Genre Oriented Summarization

Jade Goldstein Stewart

December 2008

CMU-LTI-09-001

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213
http://www.lti.cs.cmu.edu

Thesis Committee:
Jaime Carbonell, Chair
Jamie Callan
Vibhu Mittal, Google
John Conroy, IDA/Center for Computing Sciences

Submitted in partial fulfillment of the requirements
for the degree Doctor of Philosophy

To Charles

Acknowledgements

There are a great many people who have my sincere gratitude for the assistance that they gave me during the course of my work on my dissertation. I would especially like to thank my advisor Jaime Carbonell for his support and encouragement over the years

I am also grateful to the other members of my committee, Jamie Callan, Vibhu Mittal and John Conroy for their insightful comments, discussions and advice. Many thanks also go to the late Barbara Lazarus, who provided much encouragement during this effort, which was always deeply appreciated.

My thanks also goes to my colleagues, friends and family members, who have provided much support and assistance during the process – especially (but not limited to) Roberta Sabin, Maria Alvarez-Ryan, Barb Wheatley, Mark Kantrowitz, Tina Kohler, Kathy Baker, Lisa Harper, Jon Nedel and Chad Langley.

Thanks also to the numerous people (not listed here) who assisted with the annotation of the multiple data sets and judgments of summary quality.

I owe a great deal to my husband Charles, who has supported and assisted greatly in this endeavor. Thanks also to my parents who fostered an atmosphere of experimentation and exploration in my childhood.

Abstract

Summaries are used in daily life to condense information in a manner suitable for the intended recipient's use and ideally suit the recipient's information seeking goals. In the case of text, examples of such summaries include newswire articles, headlines, and the information snippets returned by Google. Previous research has focused on summarizing newswire articles or clusters of newswire documents, scientific articles, books, and extracting opinion (sentiment) sentences from reviews.

Our research addresses how to create short multi-sentence summaries meeting user's goals within specific genres. The methodology is first to determine the genre of the document and then, based on the genre, present applicable summaries designed to address the user's information seeking goals. For example, in the movie review genre, a goal-focused summary for a review could be an overview summary, a plot summary, or an opinion summary, each of which has a different focus and hence a different summary composition.

We describe our experiments with genre identification using different sets of features and varying numbers of training documents and show that genre tagging using classifiers (Support Vector Machines and Random Forests) is probably at a sufficient level of accuracy to inform a summarization system. We discuss the creation of goal focused single document summaries for seven genres (newswire articles, editorials, interviews, biographies, movie reviews, product reviews, and product press releases). Our results indicate that genre oriented goal-focused summarization algorithms perform better than our two baselines, lead sentence and the newswire summarization algorithm.

We also examine email summarization and, based on previous research in speech acts, present categories of the communicative intent of the sender. We discuss our experiments in identifying these email speech acts using a small annotated corpus of personal emails. In addition, for textual summaries of a sender's email, we analyze a human annotated subset of the Enron corpus and based on a user study, suggest that the subject line and one sentence extracted from the email text body may be an effective summary length.

We briefly explore multi-document summarization for the newswire genre and present results indicating that, by using maximal marginal relevance (MMR) to eliminate redundancy, there is more coverage of the subtopics in a cluster than our baseline - which uses our single document newswire summarization algorithm on the concatenation of all articles and no MMR. MMR based summaries were also preferred in a ranking produced by one unbiased human evaluator.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

> **"We are drowning in information, while starving for wisdom. The world henceforth will be run by synthesizers, people able to put together the right information at the right time, think critically about it, and make important choices wisely."**
>
> E.O. Wilson: Consilience: The Unity of Knowledge

Summaries are a method of conveying a condensed version of information to people. They can take various forms - text, graphics, video, speech or a combination of these forms. Effective summaries are tailored to the user's information seeking goals; they convey the significant facts of the original text, condense descriptive details and are dependent on the genre.

In this thesis, we examine the creation of textual summaries that take into account the user's information seeking goals as well as the genre – what we call "*genre oriented goal focused summaries"*. We explore the creation of several of these genre oriented summaries, including those for articles, reviews, press releases, email, interviews and biographies. The focus is on single document summarization with a brief analysis of multi-document summarization in the newswire domain.

1.1 Motivation

Why summarize and why focus on summaries that address user's information seeking goals?

Suppose one was busy at the time of a historic event. Let us illustrate by using an example of China's first astronaut to walk in space. If one wanted to know the details of such an event, one might search on Google or some other search engine. A first pass query might be something like "China man walk space".

For this query, Google returns the following five top ranked items, shown in Figure 1-1. The ranked items are snippets, a type of *indicative* summary, that are presented by Google to give the user an indication of the relevance of the link. They contain a link, often a date as well as a small snippet of text and the web page source.

The top results of this search appear to be news articles, which are inherently summaries of an event. We would assume that the first ranked Google retrieved result would be the most important. However, when one selects this link, it appears to no longer exist and instead a related article (Google retrieved result rank 2) appears. If we treat the first five sentences of this article as a summary of the newswire article, utilizing the principle that important items are presented first to the reader, then the following lead sentence summary for this Daily Mail article is depicted in Figure 1-2:

1. Out of this world: First Chinese **man walks** in **space** | Mail Online
   - 9:48am
   Sep 28, 2008 **...** Astronaut Zhai Zhigang became the first Chinese **man** to **walk** in **space** today, clambering out of **China's** Shenzhou VII **space** craft and waving to **...**
   *www.dailymail.co.uk/sciencetech/article-1063194/Out-world-First-Chinese-**man**-**walks**-**space**.html* - Similar pages - Note this

2. Jubilation as first Chinese **man walks** in **space** - but misery for **...**
   Sep 29, 2008 **... China** was celebrating yesterday after astronaut Zhai Zhigang became the first Chinese **man** to **walk** in **space**.
   *www.dailymail.co.uk/.../Jubilation-Chinese-**man**-**walks**-**space**--misery-U-S-NASAs-50th-anniversary-mired-funding-crisis-fears...* - Similar pages - Note this
3. Scoop: First Chinese **man** to **walk** in **space**
   Sep 28, 2008 **...** Astronaut Zhai Zhigang has become the first Chinese **man** to **walk** in **space**.
   www.scoop.co.nz/multimedia/tv/world/13560.html - 52k - Cached - Similar pages - Note this
4. 3 News > Video > World > First Chinese **man** to **walk** in **space**
   - 9:45am
   Sep 28, 2008 **...** First Chinese **man** to **walk** in **space** Astronaut Zhai Zhigang has become the first Chinese **man** to **walk** in **space**. 3news.co.nz - information on **...**
   *www.3news.co.nz/Video/FirstChinese**man**to**walk**in**space**/tabid/313/articleID/73476/Default.aspx - 80k* - Cached - Similar pages - Note this
5. The Telegraph - Calcutta (Kolkata) | International | **China** in **...**
   Sep 28, 2008 **...** 27 (Reuters): Astronaut Zhai Zhigang became the first Chinese **man** to **walk** in **space** today, clambering out of **China**'s Shenzhou VII **space** craft **...**
   www.telegraphindia.com/1080928/jsp/foreign/story_9898741.jsp - 28k - Cached - Similar pages - Note this

Figure 1-1: Top five Google results for "China man walk space"

Jubilation as first Chinese man walks in space - but misery for U.S. as NASA's 50th anniversary mired by funding crisis fears

By Barry Wigmore and Daily Mail Reporter
Last updated at 2:33 AM on 29th September 2008

[1] China was celebrating yesterday after astronaut Zhai Zhigang became the first Chinese man to walk in space.

[2] But in a week where China blasted men into orbit - and launched an ambitious project that could see them on the moon by 2017 - American space scientists were fearing they could be left behind in the space race with the news NASA is being hit hard by the credit crunch.

[3] The U.S. space agency, which celebrated its 50th anniversary this week, is short of funds and, with little hope of more government money, experts warn it will be stuck in a five-year time warp without a spaceship to take astronauts into space.

[4] The ageing, accident prone space shuttles are to be retired in 2010 - but the next U.S. space vehicle, a traditional rocket called Orion, will not be ready until 2015.

[5] Meanwhile there were happy scenes in China yesterday as Zhai clambered out of China's Shenzhou VII space craft and waved to the camera.

Figure 1-2: Daily Mail - First five sentence. Google retrieved result rank 2.

If we analyze the summary in Figure 1-2, we notice that half of the summary, i.e, half of sentence 2, as well as sentences 3 and 4 are not about the Chinese man walking in space, but about the American space program.

Google retrieved result rank 3 only contains the sentence "Astronaut Zhai Zhigang has become the first Chinese man to walk in space." There is a video on the page.

Google retrieved result rank 4 is from 3News (Figure 1-3). Compared to the Daily Mail summary in Figure 1-2, the 3News summary contains more information. In this summary, we learn from Sentence 2 that the astronaut Zhai Zhigang is a 41-year-old jet pilot and that he completed a successful 15-minute EVA. We learn that this is China's third manned space journey from Sentence 2 and that China is the third country to have sent people into space (Sentence 5).

Google retrieved result rank 5 is from The Telegraph Calcutta based on Reuters. The first five sentences are shown in Figure 1-4.

If we perform a possibly more sophisticated search by including the word "astronaut" instead of man[1], we obtain a Reuters (reuters.com) article as the first Google ranked item. The first five sentences are shown in Figure 1-5. The Calcutta Reuters summary (Figure 1-4) is contained in the reuters.com article summary (Figure 1-5), which merges two sentences allowing for an additional sentence. Both of these summaries contain more information content than the Daily Mail summary in Figure 1-2. They contain similar content to the 3News summary (Figure 1-3), but also differ in that they contain more details of the event, whereas the 3News summary relates this event to the accomplishments of other countries.

First Chinese man to walk in space

Sun, 28 Sep 2008 7:13a.m.

[1] Astronaut Zhai Zhigang has become the first Chinese man to walk in space.
[2] The 41-year-old jet pilot completed a successful 15 minute EVA from the Shenzhou VII space craft earlier tonight.
[3] The walk marked the highpoint of China's third manned space journey, which has received blanket media coverage in the Peoples Republic.
[4] The fast-growing Asian power wants to be sure of a say in how space and its potential resources are used.
[5] The only other countries that have sent people into space are Russia and the United States.

Figure 1-3: 3News - First five sentences (which comprise the entire article) of Google retrieved result rank 4

---

[1] China refers to astronauts as "yuhangyuan", the translation of which is "taikonaut".

China in space walk club

[1] Beijing, Sept. 27 (Reuters): Astronaut Zhai Zhigang became the first Chinese man to walk in space today, clambering out of China's Shenzhou VII space craft in a technological feat that Beijing wants the world to marvel about.
[2] "I'm feeling quite well. I greet the Chinese people and the people of the world," Zhai said as he climbed out of the craft, his historic achievement carried live on state television.
[3] Zhai, the 41-year-old son of a snack-seller chosen for the first "extra-vehicular activity," unveiled a small Chinese flag, helped by colleague Liu Boming, who also briefly popped his head out of the capsule.
[4] Zhai safely returned inside the craft after about 15 minutes.
[5] The walk marked the high point of China's third manned space journey.

Figure 1-4: The Telegraph Calcutta, India - First five sentences of Google retrieved results rank 5

Chinese astronaut takes historic walk in space

Sat Sep 27, 2008 11:37am EDT
By Ben Blanchard
[1] BEIJING (Reuters) - Astronaut Zhai Zhigang became the first Chinese man to walk in space on Saturday, clambering out of China's Shenzhou VII spacecraft in a technological feat that Beijing wants the world to marvel at.
[2] "I'm feeling quite well. I greet the Chinese people and the people of the world," Zhai said as he climbed out of the craft, his historic achievement carried live on state television.
[3] Zhai, the 41-year-old son of a snack-seller, unveiled a small Chinese flag, helped by colleague Liu Boming, who also briefly popped his head out of the capsule.
[4] Zhai re-entered the spacecraft safely after a walk of about 15 minutes, marking the high point of China's third manned space flight, which has received blanket media coverage.
[5] He wore a $4.4 million Chinese-made suit weighing 120 kg (265lb).

Figure 1-5: Reuters article - First five sentences (reuters.com)

Chinese astronaut walks in space

Page last updated at 09:21 GMT, Saturday, 27 September 2008 10:21 UK
[1] A Chinese astronaut has become the first in his country's history to take a walk in space.
[2] In an operation broadcast live on national TV, fighter pilot Zhai Zhigang emerged from the capsule orbiting the Earth to wave a Chinese flag.
[3] Mr Zhai, 42, stayed outside the capsule for 15 minutes while his two fellow astronauts stayed in the spacecraft.
[4] The exercise is seen as key to China's ambition to build an orbiting station in the next few years.
[5] Mr Zhai began the manoeuvre just after 1630 Beijing Time (0830 GMT) on Saturday, and completed it about 15 minutes later.
[6] "I'm feeling quite well. I greet the Chinese people and the people of the world," he said as he climbed out of the Shenzhou VII capsule.
[7] Mr Zhai wore a Chinese-made spacesuit thought to have cost between £5m and £20m ($10m-$40m) for the space walk.

Figure 1-6: BBC article - First seven sentences

The second ranked article from this new search is from the BBC (Figure 1-6). We display two extra sentences to show the similarity to the reuters.com article. However there are some notable discrepancies in the reports such as the cost of the space suit ($10 million vs. $4.4 million) as well the age (42 vs. 41) and previous occupation (fighter pilot vs. jet pilot) of the astronaut

If one is aware that the Chinese people refer to an astronaut as a "taikonaut" and performs the search 'China taikonaut space walk", the first ranked article is from a Chinese source – the English version of the People's Daily Online. The five sentence lead summary has some constructs that clearly show it was not written by a native speaker – such as the phrase "in the outer space". Note that if one didn't know that an astronaut was a taikonaut, it would have been difficult to retrieve this article.

Chinese taikonaut debuts spacewalk

16:48, September 27, 2008

[1] Chinese taikonaut Zhai Zhigang slipped out of the orbital module of Shenzhou-7 Saturday afternoon, starting China's first spacewalk or extravehicular activity (EVA) in the outer space.
[2] Donning a 4-million-U.S.dollar homemade Feitian space suit, Zhai waved to the camera mounted on the service module after pulling himself out of the capsule in a head-out-first position at 4:43 p.m. (0843 GMT), video monitor at the Beijing Aerospace Control Center (BACC) showed.
[3] "Shenzhou-7 is now outside the spacecraft.
[4] I feel well.
[5] I am here greeting the Chinese people and people of the whole world," the 42-year-old taikonaut reported to the ground control in Beijing.

Figure 1-7: People's Daily Online - First five sentences

From these six example lead sentence summaries (Figure 1-2 - Figure 1-7), all from news sources, we can see the variety of information that can be presented.

These sources show the challenges of summarization – first, that human produced news articles can vary tremendously in terms of the information that they prioritize through placement at the beginning of the news article. From our perspective it is important for summaries to address the user's information seeking goals, but these goals in themselves often present a problem in that they are difficult to define as they vary by reader just as each news article can differ in its focus.

In the example, that we have just illustrated, all the summaries provided the basic information from the query "China man/astronaut/taikonaut walk space/space walk." The rest of the articles/summaries highlighted different information in the first five sentences. Which of these pieces of information are the most important to the user, for example, future space plans, the fact that there were two fellow astronauts and/or what Mr. Zhai did during his spacewalk? How does the user know what to search for if s/he does not know the details of the event or what constitutes a good summary for his/her

information seeking goals? Furthermore, if a user does want the event details, s/he cannot just search on the phrase "event details" to retrieve such information.

Accordingly, we ask, "what are the key points a summary should include?" This is a question that we address in this thesis, by developing genre oriented summarization, which seeks to generate summaries based on the document genre and what we can determine about a user's information seeking goals.

Let us now examine these six articles for the purpose of multi-document newswire summarization. As is typical when performing multi-document newswire summarization, articles can be very similar such as in the case of the two Reuters articles. Articles are frequently updated as news stories unfold and slight edits can be made resulting in new versions. Alternately, newspapers can subscribe to news feeds and alter the articles to suit their purposes either for stylistic reasons or because of article space constraints.

If we employ a simple multi-document summarization algorithm that attempts to take one highlight of each article, one method might be to take the first sentence of each article since the first sentence often reflects the most important point of the article. Using this algorithm for the six lead sentence single document summaries would result in the multi-document summary shown in Figure 1-8.

---

[1] China was celebrating yesterday after astronaut Zhai Zhigang became the first Chinese man to walk in space.
[2] Astronaut Zhai Zhigang has become the first Chinese man to walk in space.
[3] Beijing, Sept. 27 (Reuters): Astronaut Zhai Zhigang became the first Chinese man to walk in space today, clambering out of China's Shenzhou VII space craft in a technological feat that Beijing wants the world to marvel about.
[4] BEIJING (Reuters) - Astronaut Zhai Zhigang became the first Chinese man to walk in space on Saturday, clambering out of China's Shenzhou VII spacecraft in a technological feat that Beijing wants the world to marvel at.
[5] A Chinese astronaut has become the first in his country's history to take a walk in space.
[6] Chinese taikonaut Zhai Zhigang slipped out of the orbital module of Shenzhou-7 Saturday afternoon, starting China's first spacewalk or extravehicular activity (EVA) in the outer space.

Figure 1-8: Multi-document summary for "China man/astronaut space walk" consisting of first sentence from five articles.

This multi-document summary is not a very useful summary. Each sentence repeats the fact that Zhai Zhigang is the first Chinese man to walk in space. There are only three other facts in the summary (1) that he climbed out of a Shenzhou VII spacecraft (repeated thrice), (2) the event was on Saturday September 27th (repeated thrice) and (3) that Beijing wants the world to marvel at this feat (repeated twice). Two of the sentences (3 and 4) are almost identical with the exception of a few words.

Clearly, this multi-document summary is highly redundant and repetitive, does not make effective use of the five sentences available for presenting information and does not

develop any themes as do the single document summaries (Figure 1-2 - Figure 1-7). In short, this is not a good multi-document summary. Each of the single document summaries with the exception of the summary from the Daily Mail which discusses the U.S. space woes, would have made a good multi-document summary for this event.

In the case of this historic spacewalk event, which does not unfold over a lengthy period of time, a single document summary might suffice for a multi-document summary. In the case of an event that spans a moderate or long duration time period, the summarizer must be careful about the information that is chosen. For example in the reports about the death tolls due to Hurricane Ike:

- September 16th, 1st sentence, U.S. & World: "Houston (AP) – The death toll from Hurricane Ike has risen to 48."
- September 16th, 2nd sentence, U.S. & World: "Authorities said Tuesday six more deaths are being blamed on the storm in the Houston area,' bringing the number of people killed in Texas to 17."
- October 2nd, 1st sentence, kwtx.com: "(October 2, 2008) – The death toll in Texas from Hurricane Ike rose to at least 33 Thursday after the Harris County Medical Examiner's Office reported the death of a 47-year old Houston man."
- October 2nd, 4th sentence, The Associated Press: "Batiste's death brings the total of deaths nationwide from Hurricane Ike to 68."
- October 2nd, 6th sentence, The Associated Press: "The September 13th storm killed at least 80 people in the Caribbean before pummeling the Gulf Course.""
- October 4th, 4th sentence, Houston Chronicle: "This raises the hurricane's Houston-area death toll to 35.""

The sentences about Hurricane Ike show that there are three death's toll being counted over time– one for Texas, one nationwide and one in the Caribbean. A sophisticated summarizer must be able to track the deaths in the different regions as well as recognize that the "September 13th storm" referred to in the 6th sentence from the Associated Press is referring to Hurricane Ike.

