Multi-document Summarization System Using Rhetorical Information

by

Mohammed Alliheedi

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2012

AUTHOR'S DECLARATION

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Over the past 20 years, research in automated text summarization has grown significantly in the field of natural language processing. The massive availability of scientific and technical information on the Internet, including journals, conferences, and news articles has attracted the interest of various groups of researchers working in text summarization. These researchers include linguistics, biologists, database researchers, and information retrieval experts. However, because the information available on the web is ever expanding, reading the sheer volume of information is a significant challenge. To deal with this volume of information, users need appropriate summaries to help them more efficiently manage their information needs. Although many automated text summarization systems have been proposed in the past twenty years, none of these systems have incorporated the use of rhetoric. To date, most automated text summarization systems have relied only on statistical approaches. These approaches do not take into account other features of language such as antimetabole and epanalepsis. Our hypothesis is that rhetoric can provide this type of additional information. This thesis addresses these issues by investigating the role of rhetorical figuration in detecting the salient information in texts. We show that automated multi-document summarization can be improved using metrics based on rhetorical figuration. A corpus of presidential speeches, which is for different U.S. presidents speeches, has been created. It includes campaign, state of union, and inaugural speeches to test our proposed multi-document summarization system. Various evaluation metrics have been used to test and compare the performance of the produced summaries of both our proposed system and other system. Our proposed multi-document summarization system using rhetorical figures improves the produced summaries, and achieves better performance over MEAD system in most of the cases especially in antimetabole, polyptoton, and isocolon. Overall, the results of our system are promising and leads to future progress on this research.

Acknowledgements

.

I would like to thank my supervisor, Prof. Chrysanne DiMarco, for all her support throughout my graduate studies and for all her guidance and encouragement.

I would like also to express my thanks to my supervisor, Prof. Charlie Clarke, for all his guidance and insightful comments that helped me to finish this thesis.

I would like to thank Prof. Randy Harris for his recommendations in rhetorical figures and the study procedure which were insightful and beneficial and for being my thesis reviewer.

I would like to thank Prof. Mark Sumaker for his comments and being my thesis reviewer.

I would like also to tank the participants who take part in this study.

Dedication

To my family

especially my parents

and

my wife

Table of Contents

List of Figures

List of Tables

List of Equations

Chapter 1

Introduction

1.1 What is Summarization?

Radev et al. [37] define a summary as “text that is produced from one or more texts that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that.” This definition identifies three key factors for meaningful summaries: (1) the summary can be produced from one or more documents; (2) the summary must be less than the half of the original text; and (3) the summary should contain important information. Moreover, Mani [26] defines the task of summarization as follows: “to take an information source, extract content from it, and present the most important content to the user in a condensed form and in a manner sensitive to the user’s or application’s need.”

1.2 What is Rhetorical Figuration?

Corbett [10] defines rhetorical figures as “an artful deviation." It is artfully deviating an expression from its original manner. Essentially, rhetorical figures are categorized into several types. The simplest type of categorizations divides them into two groups, figures of speech and figures of thought [5]. While a figure of speech deals with verbal expression, the latter type concerns the idea.

Schemes and tropes are other categorizations of rhetorical figures. Basically, schemes are defined as "deviations from the ordinary arrangement of words" [5]. However, tropes use a word in a manner that differentiates from its usual meaning [5]. Rhetorical figures are also grouped based on pattern of usage. For example, figures of omission elide words, phrases, or clauses, as in “I came, I saw, I conquered.” This example of asyndeton omits the conjunctions between clauses. Figures of repetition are based on patterns which repeat words, phrases, clauses or ideas, as in “Eat more vegetables. Eat less junk food. Eat small meals more often.” This example of epanaphora repeats the first in successive sentences.

1.3 Research Statement

Automated multi-document summarization is a method of summarization that aims to create a condensed version of information, which is concise yet comprehensive, from several texts written on the same topic. As the result, a summary is produced that will allow users to understand the content of a large cluster of documents. Although many statistical approaches have been proposed for multi-document summarization, these approaches have demonstrated insufficient quality in the summaries produced compared to human-based summarization. The main reason for this limitation is that these approaches calculate the significance of texts based only on relevance to the topic without using other language features. From the study of rhetoric, we can get a better understanding of the effective use of language in written texts. Rhetoric involves other aspects of language, which includes semantic content, emotional information, and author intention. Therefore, it is our contention that multi-document summarization using rhetorical figuration can improve the summarization of texts because

the system will now calculate the importance of a text based both on its relevance to the topic and aspects of semantic content.

1.4 Methodology

Our proposed multi-document summarization system is divided into two components: (1) an annotator of rhetorical figures and (2) the multi-document summarizer itself. In order to annotate rhetorical figures in texts, we use Java Annotation Tool Of Rhetoric (JANTOR) [14] to atomically annotate rhetorical figures in a text. Essentially, we aim to annotate various figures of repetition such as polyptoton and epanalepsis. Several experiments will be performed on different corpora of presidential speeches including campaign, state of union and inaugural speeches to evaluate the quality of the summaries produced. First, the automated annotator of rhetorical figures will be performed on each document to detect rhetorical figures. Each figure in the document will be assigned a score. This action will be repeated until each figure in the document receives a score. These documents are stored in a cluster. Then, the cluster will be entered into the summarizer, known as the MEAD system [38]. The MEAD system calculates the score of each sentence in the cluster depending on several features, which includes centroid, length, and position of the sentence in a document. Therefore, sentence selection will be based on the finial score of each sentence, which is the sum of rhetorical figure score and the MEAD features score. As the result, a summary produced from a cluster will contain the most salient sentences.

1.5 Organization of the Thesis

Chapter 2 reviews the related work on single document summarization and multi-document summarization. Chapter 3 shows a brief overview of rhetorical figurers and defines the various types of figures. Chapter 3 also presents the methodology of our approach towards multi-document summarization system in detail. It also describes both the MEAD system and the JANTOR tool. Chapter 4 shows the results of our approach and the evaluation metrics. Lastly, Chapter 5 is the conclusion and future work.

Chapter 2

Related Work

Searching for relevant information in written texts is a challenging task. It requires scanning and interpreting a very large volume of written material on many different topics, and thus is a challenge for readers to obtain pertinent information easily and conveniently. In addition, searching manually through texts is extremely time consuming. The goal of automated text summarization is to facilitate the task of manually searching for information and to assist users to rapidly acquire the desired information in a condensed form, quickly, accurately, and efficiently. For the past two decades, researchers have experimented with various approaches to automated text summarization to improve the capabilities of the summarization systems in selecting the most salient parts of a text. In this chapter, the main types of summaries will be first defined. Then, an overview of representative systems for both single and multi-document summarization will be presented.

2.1 Types of Summary

Text summarization in general is categorized by three stages in the process of summarization: input, output, and purpose [18].

At the input stage, the user needs to identify three characteristics: (1) the source size, which is either a single-document or multi-document; (2) specificity, which requires the input to be either dependent or independent of a certain domain; and (3) input language, which requires the system to select either monolingual or multilingual processing.

At the output stage, the user needs to determine whether the system should produce either an *abstract* or *extract*. An *abstractive summarization* method determines the main themes of the source text, then generates a representative summary that is a paraphrasing of the original text. In contrast, an *extractive summarization* method produces a condensed version of the original text that contains the most salient actual sentences of the original. This method of summarization extracts the n highest ranked sentences from the original text without necessarily requiring deep understanding of the original text. An extractive summarizer consists of two stages: pre-processing stage and processing stage [16]. In the pre-processing stage, several procedures are involved: (1) the boundaries of sentences in the source text are identified; (2) *stopwords* are deleted which do not contribute to the meaning of the text (*stopwords* are frequent words such as “the,” “can,” and “which"); (3) the root of each word is obtained by applying a stemming method.

In the subsequent processing stage, the importance of every sentence in the original text is determined, and then each sentence is assigned a score based on various metrics (e.g., the frequency of each word in the sentence).

At the purpose stage, the system either accepts a submitted query from the user to retrieve specific information in the text, or merely provides a general summary that is essentially the author’s perspective of the original text.

2.2 Single Document Summarization

The important information in a text is often distributed throughout the different parts of the text. That is, some sentences contain more important information than others. Therefore, the key task in summarization is to identify the key sentences in a text using an applicable method. In this section, popular extractive approaches in single document summarization are described. Early examples of summarization systems based on surface level features will first be presented. Then, several representative summarization systems using deep-language analysis will briefly be described. Lastly, summarization systems using Machine Learning methods will be reviewed.

2.2.1 Surface Level

2.2.1.1 Word Frequency Feature

Luhn [25] proposed how summarizers use various features for single-document summarization based on the distribution of words in the source text. Luhn stated that the frequency of a certain word in a document reflects its relevance. Hence, words with the highest frequencies represent significant concepts in the document.

Subsequent other researchers (e.g. Edmundson [12], Kupiec, Pedrson, and Chen [19], and Hovy and Lin [17], Strzalkowski et al. [43] proposed various frequency measures.

In Edmundson's system, the results showed that 44% of sentences in the original text were co-selected in automated summaries and human-based summaries. Edmundson demonstrated that incorporating different features would yield good performance.

Kupiec et al.'s system results showed that 79% of original-text sentences occurred in both automated and human-based summaries. Kupiec et al. demonstrated that using a combination of other metrics could result in insufficient performance.

Hovy and Lin defined the notion of a “concept signature” to identify topic words in a text. Hovy and Lin claimed that each text conveys certain concept signature(s); thus, each concept signature represents a pair of topic words along with a list of related keywords. Their metric was incorporated with other metrics in the Topic Identification Step of the SUMMARIST system [17].

Hovey and Lin also applied concept features in the Topic Interpretation Step. Their idea was to group related words in the text within certain concepts. Thus, this feature counts concepts instead of words. For example, consider the sentences “Mike bought milk, bread, and meat. Hence, Mike bought some groceries.” In this example, the term “groceries” is the general concept of specific words such as “meat”, “milk”, and “bread”.

Strzalkowski et al. proposed a new term frequency feature in addition to the original one proposed earlier by [25], called term paragraph frequency. Strzalkowski et al. in [43] made the assumption that for each word occurring in only specific paragraphs of a text, the word contributes to the concept of the text more than a word that is distributed over the entire text. Therefore, this word is weighted higher than other words that distribute over the text as a whole.

2.2.1.2 Sentence Position Feature

Baxendale [2] proposed a particularly helpful feature for an automatic text summarization system that identifies important parts of text by basically identifying the position of a topic sentence in a document. In short, this feature assumes that sentences occurring in the initial parts of the document

or individual paragraphs tend to be most relevant. In order to confirm this feature, Baxendale analyzed 200 paragraphs to determine the position of the topic sentences. The results indicate that 85% of the paragraphs introduced the topic sentence in the first sentence. However, the topic sentence also appeared as the last sentence in 7% of the paragraphs. While this feature is simplistic, it provides a reasonably accurate method to judge the importance of the position of a topic sentence.

Position feature has been used in various systems such as the systems of Edmundson [12], Brandow, Mitze, and Rau [3], Kupiec et al. [19], and Lin and Hovy [23].

Edmundson proposed a text summarization system that used four methods to extract the most salient sentences of a text. Among these methods, two methods were related to position feature, Title and Location. The Title method was based on the hypothesis that words in titles and headings tend to be relevant to the concept of the text. The Location method was concerned with the location of a sentence. Sentences that occur in the beginning or the last part of a text tend to be topic sentences, and sentences that occur under specific headings are also related to the main topic of the text.

Singer and Dolan [42], in contrast, demonstrated that the topic sentence of a paragraph can appear in different locations in the paragraph, or may not be declared explicitly at all.

Brandow et al. in [3] proposed a domain-independent system (ANES) that performed automatic extraction from news articles. This system essentially employed a position feature by modifying the weight of signature words, words with a high (tf*idf) weight, in sentences. Here tf refers to term frequency while idf refers to inverse document frequency. Thus, the system extracted the sentences with high tf*idf scores.

Lin and Hovy, however, showed the first systematic study for position features. They developed a modified method using position feature, Optimal Position Policy (OPP). For a given collection of texts with abstracts, OPP produces a list of the sentence positions in a text that cover the main topic.

2.2.1.3 Cue Phrase Feature

Edmundson [12] proposed a text summarization system that used four features to extract the most salient sentences of a text. The first feature relied on linguistic information known as the cue method. This method used pragmatic cues such as "hardly," "significant," and "impossible," to identify the important sentences. Since the cue method is a method that depends on the specific genre of text, this method requires a pre-existing cue dictionary containing the characteristic words for a specific text genre. Edmundson described three groups of words in the cue dictionary that could be used to detect the important sentences in the text: *bonus words* that are positively relevant; *stigma words* that are negatively relevant; and *null words* that are irrelevant. In order to evaluate the extracted summaries from the system, a set of 400 technical documents were manually extracted. The experimental results indicated that using a combination of the three methods (cue, title, and location) produced the best correlation between the manual and automated summaries. In particular, the results demonstrated that 44% of sentences were co-selected in both the automated extracts and the manual ones.

