowicz, S., editors, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain.

[Liu et al., 2005] Liu, B., Hu, M., and Cheng, J. (2005). Opinion Observer: Analyzing and Comparing Opinions on the Web. In *Proceedings of the* 14^{th} *international conference on World Wide Web*, WWW '05, pages 342–351, Chiba, Japan.

[Luhn, 1958] Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

[Macdonald and Ounis, 2006] Macdonald, C. and Ounis, I. (2006). The TREC Blogs06 Collection: Creating and Analysing a Blog Test Collection. Technical report. Technical Report No.: TR-2006-224.

[Mani et al., 1999] Mani, I., House, D., Klein, G., Hirschman, L., Firmin, T., and Sundheim, B. (1999). The TIPSTER SUMMAC Text Summarization Evaluation. In *Proceedings of the* 9^{th} *conference on European chapter of the Association for Computational Linguistics*, EACL '99, pages 77–85, Bergen, Norway.

[Mani and Maybury, 2001] Mani, I. and Maybury, M. T. (2001). *Automatic Summarization*. John Benjamins Publishing Company, Amsterdam, Netherlands.

[Marcus et al., 1993] Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

[Marneffe and Manning, 2008] Marneffe, M.-C. D. and Manning, C. D. (2008). The Stanford Typed Dependencies Representation. In *Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, CrossParser '08, pages 1–8, Manchester, United Kingdom.

[McClosky et al., 2006] McClosky, D., Charniak, E., and Johnson, M. (2006). Effective Self-Training for Parsing. In *Proceedings of the Main Conference on Human Language*

Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings of HLT-NAACL'06, pages 152–159, New York, USA.

[Merlo and Ferrer, 2006] Merlo, P. and Ferrer, E. E. (2006). The Notion of Argument in Prepositional Phrase Attachment. *Computational Linguistics*, 32(3):341–378.

[Miller, 1995] Miller, G. (1995). WordNet: A Lexical Database for English. volume 38, pages 39–41. ACM, New York, USA.

[Mithun, 2010] Mithun, S. (2010). Exploiting Rhetorical Relations in Blog Summarization. In Farzindar, A. and Kešelj, V., editors, *Advances in Artificial Intelligence*, volume 6085 of *Lecture Notes in Computer Science*, pages 388–392.

[Mithun et al., 2012] Mithun, S., Kosseim, L., and Perera, P. (2012). Discrepancy Between Automatic and Manual Evaluation of Summaries. In *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, pages 44–52, Montréal, Canada.

[Murray et al., 2008] Murray, G., Joty, S., and Ng, R. (2008). The University of British Columbia at TAC 2008. In *Proceedings of the Text Analysis Conference*, Proceedings of TAC 2008, Gaithersburg, Maryland USA. The National Institute of Standards and Technology (NIST).

[Nguyen and Leveling, 2013] Nguyen, D. and Leveling, J. (2013). Exploring Domain-Sensitive Features for Extractive Summarization in the Medical Domain. In Métais, E., Meziane, F., Saraee, M., Sugumaran, V., and Vadera, S., editors, *Natural Language Processing and Information Systems*, volume 7934 of *Lecture Notes in Computer Science*, pages 90–101.

[Nguyen et al., 2003] Nguyen, M. L., Phan, X. H., Horiguchi, S., and Shimazu, A. (2003). A New Sentence Reduction Technique Based on a Decision Tree Model. In *Proceedings of the* 17^{th} *Pacific Asia Conference on Language, Information and Computation*, pages 290–297. Singapore.

[Otterbacher et al., 2003] Otterbacher, J., Qi, H., Hakim, A., and Radev, D. R. (2003). The University of Michigan at TREC 2003. In *TREC*, pages 732–738.

[Owczarzak et al., 2012] Owczarzak, K., Conroy, J. M., Dang, H. T., and Nenkova, A. (2012). An Assessment of the Accuracy of Automatic Evaluation in Summarization. In *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, pages 1–9, Montreal, Canada.

[Paice and Husk, 1987] Paice, C. and Husk, G. (1987). Towards the Automatic Recognition of Anaphoric Features in English Text: The Impersonal Pronoun “it”. *Computer Speech Language*, 2(2):109 – 132.