These examples, for both the single document cases and the multi-document cases, show that in any methodology for summarization it is very important to create effective summaries that address user's information seeking goals.

We believe that development of effective summarization techniques are important with the continuing rapid expansion of online textual information on the World Wide Web as well as internal documentation and databases. The amount of electronic textual data has necessitated the importance of (1) providing improved mechanisms to assist the user in finding the appropriate information for their information seeking goals and (2) useful techniques to present such information.

Addressing these issues may involve a two step process: (1) finding the relevant documents based on maximizing relevance to a user query as performed by conventional information retrieval systems such as modern search engines and (2) finding the relevant information within the documents. In order to satisfy the second step, there has been an increasing focus on summarization systems which aid the user to rapidly scan and identify important documents as well as locate the relevant sections of documents to

retrieve specific answers to their information seeking goals [Tombros 1998, Mani 1998, DUC, NTCIR, TREC-QA, TAC, GALE, ACL Workshops].

Automated document summarization dates back at least to Luhn's work at IBM in the fifties [Luhn 1958]. Several researchers continued investigating various approaches to this problem through the seventies and eighties [Tait 1983, Paice 1990]. The resources devoted to addressing this problem grew by several orders of magnitude with the advent of the world-wide web and large scale search engines and hence the necessity to rapidly scan information to determine relevance. The initial focus was on single document summarization. Several innovative approaches began to be explored: linguistic approaches [Breck and Morton 1998, Marcu 1997, Boguraev and Kennedy 1997, Klavans and Shaw 1995, McKeown et. al. 1995, Radev and McKeown 1998, Aone et. al. 1997], statistical and information-centric approaches [Kupiec et. al. 1995, Hovy and Lin 1997, Mitra et. al. 1997, Strzalkowski et. al. 1998, Goldstein et. al. 1999], and combinations of the two [Teufel and Moens 1997, Barzilay and Elhadad 1997, Strzalkowski et. al. 1998]. The TIPSTER Phase III Program, an information retrieval initiative of the US Defense Department funded several of these projects on summarization [Mani et. al. 1998] which culminated in a formal evaluation, SUMMAC. In 2000, NIST started formal evaluation on single document and later multi-document summarization though the Document Understanding Conferences [DUC].

Summaries should provide at a minimum, an indication of the information content of a document so that a user can choose whether or not to read it, these are referred to as "*generic"* summaries. A more effective "ideal" summary will contain the content for which the user is searching – "*query relevant"* - if the user constructs a query. A "*goal focused"* summary addresses a specific user information seeking goal. Summarization systems have focused on the first two types of summaries: the *generic* summary, which gives an overall sense of the document's content (appropriate in the case where a user is browsing), or a *query-relevant* or *topic-based* summary, which presents the content that is most closely related to the initial search query or topic description.

Extrinsic evaluations [Sparck-Jones and Galliers 1996], which evaluate how well systems perform in a given task have shown that summaries can assist users in selected applications [Mani et. al. 1998]. *Intrinsic* evaluations which evaluate internal quality have shown that summarization systems can extract relevant content of the main text [DUC, NTCIR, MSE] that matches human gold standard summaries. However, these summaries have not been evaluated extrinsically in terms of their actual utility in some task. Furthermore, the focus of the intrinsic evaluations has been on the content that is presented and in later evaluations how responsive they are to the actual topic. Summaries have not been judged on their internal fluency, coherency, theme development, or successfully conveying the original author's intention.

Part of this thesis is to show that as in information retrieval systems, single document summarization systems can also extract much of the relevant sections of the document as well as the important sections of the document in genres besides newswire. In addition,

we will demonstrate that summaries should be different based on the genre and the users' information seeking goals.

In the late 1990s and early 2000s, single and multi-document summarization evaluations have primarily focused on news articles with the exception of the scientific articles [Teufel and Moens 2002] and web pages [Amitay and Paris 2000, Sun et. al 2005]. In the past few years, single document summarization has started to be applied to other genres, such as product reviews [Hu and Liu 2004, Popescu and Etzoni 2005, Jindal and Liu 2006, Liu et. al. 2007], movie reviews [Zhuang et. al 2006] and book summaries [Kazantseva and Szpakowicz 2006, Mihalccea and Ceylan 2007]. With the exception of the book summaries, the review genres have focused on either extracting the opinion sentences or the overall sentiment of the review, both the sentiment and the quality of the review [Liu et. al. 2007], the product features and the polarity and strength of opinions [Popescu and Etzoni 2005].

Our research is focused on summarization of various genres including movie review, product reviews, press releases and others [Goldstein et. al. 2007] and in contrast to these efforts have focused on what the composition of a summary should be and then attempted to extract appropriate information. In fact, many of the review oriented systems mentioned above could be used to identify information that could then be used to create a useful summary.

There has been minimal efforts on summarization for other tasks, such as extracting an author's recommendation (e.g., in a review article) with the exception of extracting opinions – sentiment analysis, besides the effort mentioned above, [Hu and Liu 2004, Wilson et. al, 2005, Titov and McDonald 2008]. There has only been little effort on adapting the summarization process to the genre of the document. Most summarization research has focused on the news events and scientific articles genres [Mani et, al. 98, Teufel and Moens 97] for which query-relevant summaries work well. Consider however, the genre of movie reviews. A user may want an overall sense of the review (generic), a specific answer to question (query-based), details of the plot, opinions on the quality of the movie and/or acting, or details on what audience for which it is suitable and the reasons. Accordingly, this genre as do others, require a new class of summary - that of the *goal-focused* summary. Part of this thesis discusses the novel technique of first performing genre identification and then using the genre to inform summarization. The utility and capabilities or goal-focused genre summarization is described and how such summarization better addresses a user's information seeking needs.

Consider another important summarization situation, that in which the user issues a search query, for instance on a news topic, and the retrieval system finds hundreds of closely-ranked documents in response. The choice of which information to select for the user presents its own set of challenges as many of these documents are likely to repeat much the same information, while differing in certain parts. Generating summaries of the individual documents would help, but are likely to be very similar to each other, unless the summarization system takes into account the other documents or other summaries that have already been generated. Multi-document summarization -- capable of summarizing

either complete documents sets, or clusters of documents in the context of previously summarized ones, referred to as *update summaries* -- is likely to be essential in such situations. Ideally, multi-document summaries should reflect the underlying information content in the overall document set while minimizing redundant information. Information repeated in most of the documents is likely to be more important than information mentioned in only one document; while it is essential that this information be conveyed in the summary, it is also essential to balance the need for emphasis (repetition) with the need for maximizing query-relevant information conveyed in the summary.

Given the enormous amounts of information that is accessible, good quality multi-document summaries are needed to save the user from the time consuming task of browsing large quantities of relevant documents. Multi-document summaries can be used for a variety of purposes including: (1) to locate the sections of text pertinent to a users' information seeking goals, such as browsing or finding specific answers to questions, (2) to indicate the content of a document collection, or (3) to provide updates to ``known'' information (a particular summary or stored representation of what an user has previously seen).

At a first pass, one might use a *generic* (overview) single document summarization system for both genre oriented summarization by ignoring the genre, and for multi-document summarization, by concatenating all the documents together and treating them as a single document. For genre oriented summarization, this methodology would not work so well. Consider for example, the movie review genre, one might want to extract rating information from the review - a genre specific item that a generic single document summarization system would not tend to include. Another example situation would be that of interviews - the reader is probably not interested in a sentence from an interviewer's question without the interviewee's answer. Thus a genre oriented summarizer must first have a component to classify to which type of genre the document belongs and then the summarizer must utilize that information to tailor the information extracted to suit the desired output summary.

In the case of multi-document summarization (MDS), at one end of the summarization spectrum, this can be considered to be just simply single-document summarization, with all the documents concatenated together. However, this view of MDS is potentially sub-optimal, since in the worst case, it can make incorrect assumptions (the same referring expression, such as “the president”, in two different documents need not necessarily refer to the same object/person), and in the best case, it ignores several additional, and possibly useful pieces of information, such as the temporal ordering of the documents, knowledge about the authorship or the source of the documents/passages, etc. Because of these facts, there are at least six issues that need to be emphasized in the design of a MDS system as compared to the design of a single document summarization system:

1. The degree of redundancy in information contained in a group of topically-related articles is much higher than the degree of redundancy within an article, as each article is apt to describe the main point as well as necessary shared background. Hence *anti-redundancy* methods are more crucial.

2. A group of articles may contain a *temporal dimension*, typical in a stream of news reports about an unfolding event. In this case later information may override earlier more tentative or incomplete accounts.
3. The *compression ratio* (i.e. the size of the summary with respect to the size of the document set) will typically be much smaller for collections of dozens or hundreds of topically related documents than for single document summaries. The SUMMAC evaluation [Mani et. al 1998] tested 10% compression summaries, but for multi-document summarization of 200-document clusters, compression in the 1% or 0.1% level is required. Summarization becomes significantly more difficult when compression demands increase. Furthermore, given such small compressions, users will need to be able to vary the summary size of the collection, access the context around a text span in the summary, and retrieve original document(s) for further information. Thus a user interface which supports these functions is a key component.
4. The *co-reference* problem in summarization presents even greater challenges for multi-document than for single-document summarization [Breck and Morton 1998]. For example, a multi document summary may contain text spans from several documents. If a text span in the summary contains a pronoun and not its preceding referent, it is most likely to given a false indication (association with a prior referent to which it is not linked) or leave the reader with insufficient information to understand the text span.
5. The *coherence* situation in multi-document summarization can cause great difficulty since the selection of articles from which information is selected can represent a variety of "subtopics' within the main multi-document cluster, whereas a single document tends to focus on one or a few subtopics. The summarization system must build and present a summary that is both coherent to a reader and appropriately focused.
6. *user interface and information presentation*: will need to address the users' information seeking goals by allowing rapid effective interaction with the summary for purposes such as viewing the context of a passage within the summary, viewing related information to the summary passages including the original document and/or single document summaries, and creating new related summaries.

Most summarization systems have been addressing the redundancy issues, many using variants of our method of eliminating redundancy, the maximal marginal relevance, presented in Chapter 3 [Carbonell and Goldstein 1998, Ye 2005]. Another issue that is being addressed is the ordering of sentences. Early work on multi-document summarization [McKeown et. al 1999, Goldstein et. al 2000a] proposed several approaches. McKeown suggested sentence ordering by fluency and coherence, where we presented four different methods in which sentences from our multiple newswire documents could be summarized – document ordering (highest ranking document first), news-story principle (rank ordering of sentences and/or information), topic cohesion and time-line or chronological ordering. Informal human studies that we performed showed that people preferred the chronological ordering. Recent work on coherence has revisited sentence ordering attempting to model the structure of text [Barzilay et. al 2002, Lapata 2003, Barzilay and Lee 2004, Barzilay and Lapata 2005]. Other work has focused on

methods to learn ordering such as algorithms that use large corpora to learn an algorithm that can be applied to a summary [Bollegala et. al. 2005] and treating sentence ordering as a Traveling Salesperson Problem (TSP) [Conroy 2006a]. Recent work has reported that evaluating sentence ordering in summaries may require multiple assessors [Madnami et. al. 2007].

Co-reference has not improved significantly since the Message Understanding Conference (MUC) evaluations [MUC] ended in 1997 (MUC F score of approximately 80%). Thus referents such as pronouns and other referents, e.g., this or that, still occur in sentences in extracts and must be carefully addressed in the context of the document or they can lead to false implications due to the fact that a pronoun or reference either appears to belong to a name or fact that isn't the correct one due to the sentence containing the reference being inserted before a sentence which no longer contains the original reference or for the case of natural language generation, the reference has been incorrectly mapped.

Multi-document and summarization systems have also started to address users' information seeking goals by creating effective user interfaces to browse information. Two research efforts from the early 2000s are available online for newswire summaries – Columbia University's Newsblaster [NEWSBLASTER] and University of Michigan's NewsInEssence [NEWSINESSENCE]. Two other systems perform multi-document summarization and visualization for documents retrieved in response to a query: Ultimate Research Assistant [ULTIMATERESEARCH] and iResearchReporter [IRESEARCH]
The Ultimate Research report creates single document summaries for relevant web pages plus a tag cloud.

Researchers have started to analyze computer mediated communication genres in the past few years as data has started to become available. Blog summarization is starting to become a focus area with the new Text Analysis Conference (TAC) evaluation in November 2008 [TAC] of blog summarization. Other communicative genres such as email and chat, are also receiving attention. In the past few years, with the public release of the Enron corpus in 2003 as well as other email corpora, research has been performed on email summarization. This includes email speech acts [Cohen and Carvalho 2004, Goldstein and Sabin 2006, Carvalho and Cohen 2006,] as well as summarization of email and email threads [Nenkova and Bagga 2003, Wan and McKeown 2004, McKeown Zajic et. al. 2008]. We are aware of two research efforts in the domain of chat both focused on educational summaries [Zhou et. al. 2005 and Joshi and Rose 2007]. Although blogs, email and chat can be found in a single document structure, their inherent composition has a multi-document flavor. People often update blogs, email typically consists of threads that are contained in the same email in a variety of formats and chat consists of a dialogue between one or many people.

At some point in the future, it will be necessary to combine the summaries of all these genres that are topically related and relevant into a multi-document summary.

An integrated genre oriented multi-document summarizer would be able to handle the following types of document collections and browsing tasks:

(1) a user is presented with a collection of dis-similar documents and wishes to access the information landscape contained in the collection,
(2) the user has a collection of topically-related documents and wishes to find the key points in the collection, and
(3) the user is looking for a piece of information using a search engine on the World Wide Web and the engine has retrieved several thousand pages, some relevant to the user and some not.

In all of these cases, a multi-document summarizer must be able to extract and indicate the key features of the information space. It would focus on sections of the documents depending of the focus of the summary (if known) and the genre of the document.

1.2 Motivation by Example

In this section, we show some examples of genre oriented goal focused summaries and multi-document summaries to motivate the utility of such enhancements to single document summarization systems.

1.2.1 Goal Focused Genre Oriented Summarization

A goal focused genre oriented summarizer determines the classification of a document and then summarizes it accordingly based on user options. Suppose we just had a generic summarization system for newswire articles. Such a system could use lead sentences (often found to be useful for summarizing news articles) to summarize the document. However, as Figure 1-9 shows, this algorithm may not produce a very useful summary for the movie review genre - depending on the user's goals. With a newswire summary algorithm which doesn't just use the lead sentences, the system provides a different type of overview summary (Figure 1-10). This summary can be compared to a summary produced if a reader wants some of the reviewer's opinions about the movie *Lucky Numbers (*Figure 1-13*)*. This example, illustrates the point, that different readers may want different summaries of a movie review.

Possible summaries for the movie review genre might be plot, opinion and overview. These summaries for the movie review genre (Figure 1-11 overview summary algorithm, Figure 1-12 plot summary algorithm, and Figure 1-13 opinion summary algorithm) show that with an appropriate algorithm we can get better summaries than just the newswire summary algorithms, Figure 1-9 and Figure 1-10. Furthermore, the movie review genre summaries are enhanced with items such as Rating and Running Time in the header due to the enhanced summarization system, which includes a genre classification label. This label allows the summarization system to provide genre specific types of extraction from the document. Thus the resulting summaries, which use both the genre of the document and user's information seeking goals, are much more tailored to a user's needs.

TITLE: REVIEW - FILM - COMEDY LUCKY NUMBERS
AUTHOR: By Ellen Futterman
DATE: Friday, October 27, 2000|2:15 a.m.

- [1] You're definitely living large when you've got your own booth at the local Denny's.
- [2] Or so thinks Russ Richards, a television weatherman in Harrisburg, Pa., and the town's biggest celebrity.
- [3] Richards (John Travolta) could easily be the poster boy for Carol House Furniture because he really does like nice things.
- [4] He drives a Jag and lives in a well-appointed mansion.
- [5] But Richards is experiencing some unluckiness; he's on the verge of bankruptcy.

Figure 1-9: Movie Review using Newswire Genre: Lead Sentence Summary. Sentence Number followed by Sentence. Sentence colored in green is unique to this algorithm.

TITLE: REVIEW - FILM - COMEDY LUCKY NUMBERS
AUTHOR: By Ellen Futterman
DATE: Friday, October 27, 2000|2:15 a.m.

- [1] You're definitely living large when you've got your own booth at the local Denny's.
- [10] Among them are a "twisted psycho bitch" (Lisa Kudrow), a strip club owner with underworld ties (Tim Roth) and a bat-toting killer (Michael Rapaport) named Dale the Thug.
- [14] Loosely based on the 1980 scheme to fix the Pennsylvania lottery, "Lucky Numbers" never quite finds the right footing.
- [17] After the critical and box-office bomb "Battlefield Earth," Travolta returns from the twilight zone to his comfort zone in comedy.
- [24] As dark comedies go, "Lucky Numbers" is no one-in-a-million, but Travolta and Kudrow do their part to keep it spinning.

Figure 1-10: Movie Review using Newswire Genre: Generic Summary. Sentence Number followed by Sentence. Sentence colored in green is unique to this algorithm.

TITLE: REVIEW - FILM - COMEDY   LUCKY NUMBERS
AUTHOR: By Ellen Futterman
RATING: * * 1/2 (out of four)
        Rating: R, language, violence, adult themes
RUNNING TIME: Running time: 1:50
DATE: Friday, October 27, 2000|2:15 a.m.

- [3] Richards (John Travolta) could easily be the poster boy for Carol House Furniture because he really does like nice things.
- [8] So with some help, Richards concocts a plan to rig the state lottery and win the $6.4 million jackpot.
- [14] Loosely based on the 1980 scheme to fix the Pennsylvania lottery, "Lucky Numbers" never quite finds the right footing.
- [17] After the critical and box-office bomb "Battlefield Earth," Travolta returns from the twilight zone to his comfort zone in comedy.
- [24} As dark comedies go, "Lucky Numbers" is no one-in-a-million, but Travolta and Kudrow do their part to keep it spinning.

Figure 1-11: Movie Review using Movie Genre: Overview Summary
Sentence Number followed by Sentence. Genre specific information from "movie" genre is in red.

TITLE: REVIEW - FILM - COMEDY   LUCKY NUMBERS
AUTHOR: By Ellen Futterman
RATING: * * 1/2 (out of four)
        Rating: R, language, violence, adult themes
RUNNING TIME: Running time: 1:50
DATE: Friday, October 27, 2000|2:15 a.m.

- [3] Richards (John Travolta) could easily be the poster boy for Carol House Furniture because he really does like nice things.
- [4] He drives a Jag and lives in a well-appointed mansion.
- [5] But Richards is experiencing some unluckiness; he's on the verge of bankruptcy.
- [6] It's almost Christmas and a heat wave has hit Harrisburg.
- [7] His snowmobile dealership is going bust.

Figure 1-12: Movie Review using Movie Genre: Plot Summary.
Sentence Number followed by Sentence. Genre specific information from "movie" genre is in red. Sentences colored in green are unique to this algorithm.

TITLE: REVIEW - FILM - COMEDY LUCKY NUMBERS
AUTHOR: By Ellen Futterman
RATING: * * 1/2 (out of four)
Rating: R, language, violence, adult themes
RUNNING TIME: Running time: 1:50
DATE: Friday, October 27, 2000|2:15 a.m.

- [12] By the same token, the film does squeeze some genuine laughs out of some nasty situations.
- [13] The picture's biggest problem is its unevenness.
- [15] The story and the direction lack the appeal and the energy of a well-conceived, well-paced caper.
- [16] Ephron, whose forte has been such romantic comedies as "Sleepless in Seattle" and "You've Got Mail," handles the light stuff fairly well but seems uncomfortable with the film's brushes with violence.
- [24] As dark comedies go, "Lucky Numbers" is no one-in-a-million, but Travolta and Kudrow do their part to keep it spinning.