Paice [34], Kupiec, Pedersen, and Chen [19], Teufel and Moens [45], and Teufel and Moens [46] also proposed automated text summarization systems that used cue-phrase features. Paice observed that cue phrase features can precisely computed. This type of feature also produces better summarizer performance than a word frequency feature.

2.2.2 Mid-level Method (Lexical Chains)

In the following section, we describe several automatic text summarization systems which are based on mid-level methods, more sophisticated than simple cues, but not requiring deep natural language analysis.

Barzilay and Elhadad [1] proposed using lexical chains as a representation of the source text to create a condensed summary. Essentially, a lexical chain is defined as a sequence of related words in a text without distance restriction [31]. In order to proceed with summarization, four steps are involved: segmentation of the source text; constructing lexical chains for each segment; identifying the strong lexical chains; and extracting the important sentences in the source text. Although co-reference and collocations are other forms of cohesion, which might enrich the quality of the produced summary, Barzilay and Elhadad focused on only one form of cohesion, lexical chaining. In order to identify lexical chains in the source text, Barzilay and Elhadad used WordNet as the knowledge base, a lexical database for the English language [30]. A three-step procedure was applied to construct lexical chains. First, a set of candidate words must be identified in the text. Second, an appropriate chain needs to be selected that relies on a semantic relationship among all words in the chain. Third, if such a chain is found, the chain is updated after inserting the candidate word in the chain. Since the same word can have various interpretations, Barzilay and Elhadad defined a component for each word that included its interpretations. In order to find the most appropriate interpretation of each word in a lexical chain that best conveys the chain sense, each word interpretation is represented as a node. If two words in the chain have a common sense, then an edge is formed between these words. Therefore, the best interpretation is the one with the most edges between members in the chain. Barzilay and Elhadad defined a score function based on two parameters: length, which is the number of occurrences of words in the chain; and homogeneity

index, which is the number of distinct occurrences divided by length then subtracted from the equation 1. Barzilay and Elhadad used a number of heuristics to extract significant sentences from the source text. To evaluate the produced summary, Barzilay and Elhadad compared their system with Microsoft's summarizer. A set of 40 documents was summarized using both systems. Recall and precision measures were used for similarity. The experimental results indicated that lexical chain summarization system outperforms the Microsoft summarizer.

Silber and McCoy [41], Nahnsen, Uzuner, and Katz [32], and Verma, Chen, and Lu [47], and others also developed automated text summarization systems using lexical chains.

Silber and McCoy extended Barzilay and Elhadad's system. They developed an efficient algorithm that was linear in both time and space to identify lexical chains in written materials. As an enhancement to Barzilay and Elhadad's method, Silber and McCoy created "metachains", which represent all possible lexical chains in a text, as a text representation by finding the relationship between words in the text. Sense disambiguation was applied to determine the relevant sense of a word among different senses. Basically, the word sense that has many relationships with other words in the text is selected; the other word senses are removed. Once all words in the text are disambiguated, the final lexical chain will have been constructed.

Recent work using ontologies in summarization was proposed by Verma et al. [47]. The authors describe a system that generated dynamic summaries based on a user query. The system used two different ontology databases, WordNet and UMLS (a large medical ontology). The following steps were executed to produce a document summary. The submitted user query is first assessed and adjusted in regards to the available ontology databases, which are both WordNet and UMLS, by deleting redundant keywords and adding relevant ones. Next, the distance between the document sentence and the revised query is calculated. Lastly, the distance among all candidate summary

sentences is calculated and the highest ranked sentences chosen to form the summary. The system was presented at the Document Understanding Conference (DUC) in 2007, and it showed a number of limitations, including insufficient redundancy reduction and a lack of syntactic analysis.

2.2.3 Machine Learning Methods

With the introduction of Machine Learning methods in Natural Language Processing, researchers in automated text summarization began to employ various Machine Learning methods, such as Naïve-Bayes, Neural Networks, and Hidden Markov Models. Several representative summarization systems using Machine Learning methods and various statistical techniques are described below. First, Naïve-Bayes methods will be presented. Then, a single-document summarization system using Neural Networks will be described. Lastly, a system using a Hidden Markov Model approach will be discussed.

2.2.3.1 Naïve-Bayes Method

Kupiec et al. [19] describe a single-document summarization system using Naïve-Bayes method. The authors proposed two new statistical features, sentence length and upper-case words, in addition to the features described by [12]. In order to evaluate the importance of text sentences, a Naïve-Bayes classifier was applied to compute the probability of each sentence. Let F1,F2,….Fn represent the features, s be a sentence in the source text, and S be a set of summary sentences. By assuming statistical independence of the features, the probability can be calculated as follow:

$$P(s \in S \mid F_i, F_2, ..F_k) = \frac{\prod_{i=1}^{k} P(F_i \mid s \in S) \bullet P(s \in S)}{\prod_{i=1}^{k} P(F_i)}$$

Equation 2.1: Naïve Bayes Classifier

The classification function determines whether the sentence will be included in the summary by assigning a score for each sentence based on Equation 1. The top-ranked sentences are extracted to perform the summary. To evaluate the produced summary, a training set of 188 technical document / manual summary pairs was used. The authors manually evaluated the sentences in the manual summaries and those in the source texts by linking them in several ways, including direct sentence match, a join of sentences match, and unmatchable sentence. Their results showed that 44% of sentences were co-selected in both the automated extracts and the manual ones.

2.2.3.2 Neural Network

The NetSum system was developed by Svore et al. [44] and is based on the RankNet algorithm [4], a neural network algorithm. NetSum extracts the three sentences from a single document that best capture the document's highlights. NetSum essentially incorporated the RankNet algorithm to rank document sentences. In order to identify salient sentences in the document, two sets of news articles from CNN.com were used to form the training set and test set. The training set was used to extract a set of features from each sentence and to train the system so that NetSum learns from the training set the distribution of features among the salient sentences. The results are a list of the highest ranked sentences for every single document. When NetSum was run on the test set, RankNet inferred the appropriate rank for sentences in a document based on the information gathered from the training set

about sentence features. In order to evaluate the system, Svore et al. used ROUGE [22], an evaluation system for text summarizers, to compare the performance of NetSum to a baseline system, SumBasic [33], which outperformed all previous systems for news article summarization and was the standard baseline system in the 2001 Document Understating Conference (DUC). NetSum outperformed the baseline and demonstrated a significant improvement in performance at 95% confidence compared to the baseline's over a dataset of 1365 documents.

2.2.3.3 Hidden Markov Models

Conroy and O'Leary [9] proposed an automated summarization system using a Hidden Markov Model (HMM) to extract the important sentences from a document. HMM-based summarization systems differ from those using Naïve-Bayes classifiers [19], which are based on independent features. A HMM summarizer uses a sequential approach that requires dependencies between sentence features. Three types of features are used in the HMM process for automated text summarization: 1) position of the sentence in a document (built into the state structure of the HMM); 2) number of terms in a sentence; 3) likelihood of sentences given the document terms. Conroy and O'Leary described their model as consisting of 2s + 1 states, where s stands for summary state and s+1 for non-summary state. The HMM model allows a "hesitation" process for non-summary states and "skipping next state" for summary states. A set of 1304 documents gathered from a TREC data set was divided into two parts, one for the training corpus and the other for the evaluation of the system. For each document, a manual summary was generated. Conroy and O'Leary compared the automated summaries that were produced from their system to the manual summaries using the following metric:

$$F_1 = 100\frac{2r}{k_h + k_m}$$

Equation 2.2: F-measure for summaries overlap

In Equation 2.2, K_h is the length of a human-based summary, K_m is the length of the system summary, and r is the number of sentences that both summaries have in common. In the evaluation, the HMM summaries co-selected at least 51 of sentences in the human-based summaries.

2.3 Multi-Document Summarization

Although single-document summarization provides the user with a short summary of a document's content, this approach provides only limited and specific information from one source. However, multi-document summarization produces a comprehensive summary from multiple sources. Since the single-document summarization track was dropped from the Document Understanding Conference (DUC) Challenge in 2003, research in this area has been declining. There are now a number of different approaches for multi-document summarization. In the following section, several multi-document summarization systems are reviewed.

2.3.1 Maximal marginal relevance (MMR)

MMR is a technique developed by Carbonell and Goldstein [7]. MMR was introduced to enrich topic-driven summarization by combining measurements for both query relevance and information novelty of a specific topic. In other words, the MMR aims to score sentences based on their redundancy among candidate sentences and their relevance to a submitted query. If a document is relevant to a query and includes certain similarities to prior selected documents, it is considered to possess high marginal relevance. The MMR for a document is calculated as follows:

$$MMR \overset{def}{=} Arg \max_{D_i \in \setminus S} \left[\lambda (Sim_1(D_i, Q) - (1 - \lambda) \max_{D_j \in S} Sim_2(D_i, D_j)) \right]$$

Equation 2.3: MMR Formula

and the standard cosine similarity is calculated as:

$$Sim_1(x, y) = Sim_2(x, y) = \frac{\langle x, y \rangle}{\|x\| \cdot \|y\|}$$

Equation 2.4: Cosine Similarity

Q is a query and R is a ranked list of retrieved documents. S is the set of previously selected documents in a particular step during the search for relevant documents, and $R \setminus S$ is the set of still

unselected documents in R. D_i is a candidate document $\in R \setminus S$ and λ is a parameter ranging between the value of (0,1). To produce the summary, documents are first segmented into sentences. Given a query, the highest-ranking sentences from a set of documents are chosen based on the relevance of a document to the given query. Thus, the MMR will reordered the sentences based on their ranks to form the resulting summary.

2.3.2 Abstractive Summarization System

McKeown and Radev [28] developed a multi-document summarization system (SUMMON), which was a knowledge-based system that produced a summary for a given cluster of news articles. Essentially, the architecture of SUMMON consisted of two components: a content planner that selected information derived from the templates of input texts to include in a text, and a linguistic component that used English grammar and a dictionary to select and arrange certain words, which referred to concepts in the text. The SUMMON system first extract the main semantic themes from the input texts in order to complete semantic templates. Thus, the system creates a summary from these templates. The input to SUMMON was a set of semantic templates that had been previously extracted from a message understanding system. McKeown and Radev used several corpora to evaluate their system, including news articles from Reuters, the Associated Press, and the Wall Street Journal. When the domain was narrow, the results of the summarization were good. However, the summaries were not as good for wide domains.

2.3.3 Centroid-based Summarization

The MEAD system was developed by Radev et al. [38]. MEAD s is a multi-document summarizer that uses cluster centroids to produce summaries. Initially, the system identifies all articles about an emerging event using modified tf*idf to group articles on the same topic in the same cluster, i.e., each cluster includes several articles that have the same topic. The grouping of articles into clusters based on the article's tf*idf: an article is included in a cluster if its tf*idf is close to the centroid of the cluster. The centroid of a cluster is a group of terms that statistically represent the cluster. Centroid-based summarization (CBS) is a new method for multi-document summarization, which uses the input of the centroids of clusters that were produced in the first step by CIDR. CIDR is the author's own extension to the more basic process of Topic Detection and Tracking (TDT).

Centroid-based summarization (CBS) works by selecting sentences that are central to the topic of the whole cluster. Two metrics are incorporated into the evaluation of single-document and multi-document summaries: (1) cluster-based relative utility (CBRU) concerns the degree of relevance of a sentence to the topic of the cluster; and (2) cross-sentence information subsumption (CSIS) refers to certain sentences duplicating the information in other sentences. In this case, these sentences will be omitted.

Three different features are used to score a sentence: (1) centroid value, the sum of centroids of all words in a sentence; (2) positional value; and (3) the first-sentence overlap with other sentences. A small corpus was used to evaluate MEAD experiments, including six clusters consisting of 27 articles in total. Radev et al. compared the performance between their MEAD and Lead-based summarizers, where Lead-based selected the first sentence from each cluster until the desired length of a summary

was achieved. The results showed that MEAD outperformed the Lead-based summarizer in 29 out of 45 cases.

2.3.4 Multi-lingual Summarization Systems

Evans [13] proposed a multi-lingual summarization system that addressed the task of summarizing articles in different languages. Such summaries are helpful to track different news articles written in multiple languages. The system extracted summaries in English while input documents were written in Arabic and English. The system aimed to summarize the English articles based on the information in both the English and Arabic articles. The Arabic documents were translated to English using IBM's statistical machine translation system. The system attempted to find similarities between sentences from the translated documents and the original ones. When the system found a relevant sentence from the translated documents that contained information similar to a sentence in the English documents, then the latter original English sentence was included in the summary instead of the translated one.

NeATS, developed by Lin and Hovy [24], was another multi-document summarizer. NeATS used three stages to produce a summary: content selection, filtering, and presentation. The idea of content selection was to determine the salient concepts in a set of documents. In order to identify the important sentences, NeATS used the likelihood-ratio λ to recognize concepts for unigram, bigram, and trigram models. The *relevant document set* was the on-topic documents, while the *irrelevant document set* was the off-topic documents. Then, concepts were clustered to group specific subtopics within the main topic. In the subsequent filtering stage, NeATS used three different filtering features:

(1) sentence position, which retained the leading sentences in texts; (2) stigma words such as conjunctions; and (3) Maximum Marginal Relevancy, which is a redundancy checker. Results showed that NeATS performed better in long summaries (200 to 400 words), and achieved the second best system in content evaluations in the DUC 2002.