[Perera and Kosseim, 2013] Perera, P. and Kosseim, L. (2013). Evaluating Syntactic Sentence Compression for Text Summarisation. In Mtais, E., Meziane, F., Saraee, M., Sugumaran, V., and Vadera, S., editors, *Natural Language Processing and Information Systems*, volume 7934 of *Lecture Notes in Computer Science*, pages 126–139. Springer Berlin Heidelberg.

[Pingali et al., 2007] Pingali, P., K, R., and Varma, V. (2007). IIIT Hyderabad at DUC 2007. In *Proceedings of the HLT-NAACL 2007 Document Understanding Workshop*, Rochester, New York.

[Porter, 1997] Porter, M. F. (1997). An algorithm for suffix stripping. pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[Radev et al., 2004] Radev, D., Allison, T., Blair-Goldensohn, S., Blitzer, J., Çelebi, A., Dimitrov, S., Drabek, E., Hakim, A., Lam, W., Liu, D., Otterbacher, J., Qi, H., Saggion, H., Teufel, S., Topper, M., Winkel, A., and Zhang, Z. (2004). MEAD - A Platform for Multidocument Multilingual Text Summarization. In *Proceedings of LREC-2004*, Lisbon, Portugal.

[Radev et al., 2002] Radev, D. R., Hovy, E., and McKeown, K. (2002). Introduction to the Special Issue on Summarization. *Computational Linguistics*, 28(4):399–408.

[Radev and McKeown, 1998] Radev, D. R. and McKeown, K. R. (1998). Generating Natural Language Summaries from Multiple On-line Sources. *Computational Linguistics*, 24(3):470–500.

[Riezler et al., 2003] Riezler, S., King, T. H., Crouch, R., and Zaenen, A. (2003). Statistical Sentence Condensation Using Ambiguity Packing and Stochastic Disambiguation Methods for Lexical-Functional Grammar. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 118–125, Edmonton, Canada.

[Taylor et al., 2003] Taylor, A., Marcus, M., and Santorini, B. (2003). The Penn Treebank: An Overview. In Abeill, A., editor, *Treebanks*, volume 20 of *Text, Speech and Language Technology*, pages 5–22. Springer Netherlands.

[Štajner et al., 2012] Štajner, S., Evans, R., Orasan, C., and Mitkov, R. (2012). What Can Readability Measures Really Tell Us About Text Complexity? In *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*, Istanbul, Turkey.

[Wan and Yang, 2007] Wan, X. and Yang, J. (2007). Single Document Summarization With Document Expansion. In *Proceedings of the* 22^{nd} *national conference on Artificial intelligence*, volume 1 of *AAAI'07*, pages 931–936, Vancouver, British Columbia, Canada.

[Zajic et al., 2007] Zajic, D., Dorr, B. J., Lin, J., and Schwartz, R. (2007). Multi-Candidate Reduction: Sentence Compression as a Tool for Document Summarization Tasks. In *Information Processing and Management*, volume 43, pages 1549–1570. Pergamon Press, Inc., Tarrytown, New York, USA.

Appendix A

Penn Treebank Tagset

Part-of-speech tags are assigned to a single word according to its role in the sentence [Marcus et al., 1993, Taylor et al., 2003].

Tag	**Description**	**Example**
CC	conjunction, coordinating	*and, or, but*
CD	cardinal number	*five, three, 13%*
DT	determiner	*the, a, these*
EX	existential there	*there were six boys*
FW	foreign word	*mais*
IN	conjunction, subordinating or preposition	*of, on, before, unless*
JJ	adjective	*nice, easy*
JJR	adjective, comparative	*nicer, easier*
JJS	adjective, superlative	*nicest, easiest*
LS	list item marker	
MD	verb, modal auxillary	*may, should*
NN	noun, singular or mass	*tiger, chair, laughter*
NNS	noun, plural	*tigers, chairs, insects*
NNP	noun, proper singular	*Germany, God, Alice*
NNPS	noun, proper plural	*we met two Christmases ago*