Figure 1-13: Movie Review using Movie Genre: Opinion Summary. Sentence Number followed by Sentence. Genre specific information from "movie" genre is in red font. Sentences colored in green are unique to this algorithm.

1.2.2 Motivation for Multi-Document Summarization Enhancements

One of the biggest issues with multi-document summarization is the repetition and duplication of information as shown in our earlier example. News documents by the nature of their genre must repeat information since each article is designed to stand alone. While this is effective from the author's perspective as well as a reader's perspective, who might not have read the original articles, a multi-document summarizer has the task of synthesizing all the information across the documents to minimize redundancy and maximize useful information content for the purpose of ensuring the continued interest of the reader. Furthermore, often the multi-document summaries are constrained by available space.

To demonstrate the effectiveness of anti-redundancy measures, consider the following output (Figure 1-14) from our summarizer not using anti-redundancy measures, for a 10 document set spanning 3 days on the January 2000 Norway Rail crash. Sentences 1 and 2 are near duplicates, Sentences 4 and 5 are also near duplicates, Sentence 9 is contained in Sentence 10, Sentence 8 is contained in Sentence 6 and Sentence 3 contains similar information to that of Sentence 7. Thus nearly 50% of the information is repeated and virtually "useless" to a reader. In contrast, the summary in Figure 1-15, generated using MMR-MD with a value of λ set to 0.3 (1.0 means no anti-redundancy measures) shows significant improvements in eliminating redundancy. The new summary retains only one

sentence from the original summary although the majority of the information in the original summary is contained in the new summary.[2]

Consider also the case of a larger document collection which spans several years as in the TIPSTER evaluation corpus which provides several sets of 200 topically based news articles.

The set of apartheid-related news-wire documents from the Associated Press and the Wall Street Journal spans a period from 1988 to 1992. The query was composed of the TIPSTER provided topic description. The 200 documents were on an average 31 sentences in length, with a total of 6115 sentences. Generating a summary 10 sentences long resulted in a sentence compression ratio of 0.2% and a character compression of .3%, approximately two orders of magnitude different with compression ratios used in single document summarization. The results of summarizing this document set with a value of λ set to 1 (effectively query relevance, but no MMR-MD) and λ set to 0.3 (both query relevance and MMR-MD anti-redundancy) are shown in Figure 1-16 and Figure 1-17 respectively. The summary in Figure 1-16 clearly illustrates the need for reducing redundancy and maximizing novel information.

Upon examining the summary shown in Figure 1-16, the fact that the ANC is fighting to overthrow the government is mentioned seven times (sentences #2,--#4, #6--#9), which constitutes 70% of the sentences in the summary. Furthermore, sentence #3 is an exact duplicate of sentence #2, and sentence #7 is almost identical to sentence #4. In contrast, the summary in Figure 1-17, generated using MMR-MD with a value of λ set to 0.3 shows significant improvements in eliminating redundancy. The fact that the ANC is fighting to overthrow the government is mentioned only twice (sentences #3, #7), and one of these sentences has additional information in it. The new summary retained only three of the sentences from the earlier summary. Counting clearly distinct propositions in both cases, yields a 60% greater information content for the MMR-MD case, though both summaries are equivalent in length.

[2] The results reported are based on the use of the SMART search engine[Buckley 85] to compute cosine similarities (with a SMART weighting of *lnn* for both queries and passages), stopwords eliminated from the indexed data and stemming turned on.

- [1 10 1] Norway's train drivers on Thursday began a boycott of a line where two trains crashed this week, killing at least 16 people, after a driver apparently passed a red stop signal.
- [2 9 1] Norway's train drivers on Thursday began a boycott of a line where two trains crashed this week, killing about 20 people, after a driver apparently passed a red stop signal.
- [3 5 1] ASTA, Norway (Reuters) - Norwegian rescuers on Wednesday recovered bodies from the burned-out wreck of two trains in which up to 33 people, including schoolchildren, were feared killed in a head-on collision.
- [4 8 1] ASTA, Norway (Reuters) - Norwegian rail controllers tried to telephone two train drivers to tell them to halt before a head-on collision that killed 20 to 30 people but had a wrong list of numbers, a television report said Wednesday.
- [5 6 1] ASTA, Norway (Reuters) - Norwegian rail controllers tried to telephone two train drivers to tell them to halt before a collision that killed up to 33 people but had an incorrect list of numbers, a television report said Wednesday.
- [6 3 6] If the death toll is as high as feared it will pass Norway's most recent comparable crash, when 27 people died further north on the same line in 1975, and be worse than Europe's last large rail accident, in which 31 people died near London's Paddington station in October.
- [7 4 1] ASTA, Norway (Reuters) - Children on a shopping trip on the last day of the Christmas holiday were feared to be among 33 people believed to have died in a head-on collision between two trains in Norway, police said on Wednesday.
- [8 4 19] If the death toll is as high as feared it will be Norway's worst rail crash since 1975, when 27 people died in an accident further north on the same line.
- [9 4 22] Officials said the line lacked some modern safety controls used on other lines in Norway, including a system to prevent trains from driving through red stop signs.
- [10 3 16] Officials said it was too early to speculate on what went wrong but the line lacked some modern safety controls used on other lines in Norway, including a system to prevent trains from driving past red stop signs.

Figure 1-14: Sample multi-document summary with λ = 1 (no anti-redundancy)
Rank order: Sentence Number, Document Number, Sentence Number in Document, Sentence.
The parameter λ [between 0 and 1] allows the user to select how much redundancy to eliminate. A setting of λ=1 means no anti-redundancy measures are applied.

- [1 2 25] ``I heard a terrible crash...(and) thought at first that we had collided with an elk,'' Jeanette Haug, 23, told Norway's NTB news agency.
- [2 3 1] ASTA, Norway (Reuters) - Norwegian rescue workers will start the search on Wednesday through the burned-out wrecks of two trains in which up to 33 people are feared to have died in a head-on collision.
- [3 3 13] Rescuers did not try to enter the trains after firefighters doused the blaze, fearing possible explosions and saying the charred carriages were still dangerously hot despite freezing temperatures outside.
- [4 5 7] Flags flew at half mast at railway stations around Norway after what could be the nation's worst rail crash, surpassing an 1975 accident in which 27 died farther north on the same line.
- [5 5 11] ``We have seen more dead bodies inside the trains'' beyond the seven known dead, Ove Osgjelten, police rescue chief, told Reuters at the site.
- [6 6 21] Police say that 67 people of the 100 aboard the two trains survived the accident, some with severe burns, leaving 33 feared dead.
- [7 8 30] At least one 12-year-old girl on a shopping trip was feared killed on the northbound train but local schools reported that several others feared missing were safe.
- [8 9 8] Police say a total of 19 people have now been reported missing, giving a guide to the likely number of dead, but down from early estimates of up to 33 killed.
- [9 10 1] Norway's train drivers on Thursday began a boycott of a line where two trains crashed this week, killing at least 16 people, after a driver apparently passed a red stop signal.
- [10 10 28] One television report said the controllers in nearby Hamar saw a crash was imminent and tried to warn the drivers but had the wrong list of phone numbers.

Figure 1-15: Sample multi-document summary with $\lambda = 0.3$, time-line ordering Sentence Number, Document Number, Sentence Number in Document, Sentence. The parameter λ [between 0 and 1] allows the user to select how much redundancy to eliminate. A setting of λ=1 means no anti-redundancy measures are applied.

- [1 WSJ910204-0176: 1] CAPE TOWN, South Africa – President F.W. de Klerk's proposal to repeal the major pillars of apartheid drew a generally positive response from black leaders, but African National Congress leader Nelson Mandela called on the international community to continue economic sanctions against South Africa until the government takes further steps.
- [2 AP880803-0082: 25] Three Canadian anti-apartheid groups issued a statement urging the government to sever diplomatic and economic links with South Africa and aid the African National Congress, the banned group fighting the white-dominated government in South Africa.
- [3 AP880803-0080: 25] Three Canadian anti-apartheid groups issued a statement urging the government to sever diplomatic and economic links with South Africa and aid the African National Congress, the banned group fighting the white-dominated government in South Africa.
- [4 AP880802-0165: 23] South Africa says the ANC, the main black group fighting to overthrow South Africa's white government, has seven major military bases in Angola, and the Pretoria government wants those bases closed down.
- [5 AP880212-0060: 14] ANGOP quoted the Angolan statement as saying the main causes of confict in the region are South Africa's``illegal occupation'' of Namibia, South African attacks against its black-ruled neighbors and its alleged creation of armed groups to carry out ``terrorist activities'' in those countries, and the denial of political rights to the black majority in South Africa.
- [6 AP880823-0069: 17] The ANC is the main guerrilla group fighting to overthrow the South African government and end apartheid, the system of racial segregation in which South Africa's black majority has no vote in national affairs.
- [7 AP880803-0158: 26] South Africa says the ANC, the main black group fighting to overthrow South Africa's white-led government, has seven major military bases in Angola, and it wants those bases closed down.
- [8 AP880613-0126: 15] The ANC is fighting to topple the South African government and its policy of apartheid, under which the nation's 26 million blacks have no voice in national affairs and the 5 million whites control the economy and dominate government.
- [9 AP880212-0060: 13] The African National Congress is the main rebel movement fighting South Africa's white-led government and SWAPO is a black guerrilla group fighting for independence for Namibia, which is administered by South Africa.
- [10 WSJ870129-0051: 1] Secretary of State George Shultz, in a meeting with Oliver Tambo, head of the African National Congress, voiced concerns about Soviet influence on the black South African group and the ANC's use of violence in the struggle against apartheid.

Figure 1-16: Sample multi-document summary with λ = 1, news-story-principle ordering (rank order)
Sentence Number, Document Number, Sentence Number in Document, Sentence. The parameter λ [between 0 and 1] allows the user to select how much redundancy to eliminate. A setting of λ=1 means no anti-redundancy measures are applied.

- [1 WSJ870129-0051 1] Secretary of State George Shultz, in a meeting with Oliver Tambo, head of the African National Congress, voiced concerns about Soviet influence on the black South African group and the ANC's use of violence in the struggle against apartheid.
- [2 WSJ880422-0133 44] (See related story: "ANC: Apartheid's Foes -- The Long Struggle: The ANC Is Banned, But It Is in the Hearts of a Nation's Blacks --- In South Africa, the Group Survives Assassinations, Government Crackdowns --- The Black, Green and Gold" -- WSJ April 22, 1988)
- [3 AP880803-0158 26] South Africa says the ANC, the main black group fighting to overthrow South Africa's white-led government, has seven major military bases in Angola, and it wants those bases closed down.
- [4 AP880919-0052 5] But activist clergymen from South Africa said the pontiff should have spoken out more forcefully against their white-minority government's policies of apartheid, under which 26 million blacks have no say in national affairs.
- [5 AP890821-0092 10] Besides ending the emergency and lifting bans on anti-apartheid groups and individual activists, the Harare summit's conditions included the removal of all troops from South Africa's black townships, releasing all political prisoners and ending political trials and executions, and a government commitment to free political discussion.
- [6 WSJ900503-0041 11] Pretoria and the ANC remain far apart on their visions for a post-apartheid South Africa: The ANC wants a simple one-man, one-vote majority rule system, while the government claims that will lead to black domination and insists on constitutional protection of the rights of minorities, including the whites.
- [7 WSJ900807-0037 1] JOHANNESBURG, South Africa -- The African National Congress suspended its 30-year armed struggle against the white minority government, clearing the way for the start of negotiations over a new constitution based on black-white power sharing.
- [8 WSJ900924-0119 20] The African National Congress, South Africa's main black liberation group, forged its sanctions strategy as a means of pressuring the government to abandon white-minority rule.
- [9 WSJ910702-0053 36] At a meeting in South Africa this week, the African National Congress, the major black group, is expected to take a tough line against the white-run government.
- [10 WSJ910204-0176 1] CAPE TOWN, South Africa -- President F.W. de Klerk's proposal to repeal the major pillars of apartheid drew a generally positive response from black leaders, but African National Congress leader Nelson Mandela called on the international community to continue economic sanctions against South Africa until the government takes further steps.

Figure 1-17: Sample multi-document summary with $\lambda = 0.3$, document time-line ordering.
(Sentence Number, Document Number, Sentence Number in Document, Sentence). The parameter λ [between 0 and 1] allows the user to select how much redundancy to eliminate. A setting of λ=1 means no anti-redundancy measures are applied.

1.3 Thesis Statement

The principal question being addressed by this thesis is: can summarization systems effectively indicate the textual content of single documents or document collections in support of users' information seeking goals?

This thesis examines the utility of single document, goal focused genre oriented summarization, email summarization in the context of email "speech" acts and short summaries and multi-document summarization for providing snippets of information to users for evaluation of the utility of the corresponding information. We show by using several human annotated data sets as a "gold standard" for evaluating our summarization system that:

- Our summarization system, in general, can extract relevant sections of documents.
- Genres identification can be performed at sufficient accuracy to inform a downstream summarization system.
- Using the genre information of the document allows the summarization system to outperform the "generic" summarization systems (tailored for news articles).
- Email "speech" acts can be an effective way to summarize short emails, especially coupled with a short one line summary.
- Multi-document summaries function as good indicative summaries by at least indicating the topic material of the data.
- Our multi-document system, which reduces redundancy, outperforms the baseline system (documents concatenated together).

1.4 Thesis Contributions

In addition to building GOLD, a functional Genre focused goal Oriented muLti-Document summarizer, this thesis has the following contributions:

- The development of three human annotated data sets for use in evaluating summarization systems: one for extracting relevant sentences for single document summaries, one for genre oriented summarization, one for multi-document summarization.
- The creation of a web corpus for genre identification studies.
- An analysis of summary quality for multi-lingual summarization.
- An evaluation of the quality of multi-document summaries based on their ability to cover key points of newswire document sets with minimal redundancy.
- The addressing of issues in multi-document summarization, genre oriented summarization and email summarization.
- The creation of a suite of summarization system for single document summarization for multiple genres, multi-document summarization and email speech acts.
- The novel technique of genre oriented goal focused summarization. The use of genre tags for categorizing data and then using a reader's information seeking

goals to create more effective summaries for addressing user's information seeking goals.
- An evaluation of genre identification on 16-42 classes. To our knowledge, this is the largest number of classes compared for genre identification.
- A comparison of the classifiers Random Forests and Support Vector Machines.
- An evaluation of the quality of goal focused genre oriented summarization.
- A brief examination of email summarization, including summarization results from annotated Enron email data.

1.5 Thesis Outline and Reader's Guide

This thesis discusses the requirements of a goal focused multi-document summarization system. It describes our studies of single document, genre oriented and multi-document summarization by human summarizers, the results of which we have applied as a *gold standard* for evaluating our summarization system. We also describe our approach to multi-document summarization that builds on our previous work in single-document summarization by using additional, available information about the document set as a whole, the relationships between the documents, as well as genre characteristics of a document collection.

The user can modify parameters to produce different types of summaries, including the trade-off between preserving emphasis (usually redundant information due to repetition) and novelty, focusing on the most recent information, and modifying the summary size. Our system does not currently address co-reference other than penalizing passages with pronominal references.

- Chapter 2: *Summarization Discussion* - provides a framework for summarization and the components necessary for summarization systems.
- Chapter 3 *Our Summarization System* - discusses our single document and multi-document summarization systems.
- Chapter 4 *Metrics* - provides an overview of metrics used for evaluating summarization systems.
- Chapter 5 *Evaluations* - discusses past and current summarization evaluations and presents the content evaluation that we performed for multi-lingual summaries.
- Chapter 6 *Genre Identification* - describes the corpora we collected for genre identification and our experiments with genre-identification on this data set.
- Chapter 7 *Genre Oriented Goal Oriented Summarization* - presents our definition of genre-oriented goal focused summaries and presents our system results.
- Chapter 8 *Email Summarization* - examines email speech act identification as well as analyzes summarization for a portion of the Enron email corpus.
- Chapter 9 *Multi-document Summarization* - highlights our work on multi-document summarization in the newswire domain.
- Chapter 10 *Conclusions and Future Work*.

Following are our recommendations for a reader who has limited time, but is interested in the following topics:

- An Overview of Summarization: Chapters 1 & 2
- A Synopsis of the Past and Present Government and Industry Sponsored Evaluations Open to the Community: Chapter 5 and if an interest in the metrics used in these evaluations, Chapter 4
- Metrics and Possible Metrics that can be used for Summarization: Chapter 4
- Summarization
    - Single document Summarization: Chapters 2 & 3
    - Genre Oriented Goal focused Summarization: Chapters 2 & 7
    - Summarization in the Email Genre: Chapter 8
    - Multi-document Summarization in the Newswire Genre: Chapters 2, 3 & 9
- Algorithms:
    - The algorithms in our suite of summarization systems: Chapters 3 & 7
    - The anti-redundancy metric MMR for single and multi-document summarization: Chapter 3
- Genre Identification Classification Experiments: Chapter 6

Chapter 2 Discussion of Summarization

> **"Both author and reader have to accept the rugged landscape of summarization for the time being, before they can advance the cultivation of the field, by further developing this or that theory, by teaching this or that summarization technique to their students, or by applying a newly acquired strategy in their summarizing practice."**
>
> Brigitte Endres-Niggeyer, *Summarizing Information* 1998.

Human-quality summarization, in general, is difficult to achieve without natural language understanding due to the fact that summarization is a cognitive task often dependent on the situation. Documents are typically composed with one of more topic themes that are elaborated on and supported through the text. Furthermore, there is much variation in writing styles, document genres, lexical items, syntactic constructions, etc., that must be accounted for in a summarization system. Generating an effective summary requires the summarizer to *select*, *evaluate*, *order* and *aggregate* items of information according to their relevance to a particular subject and purpose.

The purpose of summarization is to reduce the original amount of information. An ideal text summary includes the relevant information for which the user is looking and excludes extraneous and redundant information, while providing background to suit the user's profile. It must also be coherent and comprehensible, qualities that are difficult to achieve without using natural language processing to handle issues such as co-reference. Summaries also must be an appropriate length to fit the user's information seeking goals - as some people want to read a brief synopsis, others require more details. Fortunately, it is possible to exploit regularities and patterns -- such as lexical repetition and document structure -- to generate reasonable summaries in most document genres without having to do any natural language understanding. These summaries are certainly suitable for indicating the content of the document and often quite informative as well [SUMMAC, DUC].

This chapter discusses the dimensions of the summarization task as well as the nature of a good summary - one that meets the user's information seeking goals. This includes purpose, quality and length among other factors. We end this chapter with a high level overview of useful functionality for goal focused summarization and multi-document summarization systems.

2.1 Dimensions of Summarization

Since summaries can be used for a variety of functions from indices to abstracts, there are many factors to consider when performing summarization, human or machine. We list 12 that we have defined below.

1. *summary construct*: A natural language generated summary, ***abstract***, is created by use of a semantic representation that reflects the structure and main points of

the text, whereas an ***extract*** summary contains pieces of the original text, such as key words, phrases, paragraphs, full sentences, and/or condensed sentences. Both text extract summaries and natural language generated summaries can contain a combination of these types, such as keywords and a paragraph, which was what Textwise used in the SUMMAC evaluation [Mani et. al. 1998]. Note that summaries composed of shortened sentences are sometimes referred to as *telegraphic* since perceived extraneous words and phrases are deleted, a measure that was employed to save costs in overseas transmissions using the telegraph machine. If this technique is employed, words must be carefully eliminated by summarization systems to avoid changing the meaning of the text. For example, consider the sentence "John rarely went to the store". In telegraphic summarization, adverbs and adjectives are often deleted and in this case, the absence of the word "rarely" would indicate that John performed an action that was probably not true.