2.3.5 A Cue-based Hub-Authority Approach

Zhang et al. [49] proposed a multi-document summarization system called Hub/Authority Framework. The system used two basic steps. First, the system clustered all sentences in texts having similar content, and then identified the sub-topics in the set of documents. In doing so, the system extracted the feature words for the various sub-topics. Secondly, two types of vertices in the graph were identified: Hub and Authority. The Hub vertex uses feature words (phrases), and the Authority vertex uses all sentences in the set of documents. An edge will be linked the two vertices if a sentence has one or more words from the Hub. Therefore, a relevant Authority sentence is the one that is pointed to by many relevant Hub words, and vice versa. The system also used a Hidden Markov Model to order the sub-topics that must be contained in the summary. In the evaluation, this system showed promising performance, and exceeded the two other systems, which are Random Policy and Position Policy.

2.3.6 Progressive Summarization System

Bysani [6] proposed new features to detect novelty in the context of progressive summarization. Progressive summarization assumes that the reader has previously read some articles about a certain topic. As a consequence, this type of summary is particularly helpful for tracking a product's reviews

and news stories. Bysani's system consisted of several stages, with each stage incorporating certain features to detect the novelty of a word. At the pre-processing stage, each article was cleansed of HTML tags and news headings, then the article was split into sentences. At the scoring stage, two features, Novelty Factor (NF) and New Word (NW), were used to relate the novelty of a sentence with its relevance. At the ranking stage, feature scores were combined to obtain the final ranking of a sentence. At the summary extraction stage, candidate sentences based on the final rankings were selected to form the summary. The system outperformed most of top systems participating in the 2009 Text Analysis Conference (TAC) .

Chapter 3

Methodology

In this chapter, we describe our general approach towards multi-document summarization using rhetorical figuration metrics. First, we present an overview of the two components of our multi-document summarization system. Then, we describe how we categorized the texts in our corpus and how we sampled the texts to build training and test sets for our following experiments. Then, we provide an overview of our methodology to compare the performance of an existing multi-document summarization system with the addition of rhetorical figuration metrics.

3.1 Summarization Framework

We hypothesize that using rhetorical figuration metrics should improve the performance of a text summarization system by producing more accurate and adequate summaries. Our proposed multi-document summarization system is divided into two components: (1) an annotator of rhetorical figures; and (2) the basic multi-document summarizer itself. In order to annotate rhetorical figures in texts, many different rhetorical figures could be used to enhance the capabilities of the summarizer. We chose to focus on several types of figures of repetition, such as ploche and polyptoton because

these repetition figures in particular provide emphasis, clarity, and emotional effect. We will discuss these figures in more detail in Section 3.1.1.

3.1.1 Detection of Rhetorical Figures

Before we describe how rhetorical figuration can be used in text summarization, we will first give an overview of rhetorical figures, together with recent work in computational rhetoric that we will be applying. In this section, we give a brief overview of classical rhetoric, then explain the characteristics of rhetorical figures and their properties that can be used to provide additional metrics for our summarization system. Lastly, we review the automated annotator of rhetorical figures, "JANTOR".

3.1.1.1 Classical Rhetoric

To better explain what a rhetorical figure is, the term "rhetoric" first needs to be defined. The term "rhetoric" is derived from the Greek rhētorikós, which means "an oratorical" [39]. Essentially, rhetoric can be defined as "the art of discourse", i.e., the art that uses language for the purpose of persuading, motivating, or changing the behaviour and attitudes of audiences [10]. Aristotle defined rhetoric as "the faculty of observing in any given case the available means of persuasion" [39]. In the ancient era, rhetoric had an essential role in political and social life, especially in Western tradition [8]. Rhetoric consists of five pillars: inventio, dispositio, elocutio, memoria, and pronuntiatio [10]. Inventio means discovery or invention, a method that was used to discover arguments in Western rhetoric. It can be defined as "the concept of a process that engages a rhetor (speaker or writer) in

examining alternatives: different ways to begin writing and to explore writing situations, and different ways of framing and verifying these judgments." [20].

Dispositio is the second pillar of rhetoric and was used to organize or arrange arguments in classical rhetoric. Dispositio means organization or arrangement. Cicero and Quintilian [10] stated that dispositio consists of six parts: exordium, the Introduction; narratio concerning the statement of a case; divisio, an outline of the salient parts in an argument; confirmatio, confirming the proof of the case; refutatio, concerning the refutation of opposing arguments; and peroratio, the conclusion.

Elocutio refers to the stylistic component in written or spoken Western discourses. Certain aspects of elocutio are used in crafting and delivering speeches or written texts, such as word choice and level of style, which includes plain, middle, and high style. Each level of style is intended for specific purpose [48]. Qunitilian [48] stated that a middle style is meant for moving oration; however, plain style is used for instruction and high style for charming discourse.

Memoria means memory, and involves memorizing classical rhetoric. Although writers have paid less attention to memoria than other parts of rhetoric, they still need to recall spoken or written materials in order to answer questions, and refute opposing arguments [14].

Pronuntiatio is the last pillar of rhetoric and concerns the manner of delivery of a written or spoken discourse. Although most rhetoricians agreed that pronuntiatio is a significant aspect of persuasion, it has been paid little attention in most rhetoric references [21].

3.1.1.2 Rhetorical Figures

A rhetorical figure is defined as an artful deviation of an expression from its ordinary meaning [10], i.e., the use of a word or words to deviate from its original meaning. Moreover, McQuarrie and Mick [29] observe that the deviation occurs at the level of form rather than content. They state that a rhetorical figure occurs once an expression deviates from what is expected.

3.1.1.3 Classification of Figures

There are several ways to classify rhetorical figures. They have typically been classified into various groups based on their function, such as figures of repetition, figures of amplification, figures of grammar, and figures of omission. However, there are other classifications that are commonly used. The first classification divides the rhetorical figures into two classes: schemes and tropes. Schemes describe deviations from the normal arrangement of words, for example, "Why not waste a wild weekend at Westmore Water Park?"[5] Tropes describe deviations from the normal meaning or significance of words. "Life is a beach" is one example of a trope. Rhetorical figures can also be divided according to figures of speech and figures of thought. The purpose of a figure of speech is to convey verbal expression while a figure of thought conveys ideas.

3.1.1.3.1 Figures of Repetition

The main purpose of figures of repetition is to provide either emphasis, clarity, emotional effect, or amplification. The figures of repetition can also be used as a diagnostic device to uncover the meaning of a text. Figures of repetition occur throughout a text, at the lexical, syntactical, and

morphological levels. Since figures of repetition often occur at surface level, they can be detected and classified computationally with relative ease. Figures of repetition convey strong emphasis so people tend to use these figures more often in written or spoken material. An example of a figure of repetition is provided in Figure 3.1.

Mad world! **Mad** kings! **Mad** composition!

Figure3.1: Anaphora

Anaphora is a figure of repetition that involves repeating the same word or group of words at the beginning of successive clauses. We will be investigating the use of several figures of repetition, including *polyptoton*, *antimetabole, isocolon,* and *epanalepsis*. According to Gawryjolek [14], these figures are the most accurate, reliable, and easily detectable figures of repetition.

3.1.1.3.1.1 Antimetabole

Antimetabole is a repetition of words, in adjacent sentences or clauses but in reserve grammatical order. Antimetabole is sometimes known as chiasmus [5]. For instance, in the sentence, " Woe unto them that call evil good, and good evil; that put darkness for light, and light for darkness; that put bitter for sweet, and sweet for bitter! —Isaiah 5:20 ."

3.1.1.3.1.2 Isocolon

Isocolon is defined as a series of similar structured elements that have the same length, which is kind of parallelism. It also consider as a figure of repetition of clauses and phrases. For example, in the sentence, "I woke, I eat, I sleep", the pronoun "I" is repeated three times. Since an isocolon is intended for parallelism, the verbs "wake", "eat", and "sleep" are parallel to each other. This sentence is an example of isocolon.

3.1.1.3.1.3 Polyptoton

Polyptoton is the repetition of a word, but in a different form, using a cognate of a given word in close proximity. For example, "With eager feeding food doth choke the feeder," the words "feeding," "food "and "feeder" are derived from the same root, "food", but appear in different forms. In order to automatically detect this type of figure, the detection procedure needs to incorporate methods to reduce a term to its root. One such method uses the PORTER Stemmer [35], as well as an online dictionary or thesaurus, such as WordNet [30].

3.1.1.3.1.4 Epanalepsis

Epanalepsis is the repetition at the end of a line, phrase, or clause of a word or words that occurred at the beginning of the same line, phrase, or clause [5]. For example, in the sentence, "The king is dead, long live the king," the phrase "The king" appears twice. Although this sentence is an example of epanalepsis, the two repetitions of "The" and "king" are each examples of ploche.

3.1.1.4 JANTOR

JANTOR [14] is a computational annotation tool for detecting and classifying rhetorical figures. JANTOR has two working modes: annotation and annotator. Annotation mode provides capabilities to allow users to manually enter, update, or delete annotations for the rhetorical figures in a text. A user can manually perform the annotation of a text from an existing annotated file or create a new one and starting to add new annotation or editing the already existing annotations. At any point, the user is allowed to delete all annotations of a figure or figures, resulting in all composed parts being erased. In addition, several rhetorical figures are available to choose from, and the user can change the type of rhetorical figure for a specific annotated text. Moreover, every change or update on any annotated text will be identified by the annotator name associated with a pragmatic cue, which is small portion of information that identifies the purpose of a specific annotation.

3.1.1.4.1 File Format

HTML and XML files are the two input formats that been supported by JANTOR. Essentially, each XML file is associated with the HTML one that the annotation was performed on since the XML file contains all the annotation information of the HTML file. Moreover, a user is able to open as many files as he/she wants and annotates rhetorical figures in any one of them or identifies particular type of pragmatic evidence in some or all of them. This is one of the good features of JANTOR.

3.1.1.4.2 Annotation Schema

This is one of important features of the JANTOR. For every input HTML file is meant for annotation, this feature keeps HTML file unchanged and it creates a new XML file to save all annotations that has been made on the HTML file. This procedure is called a stand-off annotation.

3.1.1.4.3 Navigation

Several features are involved in the navigation panel. These features allow the user to walk through over all the marked rhetorical figures. The user able to select certain set of figures to be displayed. Moreover, the user, at any time, can choose which types of figures to be hidden or not. Figure 3.2 shows the navigation panel and three figures have been marked to display.

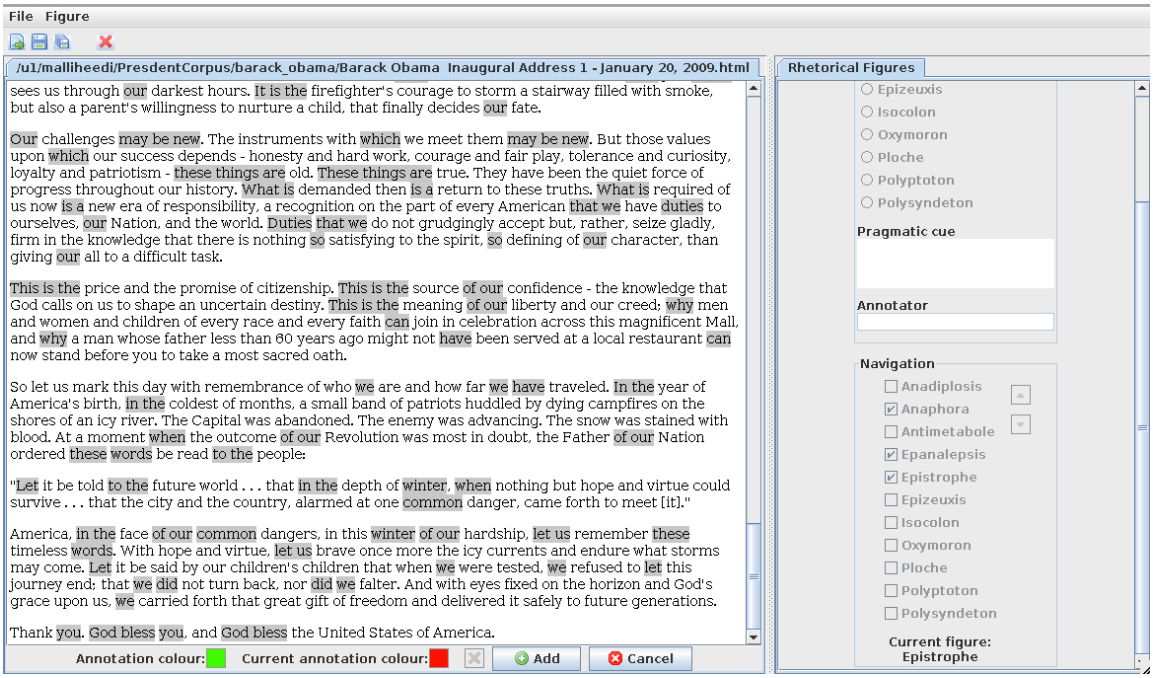

Figure 3.2: Navigation Panel

3.1.1.4.4 Detection

In the annotator mode, the computational annotator automatically detects and classifies rhetorical figures in a text. Once the HTML file is loaded, the user selects which figure to detect. The annotator then displays all annotated clauses. The annotator tends to acquire more time in the first detection since sentences in the HTML file are segmented and parsed. Moreover, the rhetorical figure "Polyptoton" is required an additional step that is meant to find the derived forms for every term that happens to be within the sentence boundary. Figure 3.3 shows several annotations in detecting rhetorical figures. Thus, the current annotation for the last figure that had been detected is red, and all previous annotations are denoted in gray. Once the annotator detects all instance of specific rhetorical figure, it lists all occurrence of that figure in a file.