PDT	predeterminer	*both his children*
PRP	pronoun, personal	*me, you, it*
PRP$	pronoun, possessive	*my, your, our*
RB	adverb	*extremely, loudly, hard*
RBR	adverb, comparative	*better*
RBS	adverb, superlative	*best*
RP	adverb, particle	*about, off, up*
SYM	symbol	%
TO	infinitival to	*what to do?*
UH	interjection	*oh, oops, gosh*
VB	verb, base form	*think*
VBZ	verb, 3rd person singular present	*she thinks*
VBP	verb, non-3rd person singular present	*I think*
VBD	verb, past tense	*they thought*
VBN	verb, past participle	*a sunken ship*
VBG	verb, gerund or present participle	*thinking is fun*
WDT	wh-determiner	*which, whatever, whichever*
WP	wh-pronoun, personal	*what, who, whom*
WP$	wh-pronoun, possessive	*whose, whosever*
WRB	wh-adverb	*where, when*
.	punctuation mark, sentence closer	.;?*
,	punctuation mark, comma	,
:	punctuation mark, colon	:
(	contextual separator, left paren	(
)	contextual separator, right paren	)

Chunk Tags Chunk tags are assigned to groups of words that belong together (i.e. phrases). The most common phrases are the noun phrase (NP, for example *"the black cat"*) and the verb phrase (VP, for example *"is purring"* [Taylor et al., 2003]).

Tag	**Description**	**Example**
NP	noun phrase	*the strange bird*
PP	prepositional phrase	*in between*
VP	verb phrase	*was looking*
ADVP	adverb phrase	*also*
ADJP	adjective phrase	*warm and cosy*
SBAR	subordinating conjunction	*whether or not*
PRT	particle	*up the stairs*
INTJ	interjection	*hello*

Appendix B

Keyword List

This keyword list is used in the sentence compression technique described as keyword based sentence compression (see Section 6.2.2).

as well as	besides	coupled with
in addition	likewise	moreover
accordingly	as a result	consequently
for this purpose	hence	otherwise
subsequently	therefore	thus
wherefore	by the same token	conversely
on one hand	on the other hand	on the contrary
still	nevertheless	in contrast
with attention to	particularly	singularly
barring	excluding	exclusive of
for example	for instance	for one thing
illustrated with	as an example	in this case
all in all	all things considered	briefly
in any case	in any event	in brief
on the whole	in short	in summary

in the long run	on balance	to sum up
finally	in the first place	as a matter of fact
not to mention	correspondingly	at the same time
then again	in reality	although
notwithstanding	in the event that	for the purpose of
with this in mind	in the hope that	for fear that
in view of	provided that	given that
in other words	to put it differently	to put it another way
by all means	important to realize	another key point
most compelling evidence	point often overlooked	to point out
notably	including	to be sure
chiefly	truly	certainly
markedly	in fact	in general
in detail	to demonstrate	to emphasize
to clarify	to explain	to enumerate
specifically	expressively	surprisingly
significantly	under those circumstances	in that case
as can be seen	generally speaking	as shown above
as has been noted	in a word	for the most part
altogether	overall	ordinarily
in either case	at the present time	from time to time
up to the present time	to begin with	in due time
as soon as	in the meantime	in a moment
all of a sudden	at this instant	immediately
straightaway	occasionally	in the middle
on this side	in the distance	here and there
in the background	in the center of	adjacent to

Appendix C

Sample Compressed Summaries

Original Summary

Topic: *Starbucks Coffee*

Query: *How has Starbucks Coffee attempted to expand and diversify through joint ventures, acquisitions, or subsidiaries?*