2. *summary purpose*: A ***generic*** or *overview* summary gives an overall sense of the document's content, whereas a ***query-relevant*** summary presents the content that is most closely related to a search query or user model, and a ***goal-focused*** summary provides information related to a particular objective, such as plot or opinion for the movie review genre. Generic and query-relevant are subcategories of goal-focused summarization, which is covered in Section 2.4.
3. *summary type:* A summary is defined as a brief account giving the main points of something. This includes a set of keywords, a headline, title, abstract, extract, goal-focused abstract, index or table of contents. News articles are inherently summaries and summarization in this genre results in a summary of a summary.
4. *summary role*: An ***indicative*** summary provides the user with an overview of the content of a document or document collection, whereas the purpose of an ***informative*** summary is to present the most relevant "informative" information which would allow the user to gain knowledge about a particular topic or answer one or more questions. An informative summary's purpose would be to act as a replacement for the original text.
5. *set for summarization*: A ***single document*** summary provides an overview of one document, whereas a ***multi-document*** summary provides this functionality for many.
6. *genre*: The information contained in a document's genre can provide linguistic and structural information useful for summary creation. Different genres for text include news articles, editorials, letters and memos, email, scientific articles, books, web pages and speech transcripts (including monologues and dialogues). A summarization system must take into account document genre and use this information to produce the most effective summary possible.
7. *modality*: Summarization has tended to focus on *text-only* written summaries. Several research groups have recently begun exploring the summarization of multi-modal and multi-media input. Recent NTCIR evaluations 2007 and 2008 are evaluating such summarization [NTCIR].
8. *user goal*: Is the user's goal to browse for information or search for specific information? The summary needs to reflect the user task. When browsing, people may prefer indicative summaries from which they can choose to explore

additional content. When searching for specific information, a well constructed informative summary would ideally directly answer their information seeking objectives, eliminating the need to refer to any original documents. Is the user looking, directly or indirectly, for a particular focus to their summary that would help them with their goals? This type of summary could be related to the genre as well. We refer to specific summaries of interest to users, which address their information seeking goals as *goal-focused*.

9. *summary length*: The length of the summary may be a *fixed length* or a *percentage of the size of the original document(s),* i.e., *compression ratio*. The best summary length is a function of the genre, the document length, the content redundancy present in the document(s), and naturally the reader's information seeking goals. For example, individual newswire articles are usually intended to be summaries of an event and therefore are short and contain minimal amounts of redundancy and therefore can easily be summarized in 3-5 sentences. Scientific articles tend to have more redundancy - they are often written to present a point, expand on the point and reiterate it in the conclusion. They still tend to be summarized in an abstract, often approximately 150 words. On the other hand, a group of newswire articles about an event contains lots of redundancy, but may require more summary sentences if the summary needs to describe the unfolding of the event. For example, in a set about a plane crash, there may be multiple references to the details of the crash followed by many updates about the search and rescue operations, retrieval of the black box and investigation into the reasons for the crash.
10. *summary presentation*: The output summary can be keywords, phrases, sentences, paragraphs. This can be presented in a text-only format, or text with hyperlink references to the original text(s). An effective user interface would allow the opportunity to expand or contract the summary length as well as provide more contextual sentences around each summary sentence or component in the case of text extract summaries.
11. *summary source language:* With the advent of translingual (or crosslingual) and multilingual information retrieval, portions of the input document set may be in a different language than the output language.
12. *summary quality:* As in any presentation, the summary must be coherent, cohesive and if it contains sentences grammatical as well. It should be readable and accurately reflect the original text, i.e., not contain false implications based on missing references or poor constructions.

Our work focuses on indicative and informative summaries using extracts for generic, goal-focused and query-relevant summarization.

Summaries can be evaluated as in the TIPSTER SUMMAC conference by how well they perform on certain **extrinsic** tasks [Mani et. al 1998]: indicating document relevance to a topic or indicating a category for the document. They can also be evaluated **intrinsically**, as to whether they contain answers to specific questions or cover the content of human gold standard summaries. Another intrinsic measure is whether or not system summaries extract the relevant portions of text [Teufel and Moens 1997, Jing et. al. 1998, Goldstein

et. al. 1999, DUC, Lin and Hovy 2003, Radev and Tam 2003] and more recently as to whether they have semantic content overlap [Nenkova and Passonneau 2004, Hovy et. al 2005]. We discuss these metrics in Chapter 4 and the formal summarization evaluations in Chapter 5.

2.2 Essentials for Useful Summaries

In the previous section, we discussed the multiple dimensions of summarization. Many of these factors are influenced by the user's goals. For example, if the user would like to create a multi-document summary of news documents reflecting worldwide opinions on a particular topic, it would be beneficial to obtain documents in the original source languages, perform machine translation and create summaries in the requestor's native language or desired output language.

In this section, we examine the fundamentals for a good summary from both an information seeking perspective and a system generation perspective. This can be thought of as how to focus summary creation for the purpose of being used in real world applications (extrinsic evaluation), or for the purpose of having a good internal composition (an intrinsic evaluation). Summaries could be subjected to a process such as the Turing Test – can a human tell that the summary was machine generated?

We conclude this section by addressing another important question - what is an ideal length for summaries? In our work, we focus on user's goals with respect to genre. As the summary length is an important part of fixed summary creation, we discuss this as well.

2.2.1 Creating a Summary Addressing the End User's Goals

One of the difficulties in creating summaries for evaluations is that often the end user and task is not specified. An expert in a field does not want summaries that include background information that s/he already knows – this is part of domain knowledge that is assumed. Similarly, a person who is familiar with current news events and following a particular event would not require an analysis of the event in a summary of novel related information (*update* summary), but a person who has been traveling in a remote area and just returned might need such information (a generic or overview summary). Summaries need to be tailored to individuals' information seeking needs. They are often difficult to constrain to a one size fits all model. Even published summaries, including news articles, book jacket reviews, Cliff notes and abstracts, although sometimes constrained to a maximum size, appear to have a typical range of sizes.

Although there is a variety of possible summaries and end users, it is possible to categorize these in a manner that shows the functionality required to address the domain of possibilities. Xie has performed an analysis of interactive intentions and information-seeking strategies in interactive information retrieval [Xie 2000]. She characterizes goals into four levels of hierarchical structure, which we present in Table 2-1: Xie's Levels of User Goals.

Summarization can be used for all goals and subgoals listed in Table 2-1: Xie's Levels of User Goals. Goal Level 1, Long Term Goal, might require summaries to be delivered daily for the user to read. Current search engines and some news sites allow one to tailor searches, the results of which are delivered daily or viewable through a browser. These results are typically a summary of the retrieved items. For a search goal, the result of any web search is a list of small web page or document summaries. Each card in the library card catalog is a summary – the summary contains information about the item. The title or subject and category classification through the Dewey Decimal System provide an indication of the content (indicative summary).

Table 2-1: Xie's Levels of User Goals

Level	Type of User Goals	Definition	Examples
1	Long-term Goal	Personal Goal	Professional achievements, personal interests
2	Leading Search Goal	Task-related goal that leads to a search	Writing a paper, preparing a project
3	Current Search Goal	Specific Search Results a user intends to obtain	Looking for a model, a book, answers to questions
4	Interactive Intention	Subgoals a user must achieve to accomplish "current Search Goal"	Identify, Learn, Find, Access, Locate, Evaluate, Record, Obtain

Summaries can be constructed to address user's specific goals in Level 4, Interactive Intentions. For example a summary used to "identify" or "find" in Level 4 is more indicative than one used to "learn". A summary for "learning" should be informative. A summary for the purpose of "recording" should have items relating to the source, but including such information in the case of summaries for "identify" or "learn" would just consume valuable summary real estate and require the user to spend more time reading or scanning the summary than necessary.

To achieve a user's interactive intention (Level 4 in Table 2-1), Xie suggests information-seeking strategies to achieve the interactive intentions. Her results indicate that these strategies can be characterized by combining eight methods and six types of resources. The eight *methods* are the techniques that users apply to interact with information: Scanning, searching, tracking, selecting, comparing, acquiring, consulting and trial and error. The six types of *resources* are meta-information, part of an item/specific information, whole item, a series of items/one location, one system/multiple databases and human.

Summarization fits in this framework through the "scanning" and "comparing" techniques. An *appropriate* summary of the above mentioned *resource* "part of an item/specific information" can be used effectively for both of the above mentioned *methods* of "scanning" and "comparing". Since the current search goal pertains to requesting specific results that meet the user's information seeking goal, a summary must

reflect such information to be effective for both the scanning and comparing methods, hence it must be *appropriately* tailored for the user's goals. This is the motivation for *goal-focused summaries*. In an ideal system, the system would either infer the user's goal(s) through observation and/or analysis of the user and data, or the user would select specific parameters to tailor the system output to address their goals. We discuss the characteristics of goal-focused summaries in more detail in section 2.4 after a discussion in the next subsection about the composition of summaries.

2.2.2 Creating a Summary with Good Internal Composition

In order to create an effective summary, one must first decide what constitutes such a summary. People summarize information and communicate such synopses to others daily, so many people might think they are capable, effective summarizers. However, as in many fields, there are experts and there are novices.

Endres-Niggemeyer states that professional summarizers are provided with explicit methods, some of which are taken from standards, textbooks, and guidelines. She presents a methodology in her book *Summarizing Information* [Endres-Niggemeyer 1998], part of which is shown in Figure 2-1.

In terms of relevance criteria in Figure 2-1, – *purpose* and *topic relevance* are the easiest items for summarization systems to address. For query relevant summaries, the purpose is to provide a topically related summary and for generic summaries, an overview. For goal-focused, the purpose is to address the user' information seeking goals. *Topic* is either provided from the topical description, query or other goal, and in the case of generic summaries, the topic is determined from the overall content of the document.

Fact relevance is difficult in that the domain knowledge may be present in the assumptions that the author is making about the reader or based on certain real world assumed knowledge. In the case of the former, the summarization system may not be able to infer this and in the case of the latter, even if the system is able to rely on a knowledge base, this repository may not contain the necessary subject matter. Relpositive, contrast, and stress *relevance* have not been a current focus of summarization systems or evaluations.

Different arguments and items of background knowledge are used to decide the importance (**relevance** and interestingness of an item) – **Relevance Criteria**:

- Fact: Relevant is what is important according to domain knowledge.
- Topic: Relevant is what relates to the text topic.
- Purpose: Relevant is what serves the purpose.
- Relpositive: Relevant is what is stated positively.
- Contrast: Relevant is what differs from other things.
- Stress: Relevant is what is characterized as relevant in the text.

M**eaning Reduction Strategies**:

- Noreason: If you have the statement, do away with its reason.
- Novoid: Leave it out if is not informative.
- Nocomment: No comments and added explanations
- Noexample: Drop Examples.

A**ssessments** of importance, relevance and interestingness:

a. The *structure of the topic* determines what is to be included in the summary
b. A summary should answer the *summary user' questions* – what fits their needs
c. A summary should have ease of knowledge assimilation and thus has to look at *prior knowledge* of its recipients
d. A summary must have *information value*, *innovation value* and *interestingness*
e. The amount of information in a summary constrains the relevance decisions
f. A summary should respect the *original author's design* whenever possible
g. A summary should convey the *author's intention*

Figure 2-1: Endres-Niggemeyer's Framework for Summaries: Relevance Criteria, Meaning Reduction Strategies and Assessments

The meaning reduction strategies in Table 2-1: Xie's Levels of User Goals are useful methods for choosing a sentence. We are not aware of any current summarization system that is utilizing these strategies. Our first pass summarization system does not incorporate methods for these strategies either and we leave the development of techniques that address these for future work.

The assessments of importance, relevance and interestingness listed in Figure 2-1 address summary structure, quality and length as well as a user's information seeking goals. One interesting item of note is that besides the structure of the topic constraining items that are selected for summaries, these specifications clearly indicate that the summary must convey the author's intention and design – items which have not been addressed in summarization systems.

Current summarization evaluations [DUC, NTCIR, TAC] are assessing information value, but do not assess innovation value or interestingness. We revisit this topic in our discussion on evaluations in Chapter 5.

In the first two evaluation sets that we created (single document relevance judgments and multi-document extract summary creation with subtopic assignments for sentences), we did not use experts in our analysis. When we began to address goal-focused summaries and the internal composition of such summaries, we decided to use the closest cost-effective approach to experts – English students who had some training in such matters. Although at the time Endres-Niggemeyer's book was not available[3], many of the concepts that the English students suggested in our discussions reflected the items listed in Figure 2-1.

2.2.3 Ideal Summary Length

There have been various lengths used for summaries in evaluations. The first SUMMAC evaluation on newswire documents compared summaries for a fixed document character size compression and a variable size [SUMMAC]; fixed length summaries consisted of 10% of the document character length. Using this criterion, for extremely long documents, such as theses and reports, a 200 page report would have an approximate 20 page summary at a 10% compression. Most summaries of reports and theses have an abstract that fits on one page. Formal evaluations post SUMMAC for single and multi-document (abstract) summarization have used a byte limit, 665 bytes, or a word limit, which ranged between 100 words to 250 words [DUC, NTCIR]. Perhaps the longer 250 word summary is effective as an informative summary from which readers glean information. However, it is perhaps too long for a quick indicative summary used for the purpose of document triage, information routing. What then is a "good" size for a multi-document summary? In our opinion it is task specific and depends on the end user.

However, what appears to be clear is that many summaries that are published by professionals have a definite fixed length range. Human produced newswire summaries tend to be 3-5 sentences (75-125 words, using an average sentence length of 25 words) [Goldstein et. al. 1999]. Book jacket reviews (summaries) tend to fit on one or both jacket flaps or the back of the book. Press releases tend to be one page or less. Office memos tend to be 1 page (2 pages at most). Abstracts for scientific conference articles tend to be about in the range of 100-200 words. Headlines also tend to have a fixed size - less than one sentence.

For multi-document summaries, we asked 10 people to generate multi-document extract summaries about the key points of 3 clusters of 10 documents. We did not specify a length restriction. Although the length of the summaries varied, they all fit on 1-2 pages supporting the hypothesis multi-document summaries also have a maximum length which people intuitively consider appropriate for summarization. This length would be the approximate size of one news article. These unlimited length "synopsis" multi-document summaries had an average sentence length of 41 - a compression of 13.6%. This summary size - within the range of the sentence length for news articles, whose average sentence length is approximately 30 [Goldstein et. al. 1999]. Thus, perhaps an ideal

[3] The publication date states 1998, but it was not published/available until sometime in 2000-2001.

summarize size for the multi-document genre is approximately the size of one newswire document.

Further investigation is required to determine if this is indeed the case and if single document and multiple document summaries tend to have a fixed length for other genres other than those mentioned above.

Besides the "genre" approach to summary length, there can be practical approaches to summary length. For certain applications there may be a fixed byte length that can be used for a particular field. In others, there can be presentation considerations that affect the chosen summary length. For example, the "summaries" returned for each web search by Google may be motivated by both how much context is necessary for a person to adequately select a link for viewing, as well as how many results can fit on one web page – especially since users prefer not to scroll.

In our experiments with genre oriented summarization, the team that discussed the composition of a summary also included in their deliberations the subject of what is an appropriate length for various types of summaries. Their assessment was that sentence lengths vary by genre and long documents such as interviews require slightly longer summaries (Chapter 7). If the user is interacting with the summary through a user interface, then the user may have the opportunity to specify the length and/or expand the length as desired if the interface is connected to a summarization system that can process summaries on demand. Otherwise it may be necessary to decide on a suitable length for a summary for any given genre.

2.3 Genres

Roussinov and colleagues, in preliminary studies of people searching the Web, discovered that the genre of the document was one of the clues used in assessing relevance, value, quality and usefulness [Roussinov et. a. 1991].

Since genre information includes the communicative purpose and form of the document, the genre can also be used to inform downstream processes, such as summarization, which can provide further information for users to select relevant documents. In particular, summaries can be formed based on the genre. For example, one would expect a store product page to contain a price and for that to be included in a summary. A product press release may not contain any price information. Product reviews are likely to contain opinions whereas both product web pages and product press releases tend to contain only factual information. For this reason, if the genre of the document is known, this information can assist the focus of the resulting summary – whether it should focus on opinion or fact, or whether to try to include price information. Such user-tailored genre summaries contain more useful information than standard summarization algorithms – we demonstrate this in Chapter 7.

The study of genres extends back hundreds of years primarily from the examination of human communication through the dimensions of content, purpose and form. The literary

community has extensively used this concept to divide literature into genres using criteria such as literary technique (purpose and form), tone (purpose), and subject matter (content) resulting in such genres which include autobiography, biography, children's literature, fairy tale, fiction romance, saga, thriller [Wikipedia]. Many of these categories have further subdivisions based on the subject matter. Information technologies have enabled the appearance of many novel genres including email, chat, forums, blogs, online-reviews, FAQs, and homepages.

Although there are many different definitions of genre used by researchers, most include two principal characterizations, that of the intended communicative *purpose* and *form* [Kwasnik and Crowston 2005], and sometimes a third characterization, that of the expected *content* of the document. Our research focuses on this triple (purpose, form, content) since the content of a particular genre informs downstream processes, such as document summarization, as well as allows the downstream processes to make decisions based on this information.

There are many different definitions of genre. The Oxford Dictionary defines genre as a style or category of art of literature. Meriam Webster (www.m-w.com) defines genre as (1) a category of artistic, musical, or literary composition characterized by a particular style, form, or content, (2) kind, sort. WordNet defines genre as a kind of literary or artistic work. www.freedictionary.com defines genre as a type of class. The Collaborative International Dictionary of English v.048 defines genre as "kind; genus; class; form; style, esp. in literature." Pcmag.com defines genre as "a French word meaning category, class, style, type or variety." Lastly, www.brittanica.com states that genre comes from French "kind" or "sort" and is a distinctive type or category of literary composition, such as the epic, tragedy, comedy, novel and short story.

Erikson defines genres [Erikson 2000]: A genre is a pattern of communication created by a combination of the individual, social and technical forces implicit in a recurring communicative situation. A genre structures communication by creating shared expectations about the form and content of the interaction, thus easing the burden of production and interpretation.

We use Erikson's broad definition of "literature", in which any communication (which clearly can be transcribed) that has a particular form and content qualifies as a genre. Thus Radio Talk Shows would be a genre.

Genres and topics are often separated. For example, the website http://books.scholastic.com/teachers allows the user to search by topic or genre. Genres include adventure; biography and autobiography; classics; comedy and humor; comic books and graphic novels; diaries and journals; drama; fables, folktales and myths; functional and how-to; general fiction; general nonfiction; historical fiction; horror and supernatural; multicultural; mystery and suspense; poetry and rhymes; science fiction and fantasy; short stores; reference. For topics, they allow animals, life experiences, math & science, holidays, multicultural, social studies, art, sports, character & values.