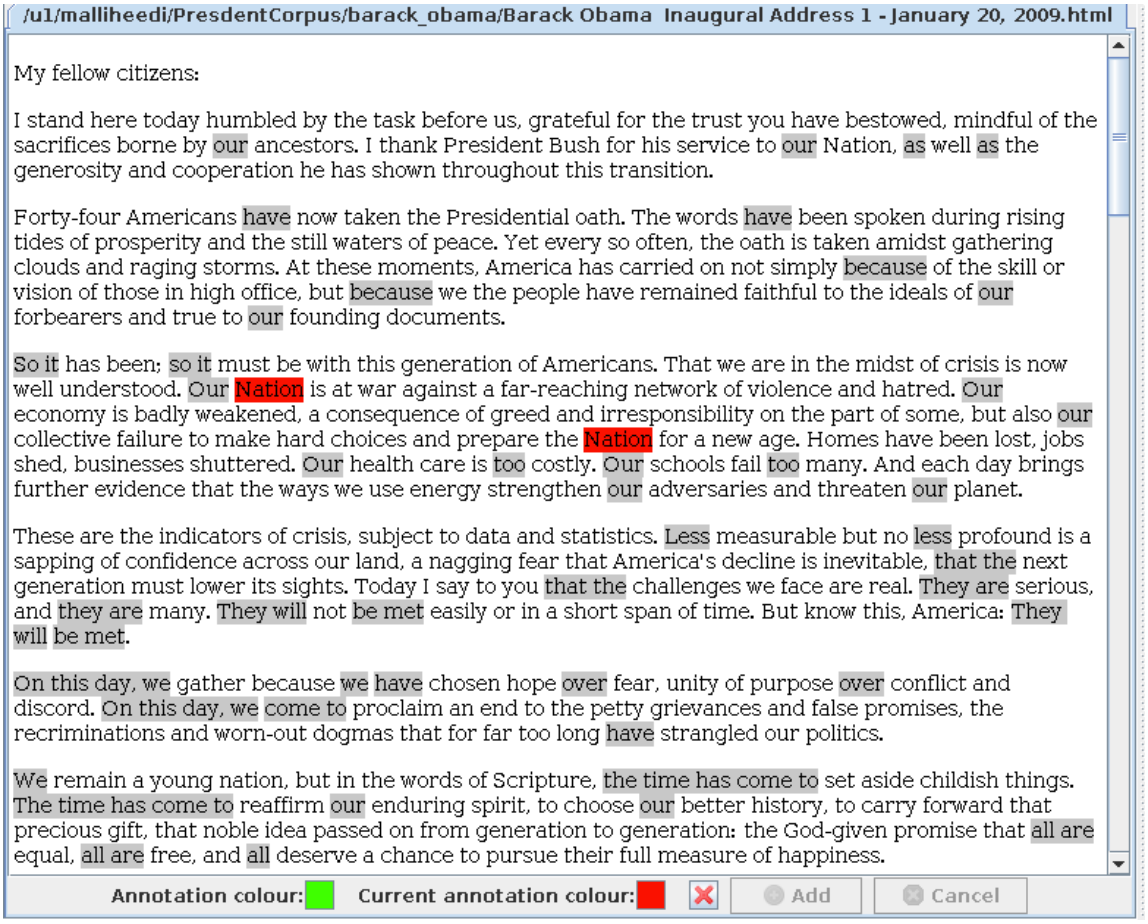

Figure3.3: Annotation Panel

3.1.1.4.5 Pre-processing

We developed this stage as an extenison to the annotator of rhetorical figures. Once the annotator detects all figures. This stage identifies the location of every rhetorical figure that occureed in a text. Thus, each sentence appears in the text is assosicated with its rhetorical figures. In other words, every sentence is accomapined with the instances of rhetorical figures that occurred in it. The idea behind that is to assign a score to each sentence depends on the occurrence and the type of rhetorical figure. It is also essential to determine the significance of sentences with higher number of rhetorical figures.

3.1.2 Multi-Document Summarization

Our approach towards summarization (RetMEAD) is based on the MEAD system [36][38] with additional new feature called Rhetorical Figure Value. MEAD is a multi-document summarizer that uses cluster centroids to produce summaries. In this subsection, we describe the MEAD system first. Then, we will explain how our new feature incorporates with other MEAD features.

Radev et al. [38] describe how the MEAD summarizer was built and evaluated. Basically, a centroid of a cluster is a group of terms that statistically represents the cluster. MEAD performs the following steps when producing summaries.

3.1.2.1 Topic Detection and Tracking (TDT)

Initially, the system identifies all articles about an emerging event using a modified TF*IDF in order to group articles with the same topic in the same cluster. In other words, each cluster includes several articles that have the same topic. Therefore, every new article is grouped into a cluster if its TF*IDF is close to the centroid of the cluster based on the following formula:

$$sim(D,C) = \frac{\sum (d_k * c_k * idf(k))}{\sqrt{\sum_k (d_k)^2}\sqrt{\sum_k (c_k)^2}}$$

Equation 3.1: Cnetiod Document Simalirty

3.1.2.2 Centroid-based Summarization (CBS)

Centroid-based summarization (CBS) was developed by Radev et al. [38] as a new method for multi-document summarization. CBS uses the centroids of clusters that were produced in the first step by TDT. Then, it selects sentences that are central to the topic of the whole cluster. Thus, the CBS selects each sentence based on its score, which is the sum of three features.

3.1.2.3 Scoring Features

Three different features are used to score a sentence: Centroid value, Positional value, and Length feature.

3.1.2.3.1 Centroid Value

A centroid is a set of words that are statistically important to a cluster of documents. Centroid value uses the centroids of the cluster that have been produced in the Topic detection and tracking step using TF*IDF to assign a score to each word in the text. Thus, the centroid value Ci is the sum of centroid values of all words Cw,i in a sentence Si. For instance, assume that we have the centroid values of words (Obama = 44.68; Hillary = 32.50; Clinton = 29.40; morning = 12.25) , and the sentence "President Obama meets with Secretary of State Hillary Rodham Clinton in the morning". Thus, the total score of this sentence is 118.83.

$$C_i = \sum_w C_{w,i}$$

Equation 3.2: Centroid Value

3.1.2.3.2 Positional Value

The positional value for all sentences in a document is calculated as follow:

$$P_i = \frac{(n - i + 1)}{n} * C_{\max}$$

Equation 3.3: Positional Value

However, the first sentence is assigned the highest-ranking sentence in the document, which is $C_{\max}$, since the first sentence tends to possess more salient information than other sentences.

3.1.2.3.3 Length Feature

The length feature is a cut-off feature that means every sentence with a length shorter than threshold will received a score of zero regardless of other feature scores. Thus, if a sentence has a length higher than the default threshold, which is 9 words, it will received a score that is the combination of other feature scores.

3.1.2.3.4 Rhetorical Figure Value

The rhetorical figure value RF_i is our new feature that added to our system RetMEAD along with MEAD features. RF_i is the sum of a rhetorical figure occurred in a document divided by the total number of all figures in the document. It is calculated as follows:

$$RF_i = \frac{n}{N}$$

Equation 3.4: Retorical Figure Value

Where i is a type of rhetorical figure such as polyptoton or isocolon, and n is the total number of i occurrence in a document. N is the total number of all figures, which includes antimetabole, epanalepsis, isocolon, and polyptoton, that occurred in the document. For example, assume that we have a document contains rhetorical figures (12 antimetaboles; 2 epanalepsises; 100 isocolons; 35 polyptoton) and we would like to have the value of each figure. Thus, we apply the rhetorical figure value as follows:

$RF_{polyptoton} = 35/149 = 0.2348$

$RF_{antimetabole} = 12/149 = 0.081$

$RF_{epanalepsis} = 2/149 = 0.0134$

$RF_{isocolon} = 100/149 = 0.671$

Thus, for every instance of rhetorical figures that occurred over two sentences, the score of that rhetorical figure is divided between the two sentences because this instance formed in two different sentences.

3.1.2.4 Sentence Score

A sentence score is the sum of scores of all words in the sentence. Since no learning algorithm has been incorporated to predicate the weight for each feature, Readv et al. assign an equal weight for all features. We also assume the weight of the new rhetorical figure feature equals to one.

$$\text{Score}(S_i) = W_c C_i + W_p P_i + W_{RF} RF_i \; ; \quad \text{if Length}(S_i) > 9$$
$$\text{Score}(S_i) = 0 \; ; \quad \text{if Length}(S_i) < 9$$

Equation3.5: MEAD Sentence Score

While the input is cluster of d documents with n sentences, the output is the number of sentences from the cluster with highest score multiplied by compression rate r .

3.1.2.5 Evaluation Metrics

Two metrics are incorporated for the evaluation of single and multi-document summaries: cluster-based relative utility (CBRU), which concerns the degree of relevance of a sentence to the topic of the cluster, and cross-sentence information subsumption (CSIS), which refers to certain sentences duplicating the information in other sentences.

3.1.2.6 Results and Performance of MEAD

A small corpus of news articles is used for MEAD evaluations by Readv et al. in [38] to test the performance of the system, made up of six clusters, 2 of which are from TDT corpus and the remaining are from the Usenet newsgroups, consisting of 27 articles in total. Radev et al. compared the performance between the MEAD summarizer and lead-based summarizer, which selects the first

sentence from each cluster until the length of a summary is achieved. In their results, the performance of the MEAD summarizer outperformed the lead-based summarizer in 29 out of 45 cases.

3.1.3 The Corpus

We perform a series of experiments using our proposed system on presidential speeches to compare the quality of the summaries produced using the MEAD summarizer with and without the aid of rhetorical figuration as an additional metric for evaluating the importance of a sentence. The corpus of presidential speeches is grouped by presidents, with each president represented by various types of presidential addresses. These speeches are organized by type: campaign, state of union, and inaugural. In order to evaluate the usefulness of rhetorical summarization as a metric, the baseline system will be first tested on the same corpora, which is the MEAD summarizer [38] alone without the new rhetorical figure value.

3.1.3.1 Data Set

The data set is composed of various types of presidential speeches: campaign, state of union, and inaugural. The data set is constructed by sampling one or more articles from each type of presidential speeches for a president every five years in the past twenty years. Then, we use this data set to calculate the frequencies of the various rhetorical figures over the different types of speeches. Then, using these frequencies, we assign a score for each figure depending on its frequency, which is the aforementioned rhetorical figure feature. Thus, we incorporate theses scores with the MEAD score to produce the summary. Moreover, rhetoricians manually summarize the original speeches. Thus, we compare their summaries with the system summaries using summarization evaluation metrics.

3.1.4 Experimental Procedure

Our multi-document summarization system incorporating rhetorical figuration as a metric will be developed as follows:

1- Use JANTOR to perform annotation of rhetorical figures on every document in our corpus.

2- Assign a value to each type of rhetorical figure occurring in the text based on the rhetorical figure feature (RF).

3- Repeat the above process for each document in the cluster.

4- After all documents in the cluster are annotated, the cluster will be entered into the MEAD summarizer [38]. Then, the score for every sentence in the cluster will be the sum of the MEAD score and the rhetorical figuration metric score.

An evaluation will then be performed comparing the results obtained using the baseline summarizers of our combined system using rhetorical figuration as the metric. We will aim to validate our hypothesis: using rhetorical figuration metrics should improve the performance of the text summarization system to create more accurate and adequate summaries.

Chapter 4

Results and Evaluation

In this chapter, we present and discuss the results that are produced from our general approach towards multi-document summarization using rhetorical figure metrics. First, we briefly describe the evaluation metrics of summaries. Next, we describe our testing corpus and how we grouped them. Then, we present our experiments that consist of: (1) the experiments of various rhetorical figures using the JANTOR tool, and (2) the experiments of the MEAD summarization system that includes the produced automated summaries using various rhetorical figures information such as Antimetabole, Polyptoton, and Isocolon. Finally, we compare and discuss the evaluation results of both the baseline summaries and our approach summaries against human summaries.

4.1 Evaluation

Evaluation of summarization requires human judgment in several quality metrics such as coherence, grammaticality, and content [26]. Although the quality metrics are significant for the evaluation of summarization systems, these metrics would require an enormous amount of human effort which is very expensive and difficult to manage [22]. Thus, over the past years, many researchers in the area of text summarization have attempted to develop an automatic evaluation metric for summarization

tasks. Although Saggion et al. [40] developed three evaluation metrics (cosine similarity, unit overlap, and longest common subsequence), the authors do not provide how the results of these methods would be correlated to human judgment.