Pasqua, which Starbucks plans to convert to the Starbucks banner, operates 48 stores in the San Francisco area, Los Angeles and New York, and has eight licensed airport locations in California. Breyer writes for the Austin American-Statesman. announced a new consumer products division for its specialty ice cream, bottled coffee drinks and other products sold away from its own shops. The new division will take over business that already brought in $206 million last year, or almost 16 percent of the company's total sales. Most of it has been joint ventures: the ice cream with Dreyer's Grand Ice Cream; the bottled Frappuccino with Pepsi-Cola; and the supermarket distribution of Starbucks coffee with Kraft Foods Inc. Schultz's Seattle-based Starbucks Coffee Co. and Johnson's Los Angeles-based Johnson Development Corp. have recently opened similar Starbucks coffee stores in West Los Angeles and New York's Harlem. In addition, Williams-Sonoma's larger stores would be good locales for small Starbucks coffee bars. Company executives had expected that strong retail sales in Starbucks core coffee business would make up the difference, but that didn't happen. The chairman and chief executive of Starbucks, Howard Schultz, the country did not need more fancy coffee shops, promised to do the same with Internet strategy, with legions of customers like Siess following Starbucks into cyberspace. By David Lazarus STARBUCKS Coffee giant Starbucks Corp. has purchased Hear Music, an upscale music retailer that until recently was based in San Francisco, for under $10 million. Under the deal, Kozmo will pay

Annotator 5

Topic: *Starbucks Coffee*

Query: *How has Starbucks Coffee attempted to expand and diversify through joint ventures, acquisitions, or subsidiaries?*

Pasqua, which Starbucks plans to convert to the Starbucks banner, operates 48 stores ~~in the San Francisco area, Los Angeles and New York,~~ and has eight licensed airport locations in California. Breyer writes for the Austin American-Statesman. announced a new consumer products division for its ~~specialty~~ ice cream, ~~bottled~~ coffee drinks and other products ~~sold away from its own shops~~. The new division will take over business that already brought in $206 million last year, or almost 16 percent of the company's total sales. Most of it has been joint ventures: the ice cream with Dreyer's Grand Ice Cream; the bottled Frappuccino with Pepsi-Cola; and the supermarket distribution of Starbucks coffee with Kraft Foods Inc. Schultz's Seattle-based Starbucks Coffee Co. and Johnson's Los Angeles-based Johnson Development Corp. have ~~recently~~ opened similar Starbucks coffee stores ~~in West Los Angeles and New York's Harlem~~. ~~In addition,~~ Williams-Sonoma's larger stores would be good locales for small Starbucks coffee bars. Company executives had expected that ~~strong~~ retail sales in Starbucks core coffee business would make up the difference, but that didn't happen. The chairman and chief executive of Starbucks~~, Howard Schultz, the country did not need more fancy coffee shops,~~ promised to do the same with Internet strategy, with legions of customers like Siess following Starbucks into cyberspace. By David Lazarus STARBUCKS Coffee giant Starbucks Corp. has purchased Hear Music~~, an upscale music retailer that until recently was based in San Francisco,~~ for under $10 million. Under the deal, Kozmo will pay

Syntax-Driven Pruning

Topic: *Starbucks Coffee*

Query: *How has Starbucks Coffee attempted to expand and diversify through joint ventures, acquisitions, or subsidiaries?*

Pasqua, ~~which Starbucks plans to convert to the Starbucks banner,~~ operates 48 stores in the San Francisco area, Los Angeles and New York, ~~and has eight licensed airport locations in California~~. Breyer writes for the Austin American-Statesman. announced a ~~new~~ consumer products division ~~for its specialty ice cream, bottled~~ coffee drinks and other products sold away from its ~~own~~ shops. The ~~new~~ division will take over business that ~~already~~ brought in $206 million ~~last~~ year, or almost 16 percent ~~of the company's total sales~~. Most ~~of it~~ has been joint ventures: the ice cream ~~with Dreyer's Grand Ice Cream~~; the ~~bottled~~ Frappuccino ~~with Pepsi-Cola~~; and the supermarket distribution ~~of Starbucks coffee~~ with Kraft Foods Inc. Schultz's ~~Seattle-based~~ Starbucks Coffee Co. and Johnson's Los Angeles-based Johnson Development Corp. have ~~recently~~ opened ~~similar~~ Starbucks coffee stores in West Los Angeles and New York's Harlem. ~~In addition~~, Williams-Sonoma's ~~larger~~ stores would be ~~good~~ locales ~~for small Starbucks coffee bars~~. Company executives had expected that ~~strong retail~~ sales ~~in Starbucks core coffee business~~ would make up the difference, but that didn't happen. The chairman and chief executive~~of Starbucks~~, Howard Schultz, ~~the country did not need more fancy coffee shops~~, promised to do the same with Internet strategy, with legions ~~of customers like Siess following Starbucks into cyberspace~~. ~~By David Lazarus STARBUCKS Coffee giant~~ Starbucks Corp. has purchased Hear Music, an ~~upscale~~ music retailer that ~~until recently~~ was based in San Francisco, for under $10 million. ~~Under the deal,~~ Kozmo will pay