The definition, of a genre as a particular purpose, form and content, is used throughout the thesis. We include broad topic categories (which we refer to as *genre-topic*) in our definition of genres as a way to further separate specific genres. Genre-topics must have a large set of writings that follow the genre guidelines for the genre, but have specific topically related items that are highly relevant to that topic. For example, a broad class of genres might be that of the "review", but we prefer to separate these further by topic, e.g., a product review vs. a movie review. These distinctions become important for summarization as a product review could have price information that would be desirable to include in a summary and a movie review might have movie ratings, audience suitable ratings and running times. We refer to these broad categories as genre-topic to distinguish them from topics and subtopics that can be the results of topics and questions. For example in a request for a multi-document summary that summarizes all positive opinions about the latest Star Wars movie, "positive opinions about the latest Star Wars movie" would be the topic and the genre-topic would be movie reviews. In this thesis, we do not separate genre-topic from genre, i.e., we just refer to the genre as movie reviews, unless there is a clear reason to distinguish topics in the manner we have described above.

2.4 Factors for Goal Focused Summarization

As discussed in Subsection 2.2.1, a user can have a variety of information seeking goals that can be categorized in many ways and one such methodology was presented. When addressing a user's specific information seeking goal, i.e. the results that a user intends to obtain – the end results ought to reflect that goal. This is the concept behind goal-focused summarization.

In the early years of summarization, summaries were created that had an *overall* summary purpose which included:

- *generic* (a survey of the information)
- *adhoc* or *query relevant* (addressing the information in the query)
- *topic relevant* (addressing the information presented in a topic query

Upon examination of these three types of summaries, these methodologies only reflect users' information seeking goals on a broad level. We refer to all these types of summaries as *Coarse Summaries*. Some of these "Coarse Summaries" can have more detail. For example, *question answering oriented* summaries, which are a subset of query relevant or topic relevant summaries, are designed to answer specific questions. These summaries are being evaluated for the first time in TAC 2008 [TAC].

If we further examine user's information seeking goals, we can see that other types of summaries can be formed to address particular situations, such as:

- *novelty* – focused on only presenting novel information with respect to a certain "known" body of information (evaluated in DUC 2003) [DUC].
- *update* – focused on presenting information that is a follow-up to previous information (evaluated in DUC 2007-2008, [DUC] and TAC 2008 [TAC].

We refer to this dimension as *summary varieties* and consider them a dimension of goal-focused summarization since they can address any type of Coarse Summary: generic, adhoc or topic relevant summaries.

As an example, let us consider a situation in which a question answering oriented summary is combined with one or more Summary Varieties. Suppose a person is interested on the movement of Hurricane Ike, the damage and the death toll from this storm. The summaries can be question answering oriented - designed to answer specific questions such as "What locations did Hurricane Ike hit?", "What is the death toll caused by this Hurricane and its remnants?", "What damage did Hurricane Ike cause?", "Did damage occur to any historic buildings and if so, how much?". The initial summary might center on the Caribbean, where Hurricane Ike first hit. Later *update* summaries might focus information updating both the death toll and damage in the Caribbean and also for the situation in Galveston, Texas. Such update summaries could also focus on *novelty*, which assumes that previous information about the Hurricane from past dates is known. One such example is a novelty summary about the effects of Hurricane Ike as it passed through Pittsburgh, Pennsylvania causing severe power disruptions. There could be several update summaries about this situation as some people were without power for several days.

"Summary varieties" are still a very broad level of categorization. We suggest another dimension to summarization, which we refer to as *summary facets or facet summaries.* The facets are certain goals for summarizations that can be used across various genres. For example, *plot* summaries would be one such facet summary and can apply to genres which contain stories, including movies, books, and reality shows. Other such types of facet summaries include opinions, thematic, professional and personal. Opinion summaries for blogs are currently being explored in TAC 2008 [TAC]. We also define an overview facet summary, where such a summary is defined specifically for the task.

Summary varieties can be combined with summary facets to address a particular type of information goal. For example, a viewer following the Survivor Reality show, might want an update plot summary after missing certain episodes in the series. An economic analyst tracking a particular company might want a novelty opinion summary with respect to that company.

Such facets tend to be specific to the nature of particular types and/or genres of information. Depending on the genre and the genre-topic, e.g., movie (defined and discussed in Section 2.3), various facet summaries are appropriate. For example, the review genres would tend to include *opinion* summaries. Natural facet summaries for the 'movie review' genre would include an *overview* summary, a *plot* summary and an *opinion* summary. We discuss our summarization system designed for genre-oriented goal-focused summarization in Chapter 7.

The three major requirements for goal focused summarization are listed below. The fourth item, *summary variety* is optional, and if present, provides a further refinement to the output summary.

1. *User Task Focused* – the summary must address the user's information seeking goals.
2. *Facet Summary* – the summary must either be a Facet Summary as described above or a Coarse Summary.
3. *Genre Appropriate* – the summary must be appropriate for the genre of information. For example, an "opinion" summary of a scientific article may not be an appropriate summary for this genre as the author of the scientific article might not express any opinions in his/her writing.
4. *Summary Variety (optional)* – addresses certain information seeking goals.

2.5 Factors for Multi-Document Summarization

Users' information seeking needs and goals vary tremendously. User satisfaction with retrieved information depends on their previous knowledge of the subject as well as the presentation of the information. We have found through our studies that people also tend to summarize articles very differently, perhaps due to perceptions and biases as well as background knowledge and intentions. Some summaries tend to reflect details of one point, others a more general overview. Thus, an ideal multi-document summarization system is able to address different types of summaries and different levels of *detail*, the latter being difficult without natural language understanding. A short summary, which significantly compresses the document set needs to minimize repetitive information in order to maximize the information presented to the user.

In the introduction we illustrated the difficulty of multi-document summarization as compared to single document summarization. In this section, we expand upon this motivation, providing a list of features that multi-document systems might need and which provide the framework for our multi-document summarization system.

At a minimum, a multi-document summarization system needs an interface to permit the user to enter information seeking goals. Such goals could be determined by the system from the user query, a stored user background interest profile (which can contain references to the users "knowledge"), user interface selections determining information to be included in a summary as well as summary format, and/or a relevance feedback mechanism. A user could also be "watched" by the system, and the system ideally could "learn" to infer preferences and choices by "noting" user behavior and tailor the summary output accordingly. For multi-document summarization, a good interface for exploring the data must allow users to view summaries of similar and dissimilar clusters of information.

Following is a list of factors to consider when designing a multi-document summarization system:

1. *clustering*: the ability to cluster similar documents/passages, not only by content, but also by document genre. Thus related information with particular characteristics is grouped as well as the genre. For example, it may be important to identify and separate a passage of an editorial article from a factual article.

2. *relevant coverage*: the ability to find and rank the main points for relevance to a query across documents within a set. If a query has not been explicitly specified, a query could be constructed based on either user context, title of documents if they are present or document-passage centroids for the document set.
3. *similarities and differences*: the ability to identify similar/dis-similar text. This includes the important task of *identification of source inconsistencies*. Articles often have errors (such as billion reported as million, etc.) or differing information (such as closing prices of stock, number of deaths). A multi-document summarizer must be able to recognize and report such source inconsistencies.
4. *anti-redundancy*: the ability to minimize information redundancy. The focus of short summaries is often novel relevant information, in which the same information is not repeated due to the desired optimization of information contained within the summary.
5. *quality*: the ability to optimize readability and coherence (as well as relevance); thus the text must have sufficient context so that the points are understandable to the reader.
    - *co-reference resolution*: If there is co-referent information in the passages in the summary - these references must refer to the proper referent, which can be difficult when selecting text extracts from multiple documents without natural language understanding.
    - *grammaticality*: The passages (syntactic unit) follow English grammar rules (independent of their content). For natural language generated summaries or extensive text span deletion methods such as telegraphic reduction summaries, this item becomes important.
    - *cohesion*: If the summary is presented as a cohesive group of sentences, the sentences must make sense with the surrounding sentences.
    - *organization/coherence*: The content of the summary is expressed and organized in an effective way. Often the summary organization is dependent on the users' information seeking goals.
6. *summary updates*: A new multi-document summary must take into account previous summaries in generating new summaries. This allows a user to obtain new relevant information. In such cases, the system needs to be able to detect, track and categorize events.
7. *indicative or informative*: A multi-document summary must either indicate points of, or facts about, the document collection or function as a more coherent informative summary, similar to an abstract.
8. *multiple genres*: A multi-document summarization system must have the functionality to handle multiple document genres and be able to summarize across genres. Thus a multi-document summary may include some selections from news articles, some from editorials and some from interviews. These include:
    - *dialogs*: Which include transcripts of meetings, interviews, threaded (multiple email sessions in one email message) email.
    - *web pages*: Which may include fragments, non-grammatical text, and hypertext.

9. *information space exploration*: Users need to be able to create different types of summaries to explore the information space (*summary varieties*). These summaries can be:
    - *Common Text Only Summary*: Each summary passage must be linked with (similar to) at least one related passage in another document in the collection, providing a summary that contains many of the main repetitive points of the document collection.
    - *Unique Text Only Summary*: Each summary passage contains a relevant point that only occurs in one document of the collection.
    - *Common and Unique Text Summary*: The summary contains a combination of passages that are linked to related passages in other documents and some relevant text that only occurs in one document of the collection.
    - *Representative Single Document Summary*: The summary is a single document summary from the centroid document in the collection.
    - *Novel Differing Points Summary*: The summary contains only passages from outlier documents (from the centroid document) in the collection, thus differing from the central focus of the main documents.
    - *Document Set Overview Summary*: The summary contains passages from the centroid document and the outlier documents.
    - *Latest Document Summary*: The summary is a single document summary from the latest document.
    - *Common Text Time Weighted Summary*: The summary contains only passages that have similar related passages in other documents and all summary passages are extracted from the most recent documents.
    - *Common Text Time Weighted plus Unique Text Summary*: The summary contains a combination of (1) passages that have similar related passages in other documents in which all such passages are extracted from the most recent documents and (2) selected passages which are unique to one document.
    - *Comparative Summaries*:
        - *Differing Points Summary*: The summary contains differing points of view or differing facts within the document collection. This includes differing opinions as well as source inconsistencies.
        - *Similar Points Summary*: The summary contains similar points of view or facts within the document collection.
        - *Combination of Differing and Similar Points Summary*
    - *Evolution Summary*: The summary contains earlier information presented in the document collection and updates to that information.
    - *Update Summary*: Given a previous summary, the summary contains updated information to the points contained in the summary.
10. *effective user interfaces*: Due to the fact that there are multiple documents and high compression, a user may often want to browse the collection or expand on information in the summary. User goals include:
    - *accessing the source of a passage*: The source of information can indicate the reliability and thus usefulness of the information.

- *eliminating sources of information from use in the summary*: Questionable sources of information may not be desirable. It may be beneficial to list the sources of information in the output summary.
- *viewing more context of a passage*: This context can take the form of the full document, the summary of the full document, or the surrounding context of this passage.
- *viewing related (similar or dissimilar) passages to the passage shown*: The user may want to view related information and so the system must be able to group similar and dissimilar passages to any given passage.
- *expanding summary length*: The summary may seem too short or too long and so the user must be able to modify the length to see more or less related information to the summary focus.
- *selecting summary cohesion criteria*: The ability to combine text passages in a useful manner for the reader. This may include ordering the passages by:
    - *time-line ordering*: Text passages ordered based on the occurrence of events in time.
    - *document ordering*: Include all text segments retrieved from the document with the highest ranking text segment. Next include all text segments from the document with the highest ranking text segment that is not already included in the summary. Continue this process for all text segments. Each group of text segments from a particular document are then ordered in the summary in the order that they appear in their source document.
    - *news-story principle (rank ordering):* Present the most relevant and diverse information first so that the readers gets the maximal information content for time spend reading the summary.
    - *topic cohesion*: Present the passages in the summary grouped together by passage similarity criteria.
- *creating new summaries*: The user must be able to create new summaries based on selecting text of the summary (the "more like this" and "not this type of information" criteria) to refine the summary information to that of interest.

Naturally, an ideal multi-document summary would include natural language understanding and generation to guarantee cohesive readable summaries [Radev and McKeown 1998, Mckeown et. al. 1999]. Although research has been ongoing since the late 1990s, much research is still needed in this area. Sentence ordering for cohesion has started to be addressed recently [Barzilay et. al 2002, Lapata 2003, Barzilay and Lee 2004, Barzilay and Lapata 2005, Bollegala et. al. 2005, Conroy 2006a, Madnami et. al. 2007].

The focus of our multi-document summarization is on outputting independent query relevant extractive summaries of an indicative nature. The summaries can sometimes appear to be somewhat disjointed due to the inclusion of extracts from various documents. These summaries and algorithms are discussed in further detail in Chapter 3 (system description) and Chapter 9 (experiments).

For newswire documents, as demonstrated in the introduction, it is necessary to employ anti-redundancy methods. To minimize repetitiveness, our system uses maximal marginal relevance (Sections 3.2 and 3.5) as well as methods for coverage and clustering. Inherent in our algorithm, is the ability to cover multiple *summary varieties,* including a form of update and novelty summaries. In fact, all of the summary varieties included in the *information space exploration* list can be created with the exception of comparative summaries.

Our summarization system at this stage does not address the important co-reference issue. Rather than attempt to substitute pronominal references, which have a F measure score of approximately 82% [Harabagiu et. al. 2001], our current system does not include any sentences that have a pronominal reference in the summary.

Chapter 3 Discussion of Our Summarization Systems

"An algorithm must be seen to be believed."
Donald Knuth

In this chapter we discuss the algorithm for our single document summarization system. We also present the metric maximal marginal relevance (MMR), designed to maximize novelty and minimize redundancy developed for both information retrieval and summarization. We conclude by presenting how we extend our single document summarization system and MMR for our multi-document summarization system.

3.1 Single Document Summarization System Description

Human summarization of documents, sometimes called abstraction, produces a fixed-length *generic* summary that reflects the key points which the abstractor deems important Section 2.1. In many situations, users are interested in facts other than those contained in the generic summary, motivating the need for *query-relevant* summaries. For example, consider a physician who wants to know about the adverse effects of a particular chemotherapy regimen on elderly female patients. The retrieval engine produces several lengthy reports (e.g., a 300-page clinical study), whose abstracts do not mention whether there is any information about effects on elderly patients. A more useful summary for this physician would contain query-relevant passages (e.g., differential adverse effects on elderly males and females, buried in page 211 of the clinical study) assembled into a summary. A user with different information needs would require a different summary of the same document.

Our approach to text summarization allows for both generic and query-relevant summaries by scoring sentences with respect to both statistical and linguistic features. This same process is used for both single and multi-document summarization. For generic summarization, a centroid query vector is calculated using high frequency document words and the title of the document. Each sentence is scored according to the following formula and then ordered in a summary according to rank order.

Equation 3-1: Summary Score:

$$Score(P_i) = w_q (Q \cdot P_i) + \sum_{s \in S} w_s (S_s \cdot P_i) + \sum_{l \in L} w_l (L_l \cdot P_i)$$

where
P_i is the passage or sentence
S is the set of statistical features
L is the set of linguistic features
Q is the query
w is the weights for the features in that set and $\sum_w w = 1$

These weights can be tuned according to the type of data set used and the type of summary desired. For example, if the user wants a summary that attempts to answer questions such as who and where, linguistic features such as name and place could be boosted in the weighting. (CMU and GE used these features for the Q&A section of the TIPSTER formal evaluation with some success [Mani et. al 1998]) Other linguistic features include quotations, honorifics, and thematic phrases [Mittal et. al. 1998].

Furthermore, different document genres can be assigned weights to reflect their individual linguistic features, a method used by GE [Strzalkowski et. al. 1998]. For example, it is a well known fact that generic summaries of newswire stories contain the first sentence of the article approximately 70% of the time [Goldstein et. al. 1999]. Accordingly, this feature can be given a reasonably high weight for the newswire genre.

We use this scoring method to weight various linguistic features for our genre oriented summarization systems (Chapter 7). These features include sentence position, passage term matches to lists of words focused on particular facets (such as opinion word lists, lists of movie genres, etc.) and numbers of named entities. The weights are determined empirically.

Statistical features include several of the standard ones from information retrieval: cosine similarity using term frequency- - inverse term frequency weights [Salton 1989]; pseudo-relevance feedback [Salton and Buckley 1990], query-expansion using techniques such as local context analysis [Xu and Croft 1996] or thesaurus expansion methods (e.g., WordNet [Feldbaum 1998]), the inclusion of other query vectors such as user interest profiles, and methods that eliminate text-span redundancy such as Maximal Marginal Relevance [Carbonell and Goldstein 1998].

3.2 Maximal Marginal Relevance

Maximal Marginal Relevance is a passage (or document) ranking method where each passage in the ranked list is selected according to a combined criterion of query relevance and novelty of information. The latter measures the degree of dissimilarity between the passage being considered and previously selected ones already in the ranked list. Of course, some users may prefer to drill down on a narrow topic, and others a panoramic sampling bearing relevance to the query. Accordingly, the best method is a user-tunable method; such as Maximal Marginal Relevance (MMR), which provides precisely such functionality, as discussed below.

Most modern IR search engines produce a ranked list of retrieved documents ordered by declining relevance to the user's query and summarization engines based on this techniques tend to include passages following the same principle. In contrast, "*relevant novelty"* is a potentially superior criterion. A first approximation to measuring relevant novelty is to measure relevance and novelty independently and provide a linear combination as the metric. The linear combination is called "*marginal relevance"* -- i.e. a document has high marginal relevance if it is both relevant to the query (created by the article in the case of generic summarization) and contains minimal similarity to

previously selected passages. The idea is to maximize marginal relevance in summarization as well as prior work in document retrieval [Carbonell and Goldstein 1998], hence the method is labeled "*maximal marginal relevance"* (MMR).

Equation 3-2: Maximal Marginal Relevance (MMR)

$$MMR = Arg \max_{P_i \in R \setminus S} \left[\lambda (Sim_1(P_i, Q) - (1-\lambda) \max_{P_j \in S} Sim_2(P_i, P_j)) \right]$$

Where
P is a passage in a document
Q is a query or user profile
θ is a relevance threshold, which can be degree of match or number of documents
R = IR(P, Q, θ), i.e., the ranked list of passages retrieved by an IR system, given P and Q and a relevance threshold θ, below which it does not retrieve passages;
S is the subset of passages in R already selected;
R\S is the set difference, i.e, the set of as yet unselected documents in R;
Sim_1 is the similarity metric used in passage retrieval and relevance ranking between passages and a query; and Sim_2 can be the same as Sim_1 or a different metric.

Given the above definition, MMR computes incrementally the standard relevance-ranked list when the parameter $\lambda = 1$, and computes a maximal diversity ranking among the documents in R when $\lambda = 0$. For intermediate values of λ in the interval [0,1], a linear combination of both criteria is optimized. Users wishing to sample the information space around the query, should set λ at a smaller value, and those wishing to focus in on multiple potentially overlapping or reinforcing relevant documents, should set λ to a value closer to 1.

For summarization of small documents, such as newswire articles, a value of λ (e.g. λ = 0.7 or 0.8) works especially well, providing relevance and diversity of information. This value of λ was determined empirically. For large documents, on the order of 10 pages or more, a small λ (e.g. $\lambda = 0.3$) allows for an overview of the information space, and then the user may wish to focus on the most important parts using a reformulated query.

For our single document summarization system, we set Sim_1 to be the Score in Equation 3-1 and Sim_2 to be the term frequency-inverse document frequency (tf-idf) cosine similarity information retrieval matching metric (Chapter 4 Equation 4-4) between the two passages. For newswire articles in the TIPSTER dry run and TIPSTER SUMMAC evaluation [Mani et. al. 1998], we use a value of $\lambda = 0.8$ to minimize redundancy. This was determined partly by manual evaluation as well as testing on our single document relevance corpus [Goldstein et. al. 1999].