4.1.1 Evaluation Types

Essentially, evaluation methods are broadly classified into two types: extrinsic and intrinsic [27]. Extrinsic evaluations are developed to evaluate and measure the capability of summarizations to perform other tasks. For example, relevance assessment is used to ask experts to determine whether a summary and topic are related or not. Intrinsic evaluations, on the other hand, measure the quality of a summary. Such a task can be performed to compare human summaries against system summaries by measuring recall and precision of both summary units, which can be sentences or words. Recall is defined as the ratio of the number of units of system summaries that overlap human summaries to the total number of units in human summaries. Precision, however, is defined as the percentage overlap between human summaries and system summaries.

In this thesis, we use an intrinsic evaluation system called ROUGE [22] to evaluate our summarization system.

4.1.1.1 ROUGE

ROUGE stands for Recall Oriented Understudy of Gisting Evaluation [22]. ROUGE provides several metrics to automatically evaluate system summaries overlapping with "ideal" human summaries. In other words, every summary is evaluated by counting the number of n-gram, which is a sequence of n words that overlaps with human summaries. ROUGE consists of four major metrics: ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S.

4.1.1.1.1 ROUGE-N

ROUGE-N is an n-gram recall measure that calculates the number of n-grams in common between the human summaries and system summaries. It can be calculated as follows:

$$ROUGE-N = \frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

Equation 4.1: ROUGE-N Formula

Where n stands for the length of n-gram, $gram_n$, and $Count_{match}(gram_n)$ is the maximum number of n-grams occurring in system summary and reference summaries which are human summaries. This formula applies to a single reference and multiple reference; however, if multiple reference summaries are used, it also computes the pair-wise summary-level ROUGE-N between a

"model" system summary *s* and each reference summary *ri*. Then, it takes the maximum of the pair-wise summary-level ROUGE-N score as the final score as follows:

$$ROUGE - N_{multi} = \arg\max_i ROUGE - N(r_i, s)$$

Equation 4.2: ROUGE-N$_{multi}$ Formula

This approach is also applied to ROUGE-S, ROUGE-L, and ROUGE-W. To elaborate on ROUGE-N measure, suppose the following example [22]:

Example 4.1:

S1. Police killed the gunman.

S2. Police kill the gunman.

S3. The gunman kill police.

We choose ROUGE-2, which is N=2 (bigram), only for demonstration purposes. Let S1 be the reference summary sentence, and S2 and S3 be he model summary sentences.

Sentence	Bigram
Police killed the gunman	"Police killed", "killed the", "the gunman"
Police kill the gunman	"Police kill", "kill the", "the gunman"
The gunman kill police	"The gunman", "gunman kill", "kill police"

Table 4.1: Example of Bigram

Each one of these sentences consists of three bigrams. From Table 4.1, S2 and S3 share one bigram, which is "the gunman", with the reference summary sentence S1. Thus, the scores for both S2 and S3 are the same, which equal to 2/4. However, S2 and S3 possess dissimilar meanings, and S2 is closer to S1 than S3.

4.1.1.1.2 ROUGE-L

ROUGE-L is a measure based on Longest Common Subsequence (LCS) between reference summaries and model summaries. It is based on the concept that the longer subsequence of words that are in common between summaries, the greater the similarity will be. Given the sequence Y = [y1,y2,…yn] is a subsequence to another sequence K = [k1,k2, …, km], then there is an existing increasing sequence [i1,i2,…,ix] of indices of K such that for all j=1,2,…,x, we have Kij = Yj [11] [22]. For two sequences C and D, LCS is the sequence with the maximum length.

The LCS-based F-measure is used to estimate the similarity between two summaries where X is the reference summary with length of n and Y is model summary with length m, as follows:

$$P_{lcs} = \frac{LCS(X,Y)}{n}$$

Equation 4.3: ROUGE-L Precision

$$R_{lcs} = \frac{LCS(X,Y)}{m}$$

Equation 4.4: ROUGE-L Recall

$$F_{lcs} = \frac{(1+\beta^2)RlcsPlcs}{Rlcs + \beta^2 Plcs}$$

Equation 4.5: ROUGE-L F-measure

Where *LCS(X, Y)* is the longest common subsequence between *X* and *Y*, and $\beta = Plcs / Rlcs$ when $\partial Flcs / \partial Rlcs = \partial Flcs / \partial Plcs$. ROUGE-L is known as the LCS-based F-measure. We use example 4.1 to illustrate the ROUGE-L measure. Thus, with β=1, S2 is assigned a score of 3/4 because it shares with S1 three words in the same sequence of S1. However, S3 receives a score of 2/4 since it only has the words "the gunman" in common with the summary sentence S1 according to ROUGE-L. ROUGE-L has one serious drawback in that it considers words only in the same sequence of the

reference summary. For instance, if there is S4, the model summary sentence would be as follows [22]:

S4. The gunman police killed.

With S1 as the reference summary sentence, ROUGE-L assigns S4 a score of 2/4 because it only counts either "the gunman" or "police killed" with the reference summary sentence S1. Thus, S4 has the same score of S3; however, ROUGE-2 assigns S4 a higher score equals to 1, which is better than S3.

4.1.1.1.3 ROUGE-W

ROUGE-W is a Weighted Longest Common Subsequence (WLCS). This measure is intended to evaluate model summary sentences that have consecutive matches with reference summary sentences higher than other sentences with non-consecutive matches. For example, there are a reference summary sentence (X1) and two model summary sentences (X2) and (X3), respectively, as follows:

Example 4.2:

X1: [A B C D E F G H]

X2: [A B C J K L M N]

X3:[A J K M B L N C]

In the case of ROUGE-L, X2 and X3 are assigned the same score of 3/8. However, in ROUGE-W, X2 would be evaluated higher than X3 because X2 has successive matches with X1. For two sequences X and Y of length m, and n, respectively, WLCS-based F-measure is calculated as follows:

$$R_{wlcs} = f^{-1}\left(\frac{WLCS(X,Y)}{f(m)}\right)$$

Equation 4.6: ROUGE-W Recall

$$P_{wlcs} = f^{-1}\left(\frac{WLCS(X,Y)}{f(n)}\right)$$

Equation 4.7: ROUGE-W Precision

$$F_{wlcs} = \frac{(1+\beta^2)RwlcsPwlcs}{Rwlcs + \beta^2 Pwlcs}$$

Equation 4.8: ROUGE-W F-measure

Where f^{-1} is the inverse function of f .

4.1.1.1.4 ROUGE-S

ROUGE-S is a skip-bigram co-occurrence statistics measure that counts any pair of words in their sentence order in a model summary that overlaps with a reference summary. ROUGE-S allows arbitrary gaps between bigrams, so each sentence will have a number of skip-bigrams depending on the number of words in it. To elaborate, ROUGE-S is applied on example 4.1 as follows:

S1. Police killed the gunman.

S2. Police kill the gunman.

S3. The gunman kill police.

S4. The gunman police killed.

Since there are four words in every sentence, every sentence will have:

C(4,2) = 4!/(2!*2!) = 6 skip-bigrams

Thus, every sentence has the following skip-bigrams as follows:

ID	Skip-Bigram
S1	"Police killed", "Police the", "Police gunman", "killed the", "killed gunman", "the gunman"
S2	"Police kill" ,"Police the" ,"Police gunman" ,"kill the" ,"kill gunman" ,"the gunman"
S3	"The gunman" ,"The kill" ,"The police" ,"gunman kill" ,"gunman police" ,"kill police"
S4	"The gunman" ,"The police" ,"The killed" ,"gunman police" ,"gunman killed" ,"police killed"

Table 4.2: Examples of Skip-Bigram

From Table 4.2, S2 shares three skip-bigrams with the reference summary sentence S1 that are "Police the", "Police gunman", and "the gunman." S2 is assigned a score of 3/6. S3 has one skip-bigram overlaps with S1, which is "the gunman", so S3 receives a score of 1/6. S4 is assigned a score of 2/6 because S4 has two skip-bigrams in common with S1.

For two sequences X and Y with both having the length of m and n, respectively, a skip-bigram based F-measure is calculated as follows:

$$R_{skip2} = \frac{SKIP2(X,Y)}{C(m,2)}$$

Equation 4.9: ROUGE-S Recall

$$P_{skip2} = \frac{SKIP2(X,Y)}{C(n,2)}$$

Equation 4.10: ROUGE-S Precision

$$F_{skip2} = \frac{(1+\beta^2)R_{skip2}P_{skip2}}{R_{skip2} + \beta^2 P_{skip2}}$$

Equation 4.11: ROUGE-S F-measure

Where $SKIP2(X,Y)$ is the number of matched skip-bigram between reference summary X and model summary Y, β controls the relative importance between P_{skip2} and R_{skip2}. C is a combination function.

One disadvantage of ROUGE-S is that it does not assign any score to a model summary sentence when it does not share any bigram with a reference summary sentence. Thus, this problem will penalize sentences having the same words of the reference sentence, but in different bigrams. For example, the following sentence received a ROUGE-S score of zero [22]:

S5: gunman the killed police.

S5 has the same words of the reference summary sentence S1, but in reverse grammatical order. Thus, the skip-bigrams of S5 would be ("gunman the", "gunman killed", "gunman police", "the killed", "the police", "killed police"). None of these bigrams have a match with S1 bigrams. Lin [22] proposed ROUGE-SU as an extension of ROUGE-S with a unigram as a counting unit. Thus, ROUGE-SU would overcome this problem.

4.1.1.1.5 How ROUGE Correlates with Human Judgment

Lin [22] computed the correlation between scores of summaries that have been assigned from both humans and ROUGE. Lin has compared and computed these scores for three Document Understanding Conference (DUC) using Pearson's product moment correlation coefficients, Spearman's rank order correlation coefficients, and Kendall's correlation coefficients. The evaluation data include DUC 2001, 2002, and 2003. Therefore, the intuition was a good evaluation measure that should evaluate good summaries with higher scores and vice versa. Lin stated several conclusions:

For single document summarization tasks, ROUGE-2, ROUGE-L, ROUGE-W, and ROUGE-S performed well.

ROUGE-1, ROUGE-L, ROUGE-W, ROUGE-SU4, and ROUGE-SU9 worked perfectly for evaluating very short summaries.

ROUGE-1, ROUGE-2, ROUGE-S4, ROUGE-S9, ROUGE-SU4, and ROUGE-SU9 performed well when stopwords removal was applied.

When stopwords were eliminated, it improved the correlation with human evaluation.

When multiple reference summaries were used, the correlation with human evaluation was increased.

4.2 Data Set

We have created a set of clusters for some U.S. presidents that include presidential speeches from The American Presidency Project[1]. Thus, we have four collections of clusters for four U.S. presidents: Barack Obama, George .W. Bush, Bill Clinton, and George Bush. Furthermore, each collection contains three clusters which include various types of speeches: state of union, campaign, and

[1] http://www.presidency.ucsb.edu/

inaugural speeches, i.e., we have a cluster for each one of the three types of speeches. Each cluster contains from one to two speeches, and these speeches contain from 90 to 700 sentences. These tested clusters are used for the evaluation process in a way that each cluster is summarized in various ratios: 5%, 10%, and 20% from the actual texts in the cluster. Furthermore, these clusters are also summarized by sentence number: 10 and 20 sentences.

4.3 Experiments of Rhetorical Figures

The following sections show the experiments of various rhetorical figures. Since we are testing four types of rhetorical figures, Antimetabole, Epanalepsis, Isocolon, and Polyptoton, we use JANTOR to detect these figures in all clusters of presidential speeches. In other words, each presidential speech is annotated by these rhetorical figures before the summarization procedure. We discuss the performance of JANTOR in terms of detecting of these figures.

4.3.1 Antimetabole

Although antimetabole states that the repetition of words in adjacent clauses or phrases are in reverse grammatical order, this definition does not state whether different word forms and word types are considered or not [14]. Gawryjolek addresses these issues in the JANTOR tool by considering all types of words, such as determiners and prepositions, and by looking for only the repetition of the exact words. Although it is obvious that including prepositions and determiners will definitely produce more antimetaboles, which are not salient in a text, it is important to include them in order not to decrease the recall value [14].

Example 4.3:

"The fact is, our economy did not fall into decline overnight, nor did all of our problems begin when the housing market collapsed or the stock market sank. We have known for decades that our survival depends on finding new sources of energy, yet we import more oil today than ever before. The cost of health care eats up more and more of our savings each year, yet we keep delaying reform. " [2]

In the above example, there are some repetitions of words in reverse grammatical order, which are examples of the antimetabole.

No	Repetition
1	our…did…did…our
2	our…market …market…our
3	yet…more…more…yet
4	our…more…more…our
5	we…more…more…we

Table 4.3: Examples of Antimetabole

[2] President Barack Obama's speech to the Joint Session of Congress, February 24, 2009.

Table 4.3 shows examples of an Antimetabole that satisfy the definition with both constraints in the JANTOR tool. Thus, it presents examples of both conjunctions, such as *yet*, and determiners of a personal determiner, such as *we*, a possessive determiner such as *our*, and an additive determiner such as *more*.

Another example is presented in Example 4.3.