Syntax With Relevancy Based Pruning

Topic: *Starbucks Coffee*

Query: *How has Starbucks Coffee attempted to expand and diversify through joint ventures, acquisitions, or subsidiaries?*

Pasqua, which Starbucks plans to convert to the Starbucks banner, operates 48 stores in the San Francisco area, Los Angeles and New York, ~~and has eight licensed airport locations in California~~. Breyer writes for the Austin American-Statesman. announced a ~~new~~ consumer products division ~~for its specialty ice cream, bottled~~ coffee drinks and ~~other~~ products sold away from its ~~own~~ shops. The ~~new~~ division will take over business that ~~already~~ brought in $206 million ~~last~~ year, or almost 16 percent ~~of the company's total sales~~. Most ~~of it~~ has been joint ventures: the ice cream ~~with Dreyer's Grand Ice Cream~~; the ~~bottled~~ Frappuccino ~~with Pepsi-Cola~~; and the supermarket distribution of Starbucks coffee with Kraft Foods Inc. Schultz's ~~Seattle-based~~ Starbucks Coffee Co. and Johnson's Los Angeles-based Johnson Development Corp. have ~~recently~~ opened ~~similar~~ Starbucks coffee stores in West Los Angeles and New York's Harlem. ~~In addition,~~ Williams-Sonoma's ~~larger~~ stores would be ~~good~~ locales for ~~small~~ Starbucks coffee bars. Company executives had expected that ~~strong retail~~ sales in Starbucks ~~core~~ coffee business would make up the difference, but that didn't happen. The chairman and chief executive of Starbucks, Howard Schultz, the country did not need more ~~fancy~~ coffee shops, promised to do the same with Internet strategy, with legions of customers like Siess following Starbucks into cyberspace. By David Lazarus STARBUCKS Coffee giant Starbucks Corp. has purchased Hear Music, an ~~upscale~~ music retailer that ~~until recently~~ was based in San Francisco, for under $10 million. ~~Under the deal,~~ Kozmo will pay

Baseline 1: Random Compression 77%

Topic: *Starbucks Coffee*

Query: *How has Starbucks Coffee attempted to expand and diversify through joint ventures, acquisitions, or subsidiaries?*

~~Pasqua~~, which Starbucks ~~plans~~ to ~~convert~~ to the Starbucks banner, operates 48 stores ~~in~~ the San Francisco area, Los Angeles and New York, and has eight ~~licensed~~ airport locations in California. Breyer writes for the Austin American-Statesman. announced a new a consumer products division ~~for~~ its ~~specialty~~ ice cream, bottled coffee drinks and ~~other~~ products sold away ~~from~~ its own shops. The ~~new~~ division ~~will~~ take over ~~business that already~~ brought ~~in~~ $206 million last year, ~~or almost~~ 16 ~~percent~~ of the company's ~~total~~ sales. Most of ~~it~~ has ~~been~~ joint ventures: the ice cream with Dreyer's ~~Grand~~ Ice Cream; ~~the bottled Frappuccino~~ with Pepsi-Cola; and the supermarket distribution of Starbucks coffee with Kraft Foods Inc. Schultz's Seattle-based ~~Starbucks Coffee~~ Co. and Johnson's Los Angeles-based Johnson ~~Development~~ Corp. have ~~recently opened~~ similar Starbucks coffee stores in West Los Angeles and New ~~York~~'s Harlem. In addition, Williams-Sonoma's larger stores would ~~be good~~ locales for small Starbucks coffee bars. Company executives had ~~expected~~ that strong retail sales in ~~Starbucks~~ core ~~coffee~~ business would ~~make~~ up the difference, but that did~~n't~~ happen. ~~The~~ chairman and ~~chief~~ executive ~~of~~ Starbucks, ~~Howard~~ Schultz, the country did not need ~~more fancy coffee~~ shops, promised to do ~~the~~ same with Internet strategy, with legions of customers like Siess following ~~Starbucks into~~ cyberspace. By ~~David~~ Lazarus STARBUCKS Coffee giant Starbucks ~~Corp.~~ has purchased ~~Hear Music~~, an upscale music retailer ~~that~~ until recently was based in San Francisco, for under $10 million. ~~Under~~ the deal, Kozmo will pay