Our process for creating single document newswire summaries is as follows:

1. Segment a document into sentences.
2. Form a query. For query relevant summaries, the query is either provided by a user or created based on a topic description. For generic summarization,

the query is composed of the concatenation of the title and the fifteen most common words in the document excluding stop words.

3. Identify the sentences relevant to the query, by using the tf-idf metric with a match between query and sentence.
4. Apply the MMR metric with $\lambda = 0.8$.
5. Keep choosing sentences until the desired number of sentences is reached.
6. Reassemble the selected passages into a summary document using one of the following summary-cohesion criteria:
   a. Document appearance order: present the segments according to their order of presentation in the original document. For this type of summary, include the first sentence, if it longer than a threshold (empirically set to 10 words) as it sets the context for the new article.
   b. News-story principle: Present the information in MMR-ranked order, i.e., the most relevant and the most diverse information first. In this manner, the reader gets the maximal information even if they stop reading the summary. This allows the diversity of relevant information to be presented earlier and any topics introduced may be revisited after other relevant topics have been introduced.
   c. Topic-cohesion principle: First group together the document segments by topic clustering (usng sub-document similarity criteria). Then rank the centroids of each cluster by MMR (most important first) and present the information, a topic coherent cluster at a time staring with the cluster whose centroid ranks highest. This type of ordering is especially relevant to long documents.

For news articles, we implemented query-relevant document appearance order as this article appeared to be the most comprehensible in informal human evaluations of summary quality.

For genre oriented summarization, where the summaries are focused on extracting particular information for each sentence (Chapter 7), we do not use anti-redundancy measures as there is usually very little redundancy present.

Due to the complexity of multi-document summarization as discussed in Section 2.5, the algorithms used for a summarization system must be expanded from that of single document summarization to appropriately address the variety of conditions necessary to meet user's information seeking goals. This will be discussed in Section 3.4.

3.3 MMR and Single-Document Summarization Evaluation

As part of the TIPSTER-III effort, CMU teamed with Carnegie Group Inc., using the MMR passage selection method for text summarization, and participated in both the dry run evaluation of summarization quality (six research systems) and the full SUMMAC evaluation of summarization (16 systems).

3.3.1 Single Document Summarization Evaluation Dry Run

The dry run and formal SUMMAC evaluation was based on TREC documents. The dry run evaluation consisted of generating adhoc summaries for sets of documents grouped under 5 topics. Each topic included a topic description which could be considered as a long query and 200 documents. Most of the documents were relatively short, containing under 50 sentences each.

Summaries were generated at two compression factors (CF = summary-length/document length): 10% and a longer "best" CF for each site. For our "best" CF, we simply selected a threshold over which sentences were selected. This turned out to be approximately 33%, 1/3 of the document length. Two tests were performed, one a categorization task and the other a relevance judgment task. In the categorization task, the humans attempted to categorize the summaries. In the relevance task, the humans judged the summaries as to whether they were relevant to the provided topic and the assigned judgment was compared to the original TREC relevance assessment, which was considered "ground truth" for the purpose of this evaluation.

The F1 scores for the query relevant task are shown in below in Table 3-1:

	CMU	Mean of 6 sites	Full documents
CF = 0.1	0.42	0.41	0.57
CF = "best"	0.53	0.49	0.57

Table 3-1: Dry Run Single Document Query Relevant Results for summary length. Compression Factor (CF) = 10% of document and the "best" summary submitted by the system.

The preliminary results did not determine a "winning" site – that was not the point of the evaluation, but rather they led to two useful observations:

1. Even short summaries contain sufficient information for tasks such as categorization and relevance judgments. Performance with shorter summaries was only slightly worse than longer ones or the full text. Time for task completion (not shown here) was much reduced with the shorter summary length.
2. Human relevance judgments are not consistent. If they were, then the F1 score would be 1.0 for the full documents. Given the variability of judgments the optimal performance for a summarizer in this task is 0.57. This indicates that a more discriminating evaluation is needed.

3.3.2 Single Document Summarization MMR Analysis

As an example of single-document summarization, we use one of the topics of this dry run evaluation: Topic 110 – Black Resistance Against the South African Government. The document set consisted of 200 news articles spanning a period of 1988-1991. The query was generated from the topic description by eliminating all stop words. Single

sentences were used as passages and a summarize size of 10% of the document sentence length was used (e.g., a 40 sentence document would have a summary length of 4 sentences).

In order to evaluate what the relevance loss for the MMR diversity gain in single document summarization, three assessors examined 50 articles from the 200 articles in the topic set and marked each sentence as relevant, somewhat relevant and irrelevant. The articles were also marked as relevant and irrelevant. The assessor scores were compared against the TREC relevance judgments provided for the topic. The assessors completely agreed in their relevant judgments 68% of the time.

The results are shown in Table 3-2 for document length percentages, 25% and 10%. Two precision scores were calculated (1) that of TREC relevance, plus at least one CMU assessor marking the document as relevant (yielding 23 documents) and (2) at least two of the three CMU assessors marking the document as relevant (yielding 15 documents). (For a discussion of precision, please refer to Section 4.2.1). From the scores, there is no significant differences between the $\lambda = 1.0$, $\lambda = 0.7$ or $\lambda = 0.3$ scores. This is often explained by cases where the $\lambda = 1.0$ article failed to choose a relevant sentence whereas the $\lambda = 0.7$ or $\lambda = 0.3$ reranking selected such a sentence. Further evaluation granularity (besides relevancy) would be required to distinguish the overall quality of these summaries.

The baseline contains the first N sentences of the document, where N is the number of sentences in the summary determined by the document compression factor.

		Sentence Precision	
Compression Factor	λ	TREC and CMU Relevant	CMU Relevant
0.1	1.0	0.78	0.83
0.1	0.7	0.76	0.83
0.1	0.3	0.74	0.79
0.1	Baseline	0.74	0.83
0.25	1.0	0.74	0.76
0.25	0.7	0.73	0.74
0.25	0.3	0.74	0.76
0.25	Baseline	0.60	0.65

Table 3-2: Precision Scores for Compression Factors (0.l and 0.25) and varying λs.

3.3.3 Single Document Summarization SUMMAC Evaluation

The TIPSTER SUMMAC Summarization Evaluation consisted of three tasks [Mani et. al. 1998]:

Adhoc task – the focus was on indicative user-focused summaries, where the summaries were tailored to a particular topic. The real-world activity represented is that of an analyst who is conducting full text searches using an information retrieval system and

who wants to determine quickly and accurately the relevance of a retrieved document. The topic is provided and the evaluation seeks to measure whether the full document is relevant based on a judgment of the topic focused summary. Given a document (either a summary of a full-text source – the subject isn't told which) and a topic description, the human subject determines whether the document is relevant to the topic. The Text Retrieval (TREC) document relevance judgments are used for ground truth. The documents were obtained from the TREC CDs 4 and 5 and all topics were drawn from those used in TREC. 20 topics were selected and for each topic a 50 document subset was created from the top 200 most relevant documents.

Categorization task – the focus is on generic summaries. The real world activity represented is that of a person manually routing or filtering information who must quickly decide whether a document contains information about any of several related topic areas. The topic is not known to the summarization system. Based on the summary, the human subject chooses a single category out of five categories to which the document is relevant or else chooses "none of the above". 10 topics were selected and 100 documents used per topic.

Question-answering task – the focus is on informative summaries. An intrinsic evaluation is performed which measure the extent to which a summary provides an answer to a set of questions that must be satisfied for a document to be judged relevant to a topic. The correct answers to these questions represent the "obligatory" aspects of a topic. The task is intended to support an information analyst responsible for writing a report on a topic and the challenge for the system is to understand the topic in relation to each document and to produce an "informative" summary that covers all obligatory aspects of the topic in as short a summary as possible. The questions were not provided in advance. Three topics of the subset of the 20 adhoc topics were used and 30 relevant documents per topic.

The results for Adhoc Accuracy (variable length) and (fixed length) are shown in Table 3-3 and Table 3-4.

Participant	Precision	Recall	F-Score	Compression
CGI/CMU	0.82	0.66	0.72	25.4
Cornell/SabIR	0.78	0.67	0.70	29.6
GE	0.78	0.60	0.67	17.4
LN	0.78	0.58	0.65	20.4
Penn	0.81	0.57	0.65	20.2
UMass	0.80	0.54	0.63	18.8
NMSU	0.80	0.54	0.63	27.4
TextWise	0.81	0.51	0.61	14.4
SRA	0.82	0.49	0.60	28.9
NTU	0.80	0.49	0.59	25.0
ISI	0.80	0.46	0.56	17.1

Table 3-3: Adhoc Accuracy (variable-length) by Participant.

Participant	Precision	Recall	F-Score
CGI/CMU	0.76	0.52	0.60
Cornell/SabIR	0.79	0.47	0.56
UMass	0.81	0.47	0.56
GE	0.77	0.45	0.55
LN	0.81	0.45	0.55
Penn	0.76	0.45	0.53
TextWise	0.79	0.41	0.52
NMSU	0.80	0.40	0.52
SRA	0.79	0.37	0.48
ISI	0.82	0.36	0.47
NTU	0.82	0.34	0.46

Table 3-4: Adhoc Accuracy (fixed-length) by Participant.

Our summarization system was designed for query relevant summarization and not generic summarization. The reader is referred to the TIPSTER Summac Evaluation Report [Mani et. al. 1998] for details of the categorization task.

For the question answering task, two accuracy measures were defined to capture the accuracy of a summary in answering questions. ARL (Answer Recall Lenient) and ARS (Answer Recall Strict). ARL = (n1 + (0.5 * n2))/n3, where n1 is the number of Correct answers in the summary, n2 is the number of partially correct answers in the summary and n3 is the number of questions answered in the key. ARS = n1/n3.

A third measure ARA (Answer Recall Average) was defined as the average of ARL and ARS and was used a a succinct way of reporting the results.

The results for the seven participants are shown in Table 3-5. Each system had a different compression factor.

Participant	ARS	ARL	ARA
CGI/CMU	0.66	0.75	0.71
Cornell/SabIR	0.62	0.71	0.66
GE	0.55	0.66	0.60
NMSU	0.54	0.65	0.59
SRA	0.44	0.53	0.49
ISI	0.36	0.45	0.40
Penn	0.30	0.38	0.34

Table 3-5: Answer Recall by Participant.

3.4 Summary Compression Rates

For single document newswire articles, many human created (by journalists) short summaries are available and these typically range from 3-5 sentences or created by

aligning text spans to appropriate sentences [Banko et. al. 1999, Goldstein et. al. 1999]. There are also model summaries that were created for the TIPSTER Question Answering task. The statistics of some single document summary collections are shown in Table 3-6:

Property	**QA Summaries**	**Reuter Summaries**	**LA Times Summaries**
Task	Q&A	generic summaries	generic summaries
Source	TIPSTER (TREC)	human => extracted	human => extracted
Document Features			
# of docs	128	1000	1250
average sent/doc	32.1	23.1	27.9
Summary Features			
% of doc length	19.6%	20.1%	20.0%
Includes 1st sentence	61.7%	70.5%	68.3%
Average size (sent)	5.8	4.3	3.7

Table 3-6: Single Document Data Set Characteristics.

Table 3-6 shows that single document summaries tend to have a compression of 20% of the document length (sentences). Using these statistics, we can compute various multi-document sentence compression factors. For our purposes, we use the number of documents in the various TREC document cluster sets that have been provided for use in summarization – these range from 50-200 documents. If we use 23 sentences as a minimum number of sentences and 32 as a maximum per document (based on Table 3-6), we can compute various multi-document summarization compression factors, shown in Table 3-7.

	Cluster Size (Using 23 sentences and 32 sentences)		
Summary Size	50 documents	100 documents	200 documents
10 sentences	0.6% to 0.8%	0.3% to 0.4%	0.2%
One article (30 sent)	1.9% to 2.6%	0.9% to 1.3%	0.5% to 0.7%

Table 3-7: Multi-Document Summaries Compression of Full Document Length. Two average document sentence lengths - 23 sentences and 32 sentences were used to compute the number of sentences per cluster and account for the ranges depicted.

The sentence compression for multi-document summaries ranges from 0.2% (10 sentence summary for a 200 document cluster) to 2.6% (30 sentence summary for a 50 document cluster) – two orders of magnitude to one order of magnitude different that single document summarization compression. Even a 10 sentence summary for a 10 document cluster would result in a sentence compression of less than 5%, as compared to the single document summary sentence compression factor of 20%. Since the sentence compression ratio is such a small percentage of the document cluster size, as previously mentioned, it is important to carefully select the sentences in the summary to maximize information content in the space available.

3.5 Multi-Document System Design

This section discusses our multi-document summarization system. As a first pass, we are focusing only on indicative summaries, created by sentence extracts from the documents. Ideally our system would have the user interface abilities mentioned in Section 2.5. However, we did not construct such a system and for purposes of multi-document evaluations in this thesis, we focus only on ten sentence summary outputs. Our multi-document summarizer assumes a document cluster of topically related documents, as produced by a topic detection and tracking system [Yang et. al. 1998, Allan et. al. 1998], from a search on a data collection [Salton 1983] or from topically related documents such as those in the TREC collections. Our summarizer focuses on coverage, anti-redundancy, and the information space exploration summary criteria mentioned in Section 2.5 with the exception of Comparative Summaries. As discussed in the previous section – se want to carefully use the summary space available and thus maximize relevance and novelty.

Accordingly, we want to tailor the MMR metric for multi-document summarization - the Maximal Marginal Relevance Multi-Document (MMR-MD) metric is defined in Equation 3-3. For a multi-document collection, we want have the potential to diversify among the documents as well as the passages. For example, we might want to ensure selections from unique documents and the weights in our formula can be adjusted accordingly. We have expanded the metric to allow through the weights to give a higher score to passages that contain information that seems to occur in many documents (calculated by creating passage clusters and counting the number of passages in a cluster). Once such a passage is chosen, if it belongs to multiple passage clusters, the passage can penalized through the weights other passages from the clusters.

To cluster the passages, as a first pass we are using the cosine similarity score since this computation is already required for the MMR and MMR-MD metric. Passages are in the same cluster if the similarity scores among all pairs all above a certain threshold. For example, if we have the following four sentences:

1. Massimo D'Alema is set to resign as Italy's prime minister today, initiating a formal government crisis at the end of which he hopes to form a new centre-left administration.
2. ROME (AP) - Premier Massimo D'Alema on Saturday night announced he was resigning, contending he has enough parliamentary support to put together a new, stronger government, Italian news agencies reported.
3. Italy looks set for a quick change of government following the resignation of Prime Minister Massimo D'Alema.
4. While D'Alema was meeting with Ciampi, seven of the 11 parties in the premier's coalition approved a declaration to support him.

Cluster 1 contains sentences 1, 2 and 3 (all have the D'Alema's resignation in common) and Cluster 2 has sentences 2 and 4 (parliamentary support in common). Thus if sentence 2 was picked as a passage, any other passage from Cluster 1 or Cluster 2 could be penalized for containing similar information.

Equation 3-3: Maximal Marginal Relevance Multi-Document (MMR-MD)

$$MMR-MD = Arg \max_{P_i \in R \setminus S} \left[\lambda(Sim_1(P_{ij}, Q, C_{ij}, D_i, D) - (1-\lambda) \max_{C_j \in S} Sim_2(P_{ij}, P_{nm}, C, S, D_i)) \right]$$

$$Sim_1(P_{ij}, Q, C_{ij}, D_i, D) = w_i \cdot (P_{ij} \cdot Q) + w_2 * \text{cov}erage(P_{ij} \cdot C_{ij}) + w_3 * content(P_{ij}) + w_4 * time\_sequence(D_i, D)$$

$$Sim_2(P_{ij}, P_{nm}, C, S, D_i) = w_a * (P_{ij} \cdot P_{nm}) + w_b * clusters\_selected(C_{ij}, S) + w_c * (documents\_selected(D_i, S)$$

$$\text{cov}erage(P_{ij}, C) = \sum_{k \in C_{ij}} w_k * |k|$$

$$content(P_{ij}) = \sum_{W \in P_{ij}} w_{type}(W)$$

$$time\_sequence(D_i, D) = \frac{timestamp(D_{\max\_time}) - timestamp(D_i)}{timestamp(D_{\max\_time}) - timestamp(D_{\min\_time})}$$

$$clusters\_selected(C_{ij}, S) = \left| C_{ij} \cap \bigcup_{v,w:P_{vw} \in S} C_{vw} \right|$$

$$documents\_selected(D_i, S) = \frac{1}{D_i} * \sum_w [P_{iw} \in S]$$

where
Sim_1 is the similarity metric for relevance ranking;
Sim_2 is the anti-redundancy metric;
D is a document collection;
P is the passages from the documents in that collection (e.g., P_{ij} is passage j from document D_i;
Q is a query or user profile;
R = IR(D, P, Q, θ), i.e., the ranked list of passages from documents retrieved by an IR system, given D, P, Q and a relevance threshold θ, below which it does not retrieve passages (θ can be degree of match or number of passages);
S is the subset of passages in R already selected;
R\S is the set difference, i.e, the set of as yet unselected passages in R;
C is the set of passage clusters for the set of documents;
C_{vw} is the subset of clusters of C that contains passage P_{vw};
C_v is the subset of clusters that contain passages from document D_v;
|k| is the number of passages in the individual cluster k;
$|C_{vw} \cap C_{ij}|$ is the number of clusters in the intersection of C_{vw} and C_{ij};
w_i are weights for the terms, which can be optimized, e.g., $w_{type\text{-}genre}$ is a weight for a particular word for a particular genre; and $w_{position\text{-}genre}$ is a weight for a passage position for a particular genre;

W is a word in the passage P_{ij};
type is a particular type of word, e.g., city name;
$w_{type\text{-}genre}$ is a weight for a particular word for a particular genre;
$w_{position\text{-}genre}$ is a weight for a passage position for a particular genre;
$|D_i|$ is the length of document i.

Using MMR-MD we can adjust the weights and define Sim_1 and Sim_2 to cover the creation of summaries according to the properties in Section 2.5[4]. The weights and value of λ were determined empirically.

For Sim_1, the first term uses the cosine similarity metric to compute the similarity between the query and passages in the document. The second term computes a *coverage* score for the passage based on whether the passage is in one or more clusters and the size of the cluster. The third term reflects the information content of the passage by taking into account both statistical and linguistic features for summary inclusion (such as query expansion, position of the passage in the document and presence/absence of named-entities in the passage). This term also has weighting according to the genre characteristics of the set. The final term indicates the temporal sequence of the document in the collection allowing for more recent information to have higher weights. Thus, to weight the last document as the highest, w_4 would be assigned a very high value, ensuring that all sentences are selected from the last document. To minimize redundancy one would assign λ a small value (to maximize diversity), and to allow redundancy (i.e., allow the user to create a summary exploring the similarity in the data set for a query), one would assign λ a value of 1.0 - no redundancy elimination.

For Sim_2, the first term uses the cosine similarity metric to compute the similarity between the passage and previously selected passages. (This helps the system to minimize the possibility of including passages similar to ones already selected.) The second term penalizes passages that are part of clusters from which other passages have already been chosen. The third term penalizes documents from which passages have already been selected; however, the penalty is inversely proportional to document length, to allow the possibility of longer documents contributing more passages. These latter two terms allow for a fuller *coverage* of the clusters and documents.

We segment the documents into passages (sentences, n-sentence chunks or paragraphs) and apply the MMR-MD metric. In order to minimize computation time of an algorithm of complexity $O(n^2)$, where n is the number of passages, we only apply the MMR-MD metric to a selected number of passages. The number of passages chosen depends on the user's selected length for the output summaries. Summaries are output according to the users' presentation choices.