Example 4.4:

"To be **kissed** by a **fool** is stupid; To be **fooled** by a **kiss** is worse.**"**[3]

In Example 4.4, there are two words that are repeated, but in different word forms. Since the JANTOR tool restricts the definition of an antimetabole by only considering the repetition of exact words in reverse grammatical order, the word "*kissed*" and "*kiss*" are not considered as an example of an antimetabole. The other words "*fool*" and "*fooled*" are also ignored.

4.3.2 Epanalepsis

Epanalepsis is defined as the repetition of the same word or phrase that occurs in the beginning and the end of a sentence, clause, or line. JANTOR was able to detect correctly most of the epanalepsis instances in our corpus. However, there are a few false positives resulting from detecting epanalepsis. Example 4.5 illustrates this situation.

[3] Ambrose Redmoon

Example 4.5:

"*Starting this year, no American will be forbidden from serving the country they love because of who they love*."[4]

In Example 4.5, the group of words "they love" is repeated in the sentence twice. It is actually a false positive example of an epanalepsis because it appears at the beginning and the end of the same phrase, "they love because of who they love." Thus, this group of words is not qualified to be an example of epanalepsis. However, this group is a perfect example of another rhetorical figure namely, anadiplosis.

4.3.3 Isocolon

Isocolon is the only figure of parallelism in our figure set. Isocolon requires a set of similar structured phrases or sentences that have the same length. In order to discover isocolon, the JANTOR tool is used for the word count and parse tree structure. Since it is not obvious what the similar structured phrases would be, Gawryjolek [14] made two assumptions to resolve this issue. The first assumption is that both phrases should have the identical parse-tree structures, which means they have the same parts of speech and the depth of tree nodes. Thus, the differences between the two phrases are the words that used. If the first assumption is not satisfied, then the difference between part-of-speech labels of words in the phrases is smaller than certain set threshold, which is usually 1.

[4] President Barack Obama's speech to the Joint Session of Congress, January 25, 2011.

Example 4.6:

"They have done so during periods of prosperity and tranquility, and they have done so in the midst of war and depression, at moments of great strife and great struggle." [5]

In Example 4.6, JANTOR was able to detect several examples of Isocolon. The first instance of isocolon is the similarly parallel structures in both groups of words "*prosperity and tranquility*" and "*war and depression*". JANTOR also considered another example of Isocolon, which are the words "*of prosperity and tranquility* " and "*of war and depression*".

4.3.4 Polyptoton

Polyptoton is one of the figures of repetition that requires a search for various word forms. Essentially, the JANTOR tool is able to detect correctly several instances of polyptoton in our test corpus. From our observation of the detection of polyptoton, however, we find that the JANTOR tool has only one issue in detecting polyptoton. JANTOR fails to detect some words that are in different stem forms. Examples are presented to elaborate how the detection of polyptoton is proceed.

[5] President Barack Obama's speech to the Joint Session of Congress, January 27, 2010

Example 4.7:

"With eager **feeding food** doth choke the **feeder**." [6]

In this example, the JANTOR tool was able to successfully detect the word "***feeding***" and the word "***feeder***". However, it was not able to discover the word "***food***" which has a different stem form from the word "***feed***". Thus, Gawryjolek [14] suggested a solution for this issue by using glosses of words, which is basically a set of words that accompanies word definitions in WordNet.

Example 4.8:

"*After one of the most difficult years in our history, they remain **busy** building cars and teaching kids, starting **businesses** and going back to school.*"[7]

In Example 4.8, JANTOR detected the word "busy" and "business" because both words share a common word stem. The definition of polyptoton states that polyptoton is the repetition of a word or words in different cognate forms, so this example indicates one of the Polyptoton instances in our test corpus.

4.4 Multi-Document Summarization Experiments

As we discussed in Section 4.2, the tested presidential speeches are annotated by the JANTOR tool for different rhetorical figures that include antimetabole, epanalepsis isocolon, and polyptoton .

[6] Shakespeare.
[7] Barak Obama's address before a joint session of the Congress on February 24,2009.

Thus, we have two types of summaries for our data set. The first summaries are MEAD summaries, which are the baseline summaries, and the second ones are our approach summaries (RetMEAD).

4.4.1 Baseline Summaries

The baseline summaries are produced by the basic MEAD features without the rhetorical figure feature. Thus, we referred to the basic MEAD system as MEAD or the baseline. Figure 4.1 shows a baseline summary with a ratio of 5% from the actual speech.

[1] My fellow citizens: I stand here today humbled by the task before us, grateful for the trust you have bestowed, mindful of the sacrifices borne by our ancestors.
[2] I thank President Bush for his service to our Nation, as well as the generosity and cooperation he has shown throughout this transition.
[3] At these moments, America has carried on not simply because of the skill or vision of those in high office, but because we the people have remained faithful to the ideals of our forbearers and true to our founding documents.
[4] The success of our economy has always depended not just on the size of our gross domestic product, but on the reach of our prosperity, on our ability to extend opportunity to every willing heart, not out of charity, but because it is the surest route to our common good.
[5] And because we have tasted the bitter swill of civil war and segregation and emerged from that dark chapter stronger and more united, we cannot help but believe that the old hatreds shall someday pass; that the lines of tribe shall soon dissolve; that as the world grows smaller, our common humanity shall reveal itself; and that America must play its role in ushering in a new era of peace.
[6] And to those nations like ours that enjoy relative plenty, we say we can no longer afford indifference to suffering outside our borders, nor can we consume the world's resources without regard to effect, for the world has changed, and we must change with it.

Figure 4.1: A baseline summary form President Barack Obama Inaugural Speeches

4.4.2 MEAD Summaries using Rhetorical Information

This type of summary is produced by the modified MEAD system, which includes the additional feature of rhetorical figures. We called our approach summarization system "RetMEAD". Figure 4.2 shows our system summary using Isocolon information.

[1] The time has come to reaffirm our enduring spirit, to choose our better history, to carry forward that precious gift, that noble idea passed on from generation to generation: the God given promise that all are equal, all are free, and all deserve a chance to pursue their full measure of happiness.
[2] The success of our economy has always depended not just on the size of our gross domestic product, but on the reach of our prosperity, on our ability to extend opportunity to every willing heart, not out of charity, but because it is the surest route to our common good.
[3] And because we have tasted the bitter swill of civil war and segregation and emerged from that dark chapter stronger and more united, we cannot help but believe that the old hatreds shall someday pass; that the lines of tribe shall soon dissolve; that as the world grows smaller, our common humanity shall reveal itself; and that America must play its role in ushering in a new era of peace.
[4] This is the meaning of our liberty and our creed; why men and women and children of every race and every faith can join in celebration across this magnificent Mall, and why a man whose father less than 60 years ago might not have been served at a local restaurant can now stand before you to take a most sacred oath.
[5] In the year of America's birth, in the coldest of months, a small band of patriots huddled by dying campfires on the shores of an icy river.
[6] At a moment when the outcome of our Revolution was most in doubt, the Father of our Nation ordered these words be read to the people: Let it be told to the future world that in the depth of winter, when nothing but hope and virtue could survive that the city and the country, alarmed at one common danger, came forth to meet it.

Figure 4.2: Our system summary form President Barack Obama Inaugural Speeches

Although both summaries in Figure 4.1 and Figure 4.2 from the same speech with a ratio of 5%, it is obvious that each summary contains different sentences. It also appears that a human-based summary in Figure 4.4 possess some dissimilar sentences to both summaries in Figure 4.1 and Figure 4.2.

4.4.3 Human Summaries

4.3.3.1 Field Study

The purpose of our study is to develop an automated multi-document summarization system that is able to produce accurate and adequate summaries for specific texts. This study was conducted using written materials and then theses written materials are entered using Graphical User Interface (GUI) for sentence extraction. Figure 4.3 shows an example of Inaugural speech for Barack Obama.

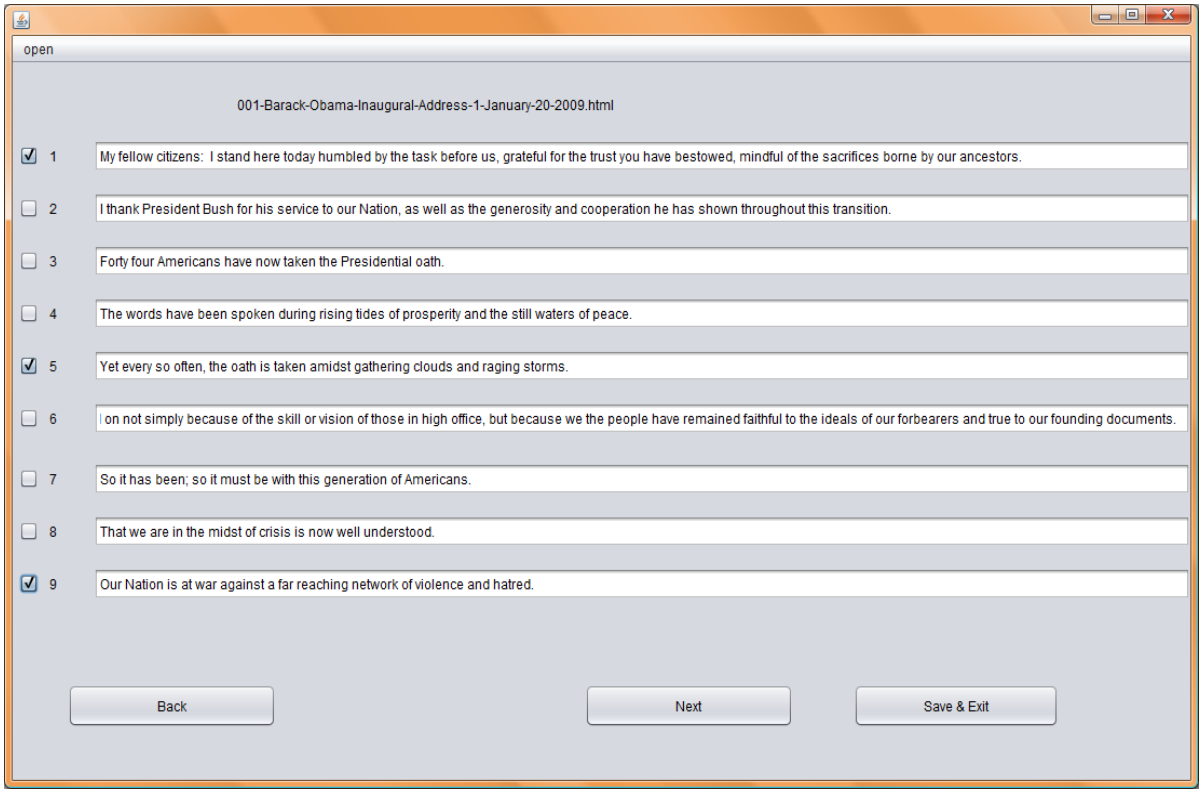

Figure 4.3: Graphical User Interface for Sentence Extraction

In Figure 4.3, the GUI presents each sentence of a presidential speech in a line associated with a number that reflects the position of the sentence in the speech. Then, a user can easily select one or

more sentences from the speech in order to be included in the summary. Three buttons are incorporated in order to perform the sentence extraction for a specific speech. When the button "Next" is clicked, it will present the next sentences in the speech. Thus, the user is able to choose sentences and so on till reach the end of speech. Once the end of speech is reached, the button "Save & Exit" is used to create a text file that contains the selected sentences by the user, which is called model summary. For any reason, if the user would like to change the selected sentences, the button "Back" is provided to help the user to show the previous sentences and the user can easily unselect the chosen sentences.

4.3.3.2 Participant Recruitment

The recruitment of participants was performed through posters and email scripts (see Appendix A). Emails were sent to mailing lists of graduate and undergraduate students in the English Language and Literature Department at the University of Waterloo. The field study was held at the campus of University of Waterloo from April 16, 2012 to April 26, 2012. Two participants agreed to take part in the field study. Every participant was asked to sign a consent form that explaining the purpose and procedure of the study, and each participant was required the permission to use his/her data.

4.3.3.3 Study Procedure

Participants were provided a hard copy of each presidential speech along with a list of questions (see Appendix B). These questions are natural to avoid swaying the participants decision in selecting sentences. The study includes one session that one hour long. The data set that we provided has 33 presidential speeches in total which include Campaign, State of the Union, and Inaugural speeches. Initially, participants read and review a presidential speech. Then, they answers the provided

questions by selecting sentences from the speech. Thus, each presidential speech is manually summarized with different ratios, 5% and 10%, by the participants.

4.3.3.4 Study Completion

The participants were rewarded ten dollars per hour ($10/hour) for their participation. Participants were given a feedback letter, and were asked to provide their contact information in order to provide them a copy of the study results.

These participants are two graduate students from the English Language and Literature Department at the University of Waterloo. One is a PhD student and the other is a Master's student. The age of participants ranges from 24 to 28. The participants have good knowledge of both grammar and rhetoric.

Figure 4.4 shows an example of participant summary from the Inaugural speech of President Barak Obama .