Baseline 2: Keyword Based

Topic: *Starbucks Coffee*

Query: *How has Starbucks Coffee attempted to expand and diversify through joint ventures, acquisitions, or subsidiaries?*

Pasqua, ~~which Starbucks plans to convert to the Starbucks banner~~, operates 48 stores in the San Francisco area, Los Angeles and New York, and has eight licensed airport locations in California. Breyer writes for the Austin American-Statesman. announced a new consumer products division for its specialty ice cream, bottled coffee drinks and other products sold away from its own shops. The new division will take over business ~~that already brought in $206 million last year, or almost 16 percent of the company's total sales~~. Most of it has been joint ventures: the ice cream with Dreyer's Grand Ice Cream; the bottled Frappuccino with Pepsi-Cola; and the supermarket distribution of Starbucks coffee with Kraft Foods Inc. Schultz's Seattle-based Starbucks Coffee Co. and Johnson's Los Angeles-based Johnson Development Corp. have recently opened similar Starbucks coffee stores in West Los Angeles and New York's Harlem. ~~In addition,~~ Williams-Sonoma's larger stores would be good locales for small Starbucks coffee bars. Company executives had expected that strong retail sales in Starbucks core coffee business would make up the difference, but ~~that didn't happen~~. The chairman and chief executive of Starbucks, Howard Schultz, the country did not need more fancy coffee shops, promised to do the same with Internet strategy, with legions of customers like Siess following Starbucks into cyberspace. By David Lazarus STARBUCKS Coffee giant Starbucks Corp. has purchased Hear Music, an upscale music retailer ~~that until recently was based in San Francisco, for under $10 million~~. Under the deal, Kozmo will pay

Baseline 3: Machine Learning Based

Topic: *Starbucks Coffee*

Query: *How has Starbucks Coffee attempted to expand and diversify through joint ventures, acquisitions, or subsidiaries?*

Pasqua, ~~which~~ Starbucks plans to convert to the Starbucks banner, operates 48 stores in the San Francisco area, ~~Los Angeles~~ and New York, and has~~eight~~ licensed ~~airport~~ locations in California. Breyer writes for the Austin American-Statesman. ~~announced a new consumer products division for its specialty ice cream, bottled coffee drinks and other products sold away from its own shops.~~ The new division will take over business that ~~already~~ brought in $206 million ~~last year~~, or ~~almost~~ 16 percent of ~~the company's~~ total sales. Most of it has been joint ventures~~: the ice cream with Dreyer's Grand Ice Cream; the bottled Frappuccino with Pepsi-Cola;~~ and the supermarket distribution of Starbucks coffee with ~~Kraft~~ Foods Inc. Schultz's ~~Seattle-based~~ Starbucks Coffee Co. and ~~Johnson's~~ Los Angeles-based Johnson Development Corp. have ~~recently~~ opened ~~similar~~ Starbucks ~~coffee~~ stores in West Los Angeles and New York's Harlem. ~~In addition~~, Williams-Sonoma~~'s larger~~ stores would be good locales for small ~~Starbucks~~ coffee bars. Company executives had expected that ~~strong~~ retail sales in Starbucks ~~core~~ coffee business would make up the difference~~, but that didn't happen~~. The chairman and chief executive ~~of Starbucks, Howard Schultz,~~ the country did not need ~~more fancy~~ coffee shops, promised to do the same with ~~Internet strategy, with~~ legions of customers like Siess following Starbucks into cyberspace. By David ~~Lazarus~~ STARBUCKS Coffee giant Starbucks Corp. has purchased Hear Music, an upscale ~~music~~ retailer that until recently was based in San Francisco~~, for under $10 million~~. ~~Under the deal,~~ Kozmo will pay