3.6 Multi-Document Update Summaries

As an example of how our system can operate, let us consider the creation of update summaries. Using a set of 10 documents from Yahoo collected on the first day of the Egypt Air crash (from multiple sources, some of which contributed more than one article), we produce a generic summary for this set (Figure 3-1)) by using as a query the titles of the documents and 15 high frequency words within the document collection. Since we use the titles to compute this summary - this summary tends to produce a

[4] Sim_1 and Sim_2 as previously defined in MMR for single-document summarization contained only the first term of each equation.

``flavor" of the document set as a whole. This summary contains some redundancy (sentences 2,6,8,10 mention the crash of Egypt Air although they also contain other information as well). This summary has five sentences in common with the three summaries of the human summarizers (sentences 3,4,5,8,10). Two of the human summarizers have one sentence in common for this set.

Now let us consider an update set on the topic of EgyptAir spanning the dates of 12/99-8/00. The resultant document collection update summary (DUS) (query from title and 15 high frequency words) is shown in Figure 3-2. We can also create a specific update summary (SUS) by using the initial document set summary as a query (Figure 3-3). As compared to the initial EgyptAir Summary (Figure 3-1) – "feared dead" has now been confirmed as "dead". There are also updates on the results of the investigation. If we compare the DUS (Figure 3-2) to the SUS (Figure 3-3), there is only one sentence in common, article 5 sentence 1. The SUS focuses more on the investigation, crash and cause of crash of EgyptAir 990 as did the original summary from which the query was taken, whereas the US also mentions a possible suit, insurance payouts, EgyptAir and Egypt officials strong reaction to a possible suicide as well as an EgyptAir pilot who claimed to have information about the crash.

If we compare the set of three human generated ten sentence text extract multi-document summaries for this set to the DUS (Figure 3-2), two people have one sentence in common with DUS - article 3 sentence 9, and one person has article 2 sentence 7 in common (a total of 3 sentences). If we compare the human generated summaries for this set to the SUS (Figure 3-3), one person has article 2 sentence 17 in common, another has three sentences article 8 sentences 1 and 3 as well as article 3 sentence 10, and the third person (shown in Figure 3-4) has article 7 sentence 1 in common (for a total of 5 sentences). The human multi-document summaries have 3 sentences in common with each other.

In the next two chapters, we discuss metrics that can be used for evaluating summarization systems and the formal evaluations of summarization systems.

1. **991031 4 4** EgyptAir flight 990 carrying 214 people -- 199 passengers and 15 crew -- took off from JFK airport around 1:20 a.m. (0620 GMT) after an unexplained two-hour delay, according to airport officials.
2. **991031 5 1** CAIRO, Oct 31 (AFP) - Egyptian officials were quick to rule out the possibility of terrorism after an EgyptAir passenger plane bound for Cairo crashed into the waters off the US coast Sunday, with all 214 people aboard feared dead.
3. **991031 6 1** BOSTON, Oct 31 (AFP) - More than 200 people aboard an EgyptAir Boeing 767 were feared dead Sunday after the Cairo-bound passenger jet disappeared shortly after taking off from New York's John F. Kennedy airport.
4. **991031 6 10** An anti-terrorism task force to investigate the disaster was set up in New York, including New York City police, the FBI and State Department officials found no immediate signs of criminal activity, New York Port Authority police inspector Anthony Infante told a news conference in New York.
5. **991031 6 37** The disappearance of the EgyptAir plane comes just three years after TWA flight 800 -- a Boeing 747 -- exploded in mid-air off the east coast.
6. **991031 7 1** CAIRO, Oct 31 (AFP) - Cairo airport echoed with the screams of hysteria and horror Sunday as distraught relatives learned their loved ones were aboard EgyptAir Flight 990, which crashed off the US coast with all 214 aboard feared dead.
7. **991031 8 1** NANTUCKET, Mass., Oct. 31 - Coast Guard ships are pouring over a 49-square-mile area for any survivors, bodies or debris from the EgyptAir flight that went down this morning 60 miles south of Nantucket.
8. **991031 9 1** Oct. 31 - Searchers combed the seas off Nantucket Island today looking for bodies and wreckage after an EgyptAir jetliner with 217 people aboard, including dozens of Americans, plunged 14,000 feet in 16 seconds and crashed into the Atlantic Ocean.
9. **991031 9 34** On Nantucket, the focus is on getting Coast Guard helicopters refueled and continuously going out to the crash site to search the debris field.
10. **991101 10 1** NANTUCKET, Massachusetts (CNN) -- An EgyptAir plane with 217 people on board crashed at sea early Sunday off the island of Nantucket, Massachusetts, en route from New York to Cairo, Egypt.

Figure 3-1: Multi-document summary EgyptAir, query=titles+15 high freq words with λ = .6 (minimal anti-redundancy), time line ordering: Sentence Number, TimeStamp, Document Number, Sentence Number in Document, Sentence.

1. **991208 1 1** Egyptian pilots have threatened a US television station with legal action over a news report into the theory that EgyptAir Flight 990 was brought down by a co-pilot.
2. **000121 2 1** Investigators examining the debris from EgyptAir 990 are still working on the theory that the aircraft was deliberately crashed into the Atlantic Ocean, according to reports from the United States.
3. **000121 2 7** EgyptAir dismissed reports that the crash was connected with the suicide of the co-pilot, citing problems with the aircraft's tail apparatus.
4. **000123 3 1** EgyptAir is to offer the families of the 217 people who died in a plane crash off the United States a total of up to \$116m in insurance payments, it has been reported.
5. **000123 3 9** The cause of the tragedy so far remains a mystery but speculation continues that one of the pilots deliberately crashed the plane.
6. **000209 4 1** The United States is sending two investigators to London to interview an EgyptAir pilot who says he has information about the crash of Flight 990 into the Atlantic last year.
7. **000209 4 5** EgyptAir vice president for operations, Hassan Musharafa, said Mr Taha was ``one of 500 EgyptAir pilots and had no access to information about the crash".
8. **000413 5 1** US investigators into the crash of EgyptAir flight 909 say they have not ruled out the possibility that the plane was crashed deliberately, and have called for cockpit cameras to be installed to help future investigations.
9. **000413 5 5** Egyptian officials have angrily rejected suggestions that a co-pilot deliberately crashed the plane.
10. **000817 10 7** Egypt has always strongly refuted an earlier theory that a co-pilot, Gameel al-Batouti, may have deliberately crashed the plane.

Figure 3-2: Multi-document summary on EgyptAir Update Set (DUS), query=titles+15 high freq words with λ = .6 (minimal anti-redundancy), time line ordering.
Sentence Number, TimeStamp, Document Number, Sentence Number in Document, Sentence.

1. **991208 1 2** All 217 people aboard the EgyptAir flight to Cairo died when the Boeing 767 crashed into the Atlantic Ocean, less than an hour after taking off from New York's Kennedy airport.
2. **000121 2 16** Last month, a Navy submarine mapping wreckage from the crash located the plane's remaining engine on the sea bed.
3. **000121 2 17** According to the NTSB, the view of the submerged wreckage showed damage indicating the engine was generating little or no power when the aircraft hit the water, supporting evidence from the flight data recorder that the engines were shut off during the crash.
4. **000123 3 7** EgyptAir flight 990 crashed off the coast of Massachusetts on 31 October en route from New York to Cairo.
5. **000123 3 10** Officials at the National Transportation Safety Board say investigations have found no evidence that would suggest an alternative reason for the crash such as an explosion or mechanical failure.
6. **000413 5 1** US investigators into the crash of EgyptAir flight 990 say they have not ruled out the possibility that the plane was crashed deliberately, and have called for cockpit cameras to be installed to help future investigations.
7. **000722 7 1** Nearly nine months after an Egyptian plane plunged into the Atlantic Ocean killing all 217 people aboard, the US-led investigation is preparing to issue a report into the disaster.
8. **000811 8 1** Air safety officials in the United States have published details of last year's crash of EgyptAir flight 990, but are not ready to say what caused it.
9. **000811 8 3** The Boeing 767 airliner plunged into the Atlantic Ocean on 31 October off the Massachusetts coast on a flight from New York to Cairo, killing all 217 people on board.
10. **000817 10 2** All 217 passengers on board the Boeing 767 were killed when the plane crashed off the east coast of the United States.

Figure 3-3: Multi-document update summary on EgyptAir Update Set (SUS), query=previous summary with λ = .6 (minimal anti-redundancy), time line ordering. Sentence Number, TimeStamp, Document Number, Sentence Number in Document, Sentence.

1. **000722 7 1** Nearly nine months after an Egyptian plane plunged into the Atlantic Ocean killing all 217 people aboard, the US-led investigation is preparing to issue a report into the disaster.
2. **000722 7 27** The new report into the crash is almost certain to revive what has become a major controversy.
3. **000811 8 4** The Americans suspect it was a deliberate action by the co-pilot, Gameel Batouty, while the Egyptians believe it was a technical fault.
4. **991208 1 7** The sources, speaking on condition of anonymity, said relief co-pilot Gamil al-Batouty was heard asking or offering to take the controls half-an-hour after Flight 990 left New York for Cairo, CNN said.
5. **991208 1 8** He was then heard saying ``I put my faith in God's hands", which he uttered "multiple times" before a series of unexplained manoeuvres.
6. **000722 7 15** The suicide theory, which strained usually close relations between the two countries, is still hotly contested.
7. **000817 10 1** EgyptAir says it is almost completely sure that mechanical problems, not pilot suicide, caused the crash of one of its airliners last October.
8. **000817 10 3** Chairman Mohammed Fahim Rayan spoke out after calls from the US Federal Aviation Administration for extra inspections of the elevator controls of 767s, which are used to point the planes up or down.
9. **000817 10 4** Mr Rayan said: ``We are 99% sure that there was something (wrong) in the elevator system."
10. **000811 8 11** ``I want to make it perfectly clear that no determination as to the cause of this crash has been made," Mr Hall said.

Figure 3-4: Multi-document human summary EgyptAir Update, most readable ordering.
Sentence Number, TimeStamp, Document Number, Sentence Number in Document, Sentence.

Chapter 4 Evaluation Metrics for Summarization Systems

One can judge from experiment, or one can blindly accept authority. To the scientific mind, experimental proof is all important and theory is merely a convenience in description, to be junked when it no longer fits. To the academic mind, authority is everything and facts are junked when they do not fit theory laid down by authority.

"Doctor Pinero" in *Life-Line* (1939) by Robert Anson Heinlein

Evaluations can be categorized into two types. *Extrinsic* evaluations measure how well a system performs in some task. *Intrinsic* evaluations are used for evaluating quality. For summarization, an extrinsic evaluation could measure whether reading a summary is suitable for replacing reading a full document for determining document relevance or for obtaining answers. An intrinsic evaluation could evaluate the quality of the summary, for example, is it cohesive, coherent, grammatical, how much redundancy does it contain and does it suggest incorrect implications? Intrinsic evaluations also measure how well the summary matches human gold standard summaries, such as how much coverage does it have of the semantic content of these human summaries.

There are many different types of summarization systems – headline, keyword, indices, single document summaries, web page summaries, multi-document summaries. Reviews can also be thought of a summary which contains the authors' opinion. The field of question answering also intersects with the field of summarization. Summaries can be designed for the purpose of answering questions or providing topic related information, as was evaluated in some of the Document Understanding Conference (DUC) evaluations [DUC]. Summaries can also compile facts produced by question answering systems – this will be evaluated in the November 2008 NIST Text Analysis Conference (TAC). Question answering systems have previously been evaluated in the NIST Text REtrieval Conference (TREC) Question Answering (QA) track [TREC-QA]. QA systems have been evaluated on finding answers to basic facts – such as "how high is the Eiffel tower". They have also been evaluated on extracting "other" information nuggets about a person, place, etc from the documents that were not asked about in previous questions. NIST assessors assigned a value to these items - "vital" or "okay". Another evaluation in previous QA Tracks were "definitions" of a person, place, thing. These can be thought of as mini-summaries. The output "definitions" of people overlapped with what might be considered a biographical summary.

Recently, in 2006, with the advent of the DARPA Global Autonomous Language Exploitation (GALE) program [GALE], the notion of "distillation" arose. An information distiller seeks to capture the query relevant information nuggets of a set of documents. For GALE, queries are expressed in a template form. These nuggets are produced from input text in foreign languages as well as speech data that has been transcribed and translated. A thorough distiller would extract all relevant information [White et. al 2007]. For our purposes we can think of the output of a distiller as potential

items to input to a summarizer, which would want to organize and present the information in a usable, functional manner.

In this chapter, we focus on metrics and evaluations for summarization beyond the keyword and headline. We provide an overview of the metrics used for evaluating summarization tasks: Precision/Recall/F1 [Salton 1989], Relative Utility [Drago and Tam 2003], Human Judgments [DUC], Factoids (Teufel & von Halteren), ROUGE and variants [Lin and Hovy 2003, Lin 2004], The Pyramid Method [Nenkova and Passonneau 2004]. We also cover some metrics related to summarization: BLEU [Papineni et. al 2002], POURPRE [Lin and Demmer-Fushman 2005], Nuggeteer [Marton and Radul 2006], METEOR [Lavie 2004]. We conclude by discussing issues with gold standard summaries that are used for these metrics.

4.1 Metrics Overview

There are many different metrics by which one can measure summarization quality, some more relevant to certain types of summarization systems than others. For our purpose, we discuss metrics relevant to summaries that are longer than one sentence. Metrics for the classes of short words summaries, which include "headline" summaries, topical keyword summaries or index type summaries, although relevant to the field of summarization are left for other endeavors.

In the case of multi-sentence summaries, metrics can differ based on whether the summary is an extractive summary or a generated summary. Extractive summaries can often be compared to the "sentences" from which they were extracted, such as summaries of news articles. Extractive summaries also include the case of "telegraph style" summaries [Strzalkowski et. al. 1998, Hori et. al. 2003] – where words of the sentences deemed unnecessary to convey the meaning are eliminated from the message. In the past, such mechanisms were employed for cost saving.

In the case of generated summaries, comparing the components of a generated sentence to the areas from which it originated can be more difficult due to the synonyms, metaphors and the variations that occur in written (and spoken language). Accordingly, relevant metrics can be categorized as useful for:

(1) extractive summaries
(2) generated summaries, often referred to as abstracts
(3) both extracted and generated summaries

Summarization quality can then be assessed by these type of metrics or others which compare the content of the machine produced summary to a human created one or some measure which indicates the overall summarization quality. The latter includes items such as grammaticality, lack of redundancy, appropriate context, appropriate connectivity in the flow – many of the same criteria with which we would judge a human written abstract or article.

Table 4-1 shows the metrics that have been used or could potentially be used for summarization or the related field of question answering. All of these methods require the creation of a good standard data set. Some, such as the Pyramid Method [Nenkova and Passonneau 2004] and Nugget Pyramids [Lin 2005, Dang 2007], require follow-on human involvement using the gold standard data set. In the following subsections, we briefly describe these metrics as they pertain to summarization.

METRIC	EXTRACTIVE	GENERATED	Used in Formal Evaluations
Sentence Precision-Recall	x		
Relative Utility	x		
Cosine Similarity	x	x	
ROUGE & Variants	x	x	x
PYRAMID			x
Nugget Pyramids			TREC QA 2006-2007
NUGGETEER			QA definition oriented
BLEU	x	x	MT
METEOR			Proposed for MT
Human Summary Quality Assessments	x	x	x
Human Summary Responsiveness	x	x	x

Table 4-1: Summarization System Metrics and Related Metrics

4.2 Metrics for Extractive Summaries

For extractive summaries, we can either measure how well extracted summary sentences match a selected group of relevant sentences for the document or how well the sentences match a human abstract. Metrics for the latter are in the next subsection.

4.2.1 Precision-Recall

The simplest metric for extractive summaries is to measure the potential relevance of the summary by the relevance of its individual sentences. This can be done by marking a sentence as relevant or not relevant to some information seeking goal. A variant of this theme is to use a third category, somewhat or partially relevant, which is often easier for humans. This was the approach used to evaluate an early version of our single document [Goldstein et. al 1999]. The resultant summaries can then be evaluated by precision and recall metrics as proposed by Salton for information retrieval [Salton 1989].

Equation 4-1: Precision

$$\text{Precision P} = \frac{\textit{system sentences which match human selected sentences.}}{\textit{sentences selected by system}}$$

Equation 4-2: Recall

$$\text{Recall R} = \frac{\textit{system sentences which match human selected sentences.}}{\textit{sentences selected by humans}}$$

Equation 4-3: The F Measure

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

Where β can be adjusted to affect precision or recall.
When β = 1, the F-measure is the harmonic mean of precision and recall – precision and recall have equal weight. For other applications, such as in the NIST TREC question answering track, for the "other" questions (e.g., definition), which often represent facts that are not a simple answer, a higher β, i.e., β=3, is used to emphasize recall over precision. For nugget precision, a measure based on length is used. For summarization, a high value of β is also used to emphasize recall over precision. Systems may choose sentences that make good summary sentences and even convey the same information but were not selected by humans, especially in the case of newswire multi-document summarization. Having more human gold standard summaries for comparison can result in better evaluations.

When there are multiple judges selecting good relevant sentences, one can use a *strict* version of precision recall, such as all judges must mark the sentence as relevant. In a *relaxed* version, all judges do not have to mark the sentence as relevant for it to count. There are many variants, e.g., at least one judge marking the sentence as relevant, the majority of judges marking as relevant, or some combination of relevant and somewhat relevant if a three way selection (relevant, somewhat relevant, not relevant) is used.

For multi-document summarization, when humans are asked to produce an extract summary or when asked to select the most informative sentence can often disagree as to what are the top or best sentences. Their chosen sentences can be identical, but represent sentences from different documents due to the information redundancy present in the cluster. Relative Utility is a metric that attempts to address this issue.

4.2.2 Relative Utility

In order to provide more granularity and comparison with human marking, in 2003 Dragomir Radev [Radev and Tam 2003] proposed a metric which used judges' sentence assessments relative to other judges' to compute a score for extract summaries. Each human judge ranks each sentence in each document with a score of 0-10 in a multi-document cluster. High sentence scores mean a good summary sentence. A sentence obtains a final score based on a combination of all the judges' assessments. E.g., if there are two judges, and one gives the sentence an 8 and the other 9 and the *relative* score is (8/9). Each summary can then be assigned a score based on its choice of sentences. Relative Utility can also be normalized based on random summaries of the cluster. Although this appears to be a good approach for extract summarization, due to the

necessity of multiple judges and the marking of all sentences in the corpus, including equivalence, this amount of tagging results in a highly expensive effort.

In the speech summarization domain, using SWITCHBOARD spoken dialogue, Zhu and Penn showed that the relative utility metric (with 3 judges) did not distinguish five system summaries better than Precision/Recall [Zhu and Penn 2005]. However, relative utility was designed for multi-document summarization with a high level of redundancy and Zhu and Penn's evaluation was for single documents.

Both precision and recall and relative utility work for extract summaries. In the next section, we discuss metrics that can be used when the humans generate free text summaries, not extracts, often referred to as abstracts.

4.3 Metrics for Generative or Abstractive Summaries

With the advent of the NIST Document Understanding Conferences [DUC], which began with a kickoff in 2000, evaluations began to use human generated summaries as gold standards. Because these were not extracted summaries but summaries composed by the human, they require different type of evaluation metrics – those that can compare abstracts and extracts and/or partial extracts (due to summary length restrictions). Such metrics include cosine similarity, variants of ROUGE and the PYRAMID method.