[1] Our economy is badly weakened, a consequence of greed and irresponsibility on the part of some, but also our collective failure to make hard choices and prepare the Nation for a new age.
[2] The time has come to reaffirm our enduring spirit, to choose our better history, to carry forward that precious gift, that noble idea passed on from generation to generation: the God given promise that all are equal, all are free, and all deserve a chance to pursue their full measure of happiness.
[3] The state of the economy calls for action, bold and swift, and we will act not only to create new jobs but to lay a new foundation for growth.
[4] The success of our economy has always depended not just on the size of our gross domestic product, but on the reach of our prosperity, on our ability to extend opportunity to every willing heart, not out of charity, but because it is the surest route to our common good.
[5] But those values upon which our success depends honesty and hard work, courage and fair play, tolerance and curiosity, loyalty and patriotism these things are old.
[6] What is required of us now is a new era of responsibility, a recognition on the part of every American that we have duties to ourselves, our Nation, and the world.

Figure 4.4: A human-based summary from President Barack Obama Inaugural Speeches

4.5 ROUGE Evaluation

The data set summaries, which include the baseline and our approach summaries, are evaluated using various ROUGE metrics. Theses metrics contain ROUGE-N, ROUGE-L, ROUGE-S, and ROUGE-SU. We evaluate the MEAD summaries comparing them to RetMEAD summaries using the information of one of the rhetorical figures, which includes Antimetabole, Epanalepsis, Polyptoton,

and Isocolon at the time. Thus, four comparisons among the RetMEAD summaries and MEAD summaries are performed.

4.5.1 ROUGE Evaluation for Summaries with a ratio of 5%

Table 4.4, shows the results of the evaluation of MEAD summaries and RetMEAD using the Antimetabole value in a ration of 5%.

Table 4.5 identifies two values for each ROUGE method, which are Case, and Stemmed. Case is the default ROUGE parameters. Stemmed is the default ROUGE parameters with Porter's Stemmer algorithm. Table 4.4 shows that our system RetMEAD using the Antimetabole value outperforms MEAD in every single ROUGE method. We observed that when we applied Porter's Stemmer algorithm, the results are improved significantly especially in ROUGE-1, ROUGE-L, ROUGE-S*, and ROUGE-SU*.

ROUGE	MEAD		RetMEAD	
Method	Case	Stemmed	Case	Stemmed
R-1	0.486	0.505	0.519	0.537
R-2	0.206	0.210	0.238	0.242
R-L	0.464	0.481	0.499	0.516
R-S*	0.228	0.247	0.260	0.280
R-S4	0.222	0.229	0.252	0.258
R-SU*	0.229	0.248	0.261	0.281
R-SU4	0.266	0.275	0.297	0.305

Table 4.4: ROUGE evaluation using Antimetabole with ratio of 5%.

Table 4.5 demonstrates that MEAD slightly outperforms our system RetMEAD using Epanalepsis information in most of the ROUGE methods such as ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-S4. However, we observe that when we applied Porter's Stemmer algorithm, the results of RetMEAD improved specially in ROUGE-S*, and ROUGE-SU*. Therefore, the reason for the poor results of RetMEAD in some ROUGE metrics is because the instances of Epanalepsis were very few in contrast with other rhetorical figures.

ROUGE	MEAD		RetMEAD	
Method	Case	Stemmed	Case	Stemmed
R-1	0.511	0.528	0.510	0.528
R-2	0.246	0.249	0.245	0.249
R-L	0.491	0.507	0.491	0.507
R-S*	0.249	0.267	0.249	0.267
R-S4	0.260	0.266	0.260	0.266
R-SU*	0.250	0.268	0.250	0.268
R-SU4	0.302	0.310	0.302	0.310

Table 4.5: ROUGE Evaluation using Epanalepsis with a ratio of 5%

Table 4.6 shows that RetMEAD using Polyptoton information achieves significantly better results than MEAD in all ROUGE metrics. Since Polyptoton concerns with the repetition of a word in different forms, it evaluates sentences that possess many instances of Polyptoton.

Table 4.7 indicates also that RetMEAD using Isocolon information slightly outperforms the baseline, MEAD. The results of RetMEAD achieved better performance than MEAD in each ROUGE metric.

ROUGE	MEAD		RetMEAD	
Method	Case	Stemmed	Case	Stemmed
R-1	0.514	0.532	0.530	0.548
R-2	0.242	0.245	0.275	0.277
R-L	0.492	0.508	0.510	0.526
R-S*	0.252	0.270	0.268	0.287
R-S4	0.259	0.265	0.288	0.294
R-SU*	0.253	0.271	0.269	0.288
R-SU4	0.301	0.309	0.329	0.336

Table 4.4: ROUGE Evaluation using Polyptoton with ratio of 5%.

The average ROUGE evaluation of summaries for all figures, which is shown in figures (Table 4.8), demonstrates that RetMEAD achieves better ROUGE methods scores than MEAD.

We observe that our approach summarization system RetMEAD performed well in summaries with ratio of 5%.

4.5.2 ROUGE Evaluation for Summaries with a ratio of 10%

We also have evaluated the summaries with a ratio of 10% to ensure the correlation of our approach summaries with reference summaries. Tables 4.9, 4.10, 4.11, 4.12, and 4.13 represent ROUGE evaluation metrics for both systems summaries with a ratio of 10%.

ROUGE	MEAD		RetMEAD	
Method	Case	Stemmed	Case	Stemmed
R-1	0.497	0.515	0.506	0.523
R-2	0.229	0.232	0.232	0.235
R-L	0.476	0.491	0.483	0.498
R-S*	0.232	0.250	0.242	0.261
R-S4	0.244	0.250	0.244	0.253
R-SU*	0.233	0.251	0.243	0.262
R-SU4	0.286	0.294	0.290	0.298

Table 4.5: ROUGE Evaluation using Isocolon with ratio of 5%

ROUGE	MEAD		RetMEAD	
Method	Case	Stemmed	Case	Stemmed
R-1	0.503	0.520	0.517	0.534
R-2	0.232	0.235	0.248	0.251
R-L	0.482	0.497	0.497	0.512
R-S*	0.241	0.259	0.256	0.274
R-S4	0.247	0.253	0.262	0.268
R-SU*	0.242	0.26	0.257	0.275
R-SU4	0.290	0.298	0.305	0.313

Table 4.6: The average ROUGE Evaluation for all rhetorical figures with ratio of 5%.

ROUGE	MEAD		RetMEAD	
Method	Case	Stemmed	Case	Stemmed
R-1	0.588	0.610	0.606	0.625
R-2	0.322	0.326	0.325	0.329
R-L	0.573	0.594	0.590	0.608
R-S*	0.337	0.363	0.358	0.385
R-S4	0.336	0.344	0.336	0.345
R-SU*	0.337	0.364	0.358	0.385
R-SU4	0.378	0.389	0.381	0.391

Table 4.7: ROUGE Evaluation using Antimetabole with a ratio of 10%.

ROUGE	MEAD		RetMEAD	
Method	Case	Stemmed	Case	Stemmed
R-1	0.604	0.624	0.605	0.625
R-2	0.345	0.349	0.344	0.349
R-L	0.588	0.607	0.588	0.608
R-S*	0.350	0.376	0.349	0.376
R-S4	0.359	0.367	0.358	0.366
R-SU*	0.350	0.377	0.350	0.376
R-SU4	0.400	0.410	0.399	0.410

Table 4.8: ROUGE Evaluation using Epanalepsis with a ratio of 10%.

In general, ROUGGE metrics improved significantly in summaries with a ratio of 10% over summaries with a ration of 5%. We observe that the correlation between reference summaries and system summaries with a ratio of 10% is higher than summaries with a ratio of 5%.

In Table 4.9, we notice that RetMEAD achieves better performance in the 10% summaries compared to the 5% summaries. We observe that RetMEAD using Antimetabole outperforms MEAD in all ROUGE methods.

Table 4.10 shows that the performance of RetMEAD improves and achieves slightly better scores than MEAD in two ROUGE methods that are ROUGE-1 and ROUGE-L. However, poor performance of RetMEAD is also noticed in all other ROUGE methods.

We observe in Table 4.11 that RetMEAD using Polyptoton performs remarkably better than MEAD in every single ROUGE method. We also notice that when Porter's Stemmer algorithm is incorporated, the results improve very well as well as RetMEAD still surpasses MEAD.

ROUGE	MEAD		RetMEAD	
Method	Case	Stemmed	Case	Stemmed
R-1	0.597	0.617	0.608	0.627
R-2	0.338	0.342	0.349	0.353
R-L	0.581	0.601	0.593	0.611
R-S*	0.346	0.372	0.356	0.381
R-S4	0.352	0.360	0.361	0.368
R-SU*	0.346	0.372	0.356	0.381
R-SU4	0.393	0.403	0.402	0.412

Table 4.9: ROUGE Evaluation using Polyptoton with a ratio of 10%.

ROUGE	MEAD		RetMEAD	
Method	Case	Stemmed	Case	Stemmed
R-1	0.597	0.616	0.607	0.627
R-2	0.335	0.339	0.336	0.341
R-L	0.581	0.600	0.590	0.608
R-S*	0.342	0.368	0.355	0.383
R-S4	0.350	0.358	0.347	0.355
R-SU*	0.343	0.368	0.355	0.383
R-SU4	0.391	0.401	0.391	0.400

Table 4.10: ROUGE Evaluation using Isocolon with a ratio of 10%

ROUGE	MEAD		RetMEAD	
Method	Case	Stemmed	Case	Stemmed
R-1	0.588	0.610	0.606	0.625
R-2	0.322	0.326	0.325	0.329
R-L	0.573	0.594	0.590	0.608
R-S*	0.337	0.363	0.358	0.385
R-S4	0.336	0.344	0.336	0.344
R-SU*	0.337	0.363	0.358	0.385
R-SU4	0.378	0.389	0.381	0.391

Table 4.11: The average ROUGE Evaluation for all rhetorical figures with ratio of 10%.

Table 4.12 indicates that RetMEAD using Isocolon performs well compared to MEAD in most ROUGE metrics. The results of MEAD, however, improve slightly in comparison to RetMEAD in two ROUGE metrics: ROUGE-S4 and ROUGE-SU4.

The average ROUGE evaluation score for all summaries using the four rhetorical figures is showed in Table 4.13. We observe that RetMEAD demonstrates better performance than MEAD in each ROUGE metric.

4.6 Discussion

In this chapter, we described the evaluation methods to evaluate our approach summaries using rhetorical figures and compared the results with the baseline. We used an intrinsic metric called ROUGE evaluation method. We evaluated summaries with different ratios, 5% and 10%. Our system RetMEAD using different rhetorical figures achieved better results in most cases. We noticed that RetMEAD using Epanalepsis achieved poor performance compared to the baseline MEAD. In all other figures, however, the results of RetMEAD increased.

We also observed that summaries with a ratio of 10% increase the scores in every ROUGE method, and thus, the correlation with reference summaries is improved significantly. Overall, RetMEAD outperformed MEAD in both summaries with different ratios.

The above results are preliminary and they showed a good direction for future work. We described some future works in Chapter 5 which will improve the performance of RetMEAD significantly.

Chapter 5

Conclusion and future work

5.1 Conclusion

In the past two decades, there has been increased interest in the area of text summarization to build automated text summarization systems able to produce accurate and adequate summaries of specific texts. Various automated text summarization systems have been developed based on different categorizes, such as abstractive or extractive summarization systems, and methods, such as centroid-based or maximum marginal relevance methods. However, these approaches do not consider deeper aspects of meaning including semantic content, emotional information, and author intention. Thus, the quality of the summaries that are produced by these systems is still inadequate. We believe that rhetoric can improve automated text summarization systems by providing the deeper aspects of meaning through rhetorical figuration metrics.

In this thesis, we presented a multi-document summarization system using rhetorical figuration information. Along with MEAD features [38], our approach system includes the use of rhetorical figuration information to produce a summary that possesses the most salient information from multiple articles.

We identified the role of rhetorical figures in a text. We also defined different categories of rhetorical figures. Since rhetorical figures provide clarity, emphasis, and emotional content, we use a tool called JANTOR to detect the rhetorical figures in a text [14].

We created a corpus of U.S. presidential speeches, which includes Campaign, State of the Union, and Inaugural speeches because these speeches use rhetorical figures extensively. Thus, these presidential speeches are meant to be used as a data set to test the performance of our multi-document summarization system approach using rhetorical figures and to compare our system against the baseline system, MEAD.

We then produced summaries in different ratios, such as 5%, and 10%, from the data set using both systems. Next, we performed an intrinsic evaluation to show the quality of our summarization system approach. Using ROUGE metrics, we found that our system RetMEAD outperforms MEAD in most cases. We observed that our system achieved better performance than MEAD when we used antimetabole, isocolon, and polyptoton. However, when we used Epanalepsis, RetMEAD shows poor performance in some cases and acceptable performance in other cases.

The results have supported our hypothesis and have ensured that a multi-document summarization system using rhetorical information should improve the produced summaries.

5.2 Future work

In our approach towards a -document summarization system, we have used only the information of four rhetorical figures. We tried to enrich the produced summaries to contain more salient information of texts. This approach towards the summarization system can be adapted to other domains that use rhetorical figures, such as classic news articles, public speeches, and persuasive articles. Thus, we would extend the list by adding new figures and comparing the results, which, in turn, should improve the quality of the summaries.

It is also possible to test more types of rhetorical figures and assign to them a weight based on their significance in a text. This requires the judgment of rhetoricians to determine the importance of a figure and to decide whether a figure is worthy to obtain higher weight or not based on its contribution to the content of a summary.

It is feasible to combine one or more figures with the MEAD features rather than using only one figure in the summarization process in order to achieve better performance.

Another possibility is to incorporate new features along with MEAD features in our summarization system to enrich the produced summaries such as a first sentence feature, which is based on the assumption that first sentences are highly relevant to the topic of a text(s).

It is possible to incorporate a learning algorithm to predicate the weight of each feature in MEAD system because there is no learning algorithm involved and the assumed weight was 1 to all MEAD features.

Potential work can be done towards manual summaries is to obtain more human summaries from 3-5 annotators for each speech. Then, to use them as a gold standard to have more accurate and precise evaluation results because the more human summaries included, the more reliable evaluation results will be.

There are some obstacles in the area of multi-document summarization system that remain challenging and complex chore. It is feasible to investigate these problems and attempt to overcome these issues by providing every possible solution. A potential problem is the coherent of produced summaries since our summarization system does not provide high linguistic quality summaries because we do not look at the readability of produced summaries. It is possible to investigate this issue and implement a method to coherently extract sentences from each cluster to create summaries by focusing on content quality.

Most current multi-document summarization systems are extractive summarization. However, many researchers in the area of summarization seek to develop abstractive summarization systems [15]. Since abstractive summarization systems require a deeper aspect of natural language understating, it is possible to develop a hybrid abstractive summarization by using textual entailment. Textual entailment assists in finding relations between text fragments, and it detects a concise shorter version of texts that involve holding meaning of the original texts [6].

Researchers in the text summarization community continue to improve the capabilities of the summarization system to deliver accurate and adequate summaries that hold the salient information in texts.

Appendix A

Recruitment Scripts and Feedback letter

David R. Cheriton School of Computer Science
University of Waterloo

PARTICIPANTS NEEDED FOR RESEARCH IN Text summarization in Computer science

We are looking for volunteers to take part in a study of a multi-document summarization system. .

As a participant in this study, you would be asked to: *rank the most representative sentences in each document (a presidential speech).*

Participants should have good knowledge of both English grammar and rhetoric.

Your participation would involve 1 session, of approximately *1 hour and half*.

In appreciation for your time, you will receive a *$10 Williams gift card or cash.*

For more information about this study, or to volunteer for this study, please contact:

Mohammed Alliheedi
David R. Cheriton School of Computer Science
519-888-4567 Ext.
Email: mallihee@uwaterloo.ca

This study has been reviewed by, and received ethics clearance through, the Office of Research Ethics, University of Waterloo.

Verbal and email script for student recruitment

Dear student,

We are looking for graduate students to partake in a research study that use rhetorical figuration metrics to improve the produced summaries from a multi-document summarization system.

In this study, you would be asked to review a presidential speech and extract a few sentences from the original article. This study will take approximately 1 hour and half, and you will receive a $10 Williams gift card or cash as remuneration for your participation in the study.

Participants should have good knowledge of both grammar and rhetoric.

If you are interested in participating please contact us by emailing Mohammed Alliheedi at mallihee@uwaterloo.ca . Any data pertaining to you as an individual participant will be kept confidential.

This study is supervised by Charlie Clarke (School of Computer Science). The study is being conducted for Mohammed Alliheedi's M.Math at School of Computer Science. This study was reviewed by, and received ethics clearance through, the Office of Research Ethics at the University of Waterloo.

Study Feedback

Date

Dear Participant,

We would like to take this opportunity to thank you for your participation in the research study. We want to specifically acknowledge your time and commitment to the study. It would not be possible to conduct this research without your participation.

Your participation has played a significant role in our research study the results of which will enrich text summarization systems. Through this study we will investigate the effectiveness of using rhetorical figuration metrics in multi-document summarization system. This research will contribute to the body of knowledge in the area of text summarization system.

All hard copies of consent forms and surveys will be stored under lock and key in the researcher's office. This will ensure that once collected, data with personal identifiers are securely stored in a locked area, and are accessible only to the research team.

An executive summary including the aggregated results of the study will be made available sometime in April 2012. We will send you a copy of this report via email.

This project has been reviewed by, and received ethics clearance through, the Office of Research Ethics. In the event you have any comments or concerns resulting from your participation in this study, please contact the Director at 519-888-4567, Ext. 36005.

Researcher Contact Information:

Mohammed Alliheedi, Department of Computer Science
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada N2L 3G1
mallihee@uwaterloo.ca

Appendix B
Participant Task

Participant Task

Name: ID:

Speech Title:

Please Read the following speech, and then answer these question carefully:

1. Select and highlight the sentences that convey the main topic of the speech(s). Please list these sentences in their original order.

2. You have to choose _____ (10%) sentences from the speech(s). Please list the number of sentences.

3. Choose the _____ (5%) most significant sentences from question no. 3.

References

[1] Barzilay, R., & Elhadad, M. (1997). Using lexical chains for text summarization. In Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization, 10-17.

[2] Baxendale, P. B. (1958). Machine-made index for technical literature: An experiment. IBM J.Res.Dev., 2(4), 354-361. doi:http://dx.doi.org/10.1147/rd.24.0354

[3] Brandow, R., Mitze, K., & Rau, L. F. (1995). Automatic condensation of electronic publications by sentence selection. Inf.Process.Manage., 31(5), 675-685. doi:10.1016/0306-4573(95)00052-I

[4] Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., & Hullender, G. (2005). Learning to rank using gradient descent. Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany. 89-96. doi:10.1145/1102351.1102363

[5] Burton, G.,Dr. (2007). Silva rhetoricae: The forest of rhetoric. Retrieved from http://humanities.byu.edu/rhetoric/Silva.htm

[6] Bysani, P. (2010). Detecting novelty in the context of progressive summarization. Proceedings of the NAACL HLT 2010 Student Research Workshop, Los Angeles, California. 13-18.

[7] Carbonell, J., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia. 335-336. doi:http://doi.acm.org/10.1145/290941.291025

[8] Conley, T. M. (1990). Rhetoric in the European tradition. New York: Longman.

[9] Conroy, J. M., & O'leary, D. P. (2001). Text summarization via hidden markov models. Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans, Louisiana, United States. 406-407. doi:http://doi.acm.org/10.1145/383952.384042

[10] Corbett, E. P. J. (1990). Classical rhetoric for the modern student.(3rd ed. New York: Oxford University Press)

[11] Cormen, T. H., Leiserson, C. E., & Rivest, R. L. (1992). Introduction to algorithms McGraw-Hill.

[12] Edmundson, H. P. (1969). New methods in automatic extracting. J.ACM, 16(2), 264-285.

[13] Evans, D. K., Mckeown, K., & Klavans, J. L. (2005). Similarity-based multilingual multi-document summarization. IEEE Transactions on Information Theory, 49

[14] Gawryjolek, J. J. (2009). Automated annotation and visualization of rhetorical figures. (Master of Mathematics, University of Waterloo). Retrieved from http://uwspace.uwaterloo.ca.proxy.lib.uwaterloo.ca/bitstream/10012/4426/1/Gawryjolek_Jakub.pdf

[15] Genest, P. E., & Lapalme, G. (2010). Text generation for abstractive summarization. Proceedings of the Third Text Analysis Conference, Gaithersburg, Maryland, USA. National Institute of Standards and Technology

[16] Gupta, V., & Lehal, G. (2010). A survey of text summarization extractive techniques. Journal of Emerging Technologies in Web Intelligence, 2(3)

[17] Hovy, E., & Lin, C. (1999). Automated text summarization in SUMMARIST. advances in automatic text summarization, 81-94 MIT Press.

[18] Jones, K. S., & others. (1999). Automatic summarizing: Factors and directions. Advances in Automatic Text Summarization, , 1-12.

[19] Kupiec, J., Pedersen, J., & Chen, F. (1995). A trainable document summarizer. Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, United States. 68-73. doi:http://doi.acm.org/10.1145/215206.215333

[20] Lauer, J. M., & Pender, K. (2003). Invention in rhetoric and composition Parlor Press.

[21] Lausberg, H., Orton, D. E., & Anderson, R. D. (1998). Handbook of literary rhetoric: A foundation for literary study Brill Academic Pub.

[22] Lin, C. Y. (2004). Rouge: A package for automatic evaluation of summaries. Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), 16.

[23] Lin, C., & Hovy, E. (1997). Identifying topics by position. Proceedings of the Fifth Conference on Applied Natural Language Processing, Washington, DC. 283-290. doi:http://dx.doi.org/10.3115/974557.974599

[24] Lin, C., & Hovy, E. (2002). Automated multi-document summarization in NeATS. Proceedings of the Second International Conference on Human Language Technology Research, San Diego, California. 59-62.

[25] Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM J.Res.Dev., 2(2), 159-165. doi:http://dx.doi.org/10.1147/rd.22.0159

[26] Mani, I. (2001). Automatic summarization J. Benjamins Publishing Company.

[27] Mani, I. (1999). Advances in automatic text summarization. In M. T. Maybury (Ed.), (pp. 1-2). Cambridge, MA, USA: MIT Press.

[28] McKeown, K., & Radev, D. R. (1995). Generating summaries of multiple news articles. Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, United States. 74-82. doi:http://doi.acm.org/10.1145/215206.215334

[29] McQuarrie, E. F., & Mick, D. G. (1996). Figures of rhetoric in advertising language. Journal of Consumer Research, 22(4), 424-438.

[30] Miller, G. A. (1995). WordNet: A lexical database for English. Commun.ACM, 38(11), 39-41. doi:http://doi.acm.org/10.1145/219717.219748

[31] Morris, J., & Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Comput.Linguist., 17(1), 21-48.

[32] Nahnsen, T., Uzuner, O., & Katz, B. (2005). Lexical chains and sliding locality windows in content-based text similarity detection. Companion Volume to the Proceedings of Conference Including Posters/Demos and Tutorial Abstracts, 150-154}.

[33] Nenkova, A., & Vanderwende, L. (2005). The impact of frequency on summarization. Microsoft Research, Redmond, Washington, Tech.Rep.MSR-TR-2005-101

[34] Paice, C. D. (1990). Constructing literature abstracts by computer: Techniques and prospects. Inf.Process.Manage., 26(1), 171-186. doi:10.1016/0306-4573(90)90014-S

[35] Porter, M. F. (1997). An algorithm for suffix stripping. In (pp. 313-316). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

[36] Radev, D., Allison, T., Goldensohn, B. S., Blitzer, J., Celebi, A., Dimitrov, S., Liu, D. (2004). MEAD-a platform for multidocument multilingual text summarization. LREC, , 2004 1-4.

[37] Radev, D. R., Hovy, E., & McKeown, K. (2002). Introduction to the special issue on summarization. Comput.Linguist., 28(4), 399-408. doi:http://dx.doi.org/10.1162/089120102762671927

[38] Radev, D. R., Jing, H., Stys, M., & Tam, D. (2004). Centroid-based summarization of multiple documents. Information Processing & Management, 40(6), 919-938.

[39] Ross, W. D., & Roberts, W. R. (2010). Rhetoric Cosimo, Incorporated.

[40] Saggion, H., Teufel, S., Radev, D., & Lam, W. (2002). Meta-evaluation of summaries in a cross-lingual environment using content-based metrics. Proceedings of the 19th International Conference on Computational Linguistics - Volume 1, Taipei, Taiwan. 1-7. doi:10.3115/1072228.1072301

[41] Silber, H. G., & McCoy, K. F. (2002). Efficiently computed lexical chains as an intermediate representation for automatic text summarization. Comput.Linguist., 28(4), 487-496. doi:http://dx.doi.org/10.1162/089120102762671954

[42] Singer, H., & Donlan, D. (1985). Reading and learning from text Lawrence Erlbaum Associates, Inc., Publishers, 365 Broadway, Hillsdale, NJ 07642.

[43] Strzalkowski, T., Wang, J., & Wise, B. (1998). A robust practical text summarization. Proceedings of the AAAI Symposium on Intelligent Text Summarization, 26-33.

[44] Svore, K., Vanderwende, L., & Burges, C. (2007). Enhancing single-document summarization by combining RankNet and third-party sources. Proceedings of EMNLP-CoNLL, 448-457.

[45] Teufel, S., & Moens, M. (1997). Sentence extraction as a classification task. Proceedings of the ACL, 97 58-65.

[46] Teufel, S., & Moens, M. (1999). Argumentative classification of extracted sentences as a first step towards flexible abstracting. ADVANCES IN AUTOMATIC TEXT SUMMARIZATION, 155-171.

[47] Verma, R., Chen, P., & Lu, W. (2007). A semantic free-text summarization using ontology knowledge. En Proceedings of the Document Understanding Conference.,

[48] Watson, J. S. (1891). Quintilian's institutes of oratory: Or, education of an orator Bell.

[49] Zhang, J., Sun, L., & Zhou, Q. (2005). A cue-based hub-authority approach for multi-document text summarization. Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE '05. Proceedings of 2005 IEEE International Conference on,642. doi:10.1109/NLPKE.2005.1598815