4.3.1 Cosine Similarity

In the late 1990s and early 2000s, metrics revolved around a bag of words and word overlap. Cosine similarity was one such approach. Cosine similarity was originally used to compare documents in text mining. It can be used in the same manner to compare summaries. The equation is given below:

Equation 4-4: Cosine Similarity

$$\theta = \arccos \frac{A \bullet B}{|A||B|} \qquad \text{and Sim(A,B)} = \text{cosine } \theta$$

For text matching, including summarization, the vectors A and B are usually the term frequency - inverse document frequency (tf-idf) vectors of the documents, defined below.

When angle $\theta = 0$, this means the two texts are exactly equal. When the angle $\theta = \pi/2$ this means that the two documents (or summaries) are independent. The values in between 0 and $\pi/2$ represent the degree of similarity or dissimilarities between the two texts. A and B are vectors. They can represent the term frequency, the normalized term frequency (either for the document or for the sentence) or the term-frequency-inverse document frequency (tf-idf) [Salton 89].

The tf-idf is a weight used to evaluate how important a word is to a document (or summary) in a collection or corpus. The importance increases proportionally to the number of time that a word appears but is offset by the corpus frequency of that particular word (the inverse document frequency – idf). Thus, the inverse document frequency can be thought of as measuring the general importance of the word in the documents. It is obtained by dividing the number of all documents by the number of documents that contain the term and then taking the logarithm of the quotient. There are many variants of tf-idf that can be used, including whether or not the term frequency is normalized. (Normalization prevents a bias towards longer documents).

Equation 4-5: Normalized Term Frequency

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

Where n_{ij} is the number of occurrences of the term in the document (summary) d_j and the denominator is the number of occurrences of all terms in the document d_j.

Equation 4-6: Inverse Document Frequency

$$idf_i = \log \frac{|D|}{|\{d_j : t_i \in d_j\}|}$$

where $|D|$ is the total number of documents in the corpus and $|\{d_j : t_i \in d_j\}|$ is the number of documents where the term t_i appears (that is for all $n_{i,j} \neq 0$).

Equation 4-7: Term Frequency – Inverse Document Frequency

$$\text{tfidf: } \text{tfidf}_{i,j} = \text{tf}_{i,j} \cdot \text{idf}_i$$

This methodology can also be used to compare summaries, where the summary is treated as a document. It also can be used to compare sentences in summaries, where a sentence in the summary (or document) takes the role of the document in Equation 2.

Since the cosine similarity is essentially a bag of words approach, measuring similarity between the words of two documents, there was a need for enhanced automated metrics for summarization – ideally taking into account word order as well as information similarity. Such metrics would need to be correlated with the human summary quality judgments that were becoming available from the formal DUC evaluations that started in 2001. This led to the development of ROUGE.

4.3.2 ROUGE

In 2003, the metric ROUGE was proposed by Chin-Yew Lin and Eduard Hovy at ISI [Lin and Hovy 2003]. It was roughly based on the machine translation metric BLEU [Papineni et. al. 2002], but unlike BLEU which focuses on precision, this metric ROUGE focused on recall and measure word overlap in sequences. ROUGE was found to

correlate fairly well with human evaluations unlike BLEU. With the availability of more data as the DUC evaluations continued, results showed that ROUGE-2 and ROUGE-SU4 (skip unit of 4) gave the best correlations with human judgments [Lin 2004].

There are several variants of ROUGE: [Lin 2004]

- ROUGE-N: N-gram based co-occurrence statistics
- ROUGE-L: Longest Common Subsequence (LCS) based statistics
- ROUGE-W: Weighted LCS-based statistic that favors consecutive LCSes
- ROUGE-S: Skip-bigram based co-occurrence statistics, where skip-bigram is defined as any pair of words in their sentence order, allowing for arbitr ary gaps
- ROUGE-SU: Skip-bigram plus unigram based co-occurrence statistics.

ROUGE-SU is important so that a sentence such as S1 "the gunman who was killed by police" has a matching score with S0: "police killed the gunman". If S0 is the reference summary, using ROUGE-S, since words have to be in sentence order, S1 would have a ROUGE-S score of zero. ROUGE-SU can be thought of as putting a begin-of-sentence marker (BSM) at the beginning of the candidate and reference sentences. Then the skip bigrams "BSM police", "BSM killed" and "BSM gunman" all match. If the option of stopwords (common frequency words such as "the", "a", etc.) is included, "BSM the" also matches. Stopwords are not removed in the NIST DUC evaluations, whose official metrics have been ROUGE-2 and ROUGE-SU4. Note that although ROUGE-2 and ROUGE-SU4 were shown to have the best correlations with human judgments for newswire corpora, this result may not hold true for other genres.

The formula for ROUGE-N and ROUGE S are shown in Equations 5 and 6 respectively. Readers interested in the other variants of ROUGE are referred to Chin-Yew Lin's papers [Lin and Hovy 2003, Lin 2004].

Equation 4-8: ROUGE N

$$\text{ROUGE-N recall} = \frac{\sum_{S \in \{\text{Re}\, ferenceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{\text{Re}\, ferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

Where n stands for the length of the n-gram, $gram_n$ is the actual n-gram and $Count_{match}(gram_n)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries [Chin-Yew Lin 2004].

ROUGE-S Skip-Bigram Co-Occurrence Statistics:

Equation 4-9: ROUGE S Skip Bigram Recall

$$\text{Recall } R_{skip2} = \frac{skip2(X,Y)}{C_d(m,2)}$$

Equation 4-10: ROUGE S Skip Bigram Precision

$$\text{Precision } P_{skip2} = \frac{skip2(X,Y)}{C_d(n,2)}$$

Equation 4-11: ROUGE S Skip Bigram F Measure

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

Where reference summary X is of length m and Candidate summary Y is of length n. skip2(X,Y) is the number of skip bigram matches between X and Y. d is the skip distance. If d = 0, then ROUGE-S is equivalent to bigram overlap. For example if d=4 (as in ROUGE-SU4), then only word pairs of at most 4 words apart can from skip-bigrams. The denominators of Equation 6 and Equation 7 is adjusted appropriately. The number of possible bigram counts $C_{d=m}(m,2)$ is as follows:

$$C_{d=m}(m,2) = \frac{m!}{(m-2)!2!}$$

where C_d is adjusted by d if necessary – the maximum distance that words can be apart.

Equation 4-11 is the F measure, weighted by β to favor precision or recall. A large value for β favors recall.

After ROUGE was developed and shown to correlate with human judgments on the DUC data [Lin and Hovy 2003], starting with DUC 2004, ROUGE started to be used as one of the evaluation metrics [DUC]. In 2004, ROUGE-2 and ROUGE-SU4 were shown to have the best correlation for multi-document summaries and these metrics became the "standard".

One issue with ROUGE was that it measured just overlap or presence or the similar word sequences and was not able to handle synonyms or equivalent ways to represent the same fact, e.g., "200 people were killed", as compared to "200 people died". The need for semantically equivalent nuggets led to the development of ROUGE-BE [Hovy et. al 2006] and the PYRAMID method [Nenkova and Passoneau 2004]. These methods were applied to the DUC evaluation in 2005 and then adopted for future DUC evaluations [DUC].

4.4 Metrics that take into account semantic equivalence

It is a known fact that humans chose different content when writing summaries, especially when the task is not well-defined or specific. Generic summarization falls into that category and so does topically oriented or query relevant summarization. For example, how much background information does one provide in a topic oriented summary? Research has suggested that one needs multiple gold standard abstracts to effectively rank summaries. Teufel's and van Halteren's research suggest that this may be on the order of 30 [Teufel and van Halteren 2004], whereas Nenkova's research suggests that 5 is enough with their coarser level of "summary unit" granularity [Nenkova et. al. 2007].

Unfortunately, in the DUC evaluations, the required number of human gold standards has remained a subjective process. Not only are summaries are created by a variety of assessors, they are often evaluated by the assessor who wrote the topic. It is not clear whether or not the other assessors writing summaries share the same information seeking goals or standards as the assessor who created the topics. The DUC coverage metric (covered in the next section) at DUC 2002 averaged 50% for single document summaries and less than 40% for multi-document generic (not topic focused) summaries [Harman and Over 2004].

4.4.1 Human Judgments

In the NIST DUC evaluations, summaries were evaluated by assessors using a set of quality questions through the SEE interface developed in 2001 by Chin-Yew Lin at ISI [Lin and Hovy 2002]. These questions covered coverage, coherency, redundancy, grammatically among other items. The coverage human judgments attempted to determine semantic equivalence between the system and human reference summaries.

The questions from DUC 2002 and 2003 are provided below [DUC-QUALITY 2003]

1. About how many gross capitalization errors are there?
2. About how many sentences have incorrect word order?
3. About how many times does the subject fail to agree in number with the verb?
4. About how many of the sentences are missing important components (e.g., the subject, main verb, direct object, modifier) – causing the sentence to be ungrammatical, unclear, or misleading?
5. About how many times are unrelated fragments joined into one sentence?
6. About how many times are articles (a, an, the) missing or used incorrectly?
7. About how many pronouns are there whose antecedents are incorrect, unclear, missing or come only later?
8. For about how many nouns is it impossible to determine clearly who or what they refer to?
9. About how many times should a noun or noun phrase have been replaced with a pronoun?
10. About how many dangling conjunctions are there (“and”, “however” …)?
11. About how many instances of unnecessarily repeated information are there?
12. About how many sentences strike you as being in the wrong place because they indicate a strange time sequence, suggest a wrong cause-effect relationship, or just don’t fit in topically with neighboring sentences?

All of these questions were measured on a 5 point scale.

For coverage, the assessor examines the model summary, which is composed of model units (MUs) which are human-corrected chunks provided to the assessor. The evaluator steps through all MUs and marks all peer units (PUs) in the system summary sharing content with the current MU. The assessor then indicates whether the marked PUs taken

together express what percentage (0%, 20%, 40%, 60%, 80% or 100%) of the content in the current MU. The evaluator reviews all unmarked PUs and indicates for the entire peer summary what percentage of unmarked PUs (using the same scale) are related but needn't be included in the summary.

In DUC 2004, 7 quality questions were used as well as the coverage assessment.

In DUC 2005, Evaluators used 5 linguistic quality questions (again 5 point scale) [DUC-QUALITY 2005] and the subject of focus was introduced:

1. *Grammaticality*: The summary should have no datelines, system-internal formatting, capitalization errors or obviously ungrammatical sentences (e.g., fragments, missing components) that make the text difficult to read.
2. *Non-redundancy*: There should be no unnecessary repetition in the summary. Unnecessary repetition might take the form of whole sentences that are repeated, or repeated facts, or the repeated use of a noun or noun phrase (e.g., "Bill Clinton") when a pronoun ("he") would suffice.
3. *Referential Clarity*: It should be easy to identify who or what the pronouns and noun phrases in the summary are referring to. If a person or other entity is mentioned, it should be clear what their role in the story is. So, a reference would be unclear if an entity is referenced but its identity or relation to the story remains unclear.
4. *Focus*: The summary should have a focus; sentences should only contain information that is related to the rest of the summary.
5. *Structure and Coherence*: The summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to sentence to a coherent body of information about a topic.

Since DUC 2005 moved from generic summarization to query relevant (topic focused summarization), the responsiveness evaluation was introduced. All summaries were read and then graded according to how responsive they were to the topic relative to the others (5 point scale).

Responsiveness is measured based on the amount of information in the summary that actually helps to satisfy the information need expressed in the topic statement at the level of granularity requested in the user profile. In DUC 2005, there were two levels of granularity– general and specific.

In DUC 2006 and DUC 2007, the NIST assessors used the same 5 quality measures and responsiveness, except the level of granularity was dropped.

4.4.2 Factoids

In 2003, van Halteren and Teufel were the first to present summary evaluation using content comparison [van Halteren and Teufel 2003]. They suggested a summarization unit which they named the factoid, which was based on atomic information units that

could be robustly marked in text. For two documents and 50 and 20 gold standard summaries respectively, they estimated that 20-30 gold standard summaries are needed for stable system rankings [Teufel and van Halteren 2004]. They also showed that factoid scores cannot be sufficiently approximated by unigrams.

The factoids are atomic semantic units. For example, the sentence "The police have arrested a white Dutch man" is the union of the factors "a suspect was arrested", "the police did the arresting", "the suspect is white", "the suspect is Dutch", "the suspect is male". Such fine granularity allows a very precise scoring of the resultant summaries. Note that the Pyramid Method (discussed in Section 4.4.4) does not require such fine grained units. A summary content unit (SCU) in the pyramid method would contain this entire sentence. A system summary that stated "the police arrested a Dutch man" would be treated as an match to this SCU – there are no partial matches.

4.4.3 ROUGE with Basic Elements

ROUGE started being used in the NIST DUC evaluation based on the need for automated metrics and the research which showed its correlation with human judgments [Lin and Hovy 2003]. However, as previously mentioned, its incapacity to measure semantic equivalence led to the development or ROUGE with Basic Elements, ROUBE BE. This methodology was presented to the summarization community at DUC 2005 [Hovy et. al. 2005]. Rouge BE attempts to create individual BE units from text, rate the similarity of BE units and score BEs by comparing system summary BEs with reference ones.

A BE is defined as "the head of a major syntactic constituent (noun, verb, adjective or adverbial phrases) expressed as a single item, or a relation between a head-BE and a single dependent, expressed as a triple (head | modifier | relation)" [Hovy et al. 2006]. BEs are created by decomposing parse trees by hand-built "cutting rules".

In scoring, a BE gets one point for each reference summary it participates in. The score is weighted by the completeness of the match by the matching strategies given below (in order of easiest to most difficult).

1. lexical identify – exact match of words,
2. lemma identify – roots of words match,
3. synonym identity as identified by WordNet,
4. phrasal paraphrases match,
5. semantic generalization as through an ontology (e.g., John Doe is a human).

For further details of this method, please refer to Hovy's papers [Hovy et. al 2005, Hovy et. al. 2006]. Since this method is fully automatic, if a semantically equivalent item is not matched by one of these matching strategies, the summary cannot receive credit. For example, the fact that in 2008, "a man who walks in space" is an astronaut may not result in a match. If WordNet does not have the information that a "taikonaut" is in the same set of cognitive synonyms (synset) as an "astronaut" or "cosmonaut", the content in the summaries may also not result in a match. (As of 5th October 2008, WordNet does not contain the word taikonaut, but does contain cosmonaut).

Hovy and colleagues were able to show ROUGE-BE correlation with the DUC human responsiveness measure, Spearman rank coefficient of 0.93 and Pearson coefficient of 0.98 (the closer to 1.0 in both these measures the more correlated the data) [Hovy et. al. 2005, Hovy et. al. 2006]. Naturally, BE was also highly correlated with the variants of ROUGE. ROUGE-BE appears to be an effective automated measure and started to be used as an official metric in the DUC evaluations in DUC 2006.

The advantages of ROUGE-BE is that it is a fully automatic method and thus allows for developers to improve summarization system once there is a gold standard corpus of human summaries. The other methodology which has been used in DUC, is the Pyramid method, discussed in the next subsection. This involves human annotation for every system summary version in order to assess summary improvement.

4.4.4 The PYRAMID Method

In contrast to ROUGE-BE, the Pyramid Method involves human marking of items in a summary and human grouping into related units. It attempts to address the fact that items can be chosen by summarization systems that are related to the topic but not expressed in a human "model" summary (this is what ROUGE-BE attempts to do through the matching strategies, especially #3-#5).

The pyramid evaluation method [Nenkova and Pasonneau 2004] groups related "nuggets" of information, which they call summarization content units (SCUs). In the process, one or more humans mark SCUs in the gold standard summaries to provide a "*master*" *template* for judging system summaries. Nenkova and Passoneau report that five gold standard summaries are needed for this methodology [Nenkova et. al. 2007]. After the master template is formed, the equivalent SCUs in the system summaries are marked according to this template. If a system summary has a SCU that was in all human gold standard summaries (weight of 5 if there are 5 summaries), then it receives a higher score because this is an item that all humans deemed important. Based on the first test of the Pyramid Method at DUC 2005 [Passonneau et. al. 2005], it was found that a modified pyramid score could be used, where any SCU in a system summary that did not appear in the master template did not need to be marked. These SCUs do not contribute to an overall system score that is focused on recall of the SCUs in the human summaries. Eliminating such SCUs greatly reduced the time needed to mark SCUs due to the frequency of such singletons in system summaries.

An example of a SCU is "200 people died". Let us suppose that this SCU is in the master template. A system summary might have "200 people were killed". The human annotator would mark this system summary as matching that SCU. The number of SCUs in the system summary and the weight of the SCU in the template (how many model summaries contain this SCU) are used to compute an overall score for the system summary.

The formula for the original pyramid is as follows:
Suppose a pyramid has n tiers, T with Tier T_n on top and T_i on the bottom. The top tier, T_n corresponds to the highest weight in the pyramid. If there are 4 tiers, then the weight of the top Tier is 4 and the weight of the bottom tier T_i is 1.
$|T_i|$ denotes the number of SCUs in tier T_i. D_i. is the number of SCUs in the summary that appear in T_i. SCUs in a system summary that do not appear in the pyramid are assigned weight zero.

Equation 4-12: Total SCU weight

$$D = \sum_{i=1}^{n} i * D_i$$

The optimal content score for a summary with X SCUs is shown in Equation 4-13 and the Pyramid Score in Equation 4-14.

Equation 4-13: Optimal content score for a summary

$$Max = \sum_{i=j+1}^{n} i * |T_i| + j * \left(X - \sum_{i=j+1}^{n} |T_i| \right) \quad where \;\; j = \max\left(\sum_{t=i}^{n} |T_t| \geq X \right)$$

where j is equal to the index of the lowest tier an optimally information summary will draw from and X is the summary size in SCUs.

Equation 4-14: Pyramid Score

$$P = \frac{D}{Max}$$

The Pyramid score in Equation 4-14 is a precision score, showing how many content units of a new summary match a set of summaries.

A pyramid score can be computed that corresponds more to recall. This *Modified Pyramid Score* is defined as the weight of the SCU's normalized by the weight of an ideally information summary of SCU size equal to the average SCU size of the human summaries in the pyramid. Thus the X in Equation 4-14 now corresponds to the average number of SCUs in the model summary used for the creation of the pyramid, rather than the original definition of SCU length.

As an example, let us consider the first three sentences of the initial five lead sentence summaries of the historic Chinese walk in space examined in Chapter 1. A sample SCU master template is shown in Table 4-2. There are 21 SCUs including the 3 about the American space program from the Daily Mail summary that are not shown in the table. Each SCU# has a corresponding number of human summaries that contain this SCU – the column Number. This value corresponds to the weight of the SCU:

SCU #	Daily Mail	3News	Reuters	BBC	Number
Total.	9 (6 shown)	6	10	8	
1	China celebrating				1
2	Zhai Zhigang first Chinese man to walk in space	Zhai Zhigang first Chinese man to walk in space	Zhai Zhigang first Chinese man to walk in space	Chinese astronaut first in country to take walk in space	4
3	China put men in orbit			capsule orbiting the Earth	2
4	China launch project for man on moon				1
5	Americans scientists fear left behind[5]				1
6	NASA hit by credit crunch				1
7		The 41-year old jet pilot	Zhai, the 41 year old son	Mr Zhai, 42	3
8		The … jet pilot		fighter pilot Zhai Zhigang	2
9		15 minute EVA from Shenzhou VII spacecraft	Clambered out of China's Shenzhou VII spacecraft	Emerged from the capsule (S1) … stayed outside the capsule for 15 minutes (S3)	3
10		China's third manned space journey			1
11		Blanket media coverage in China	Zhai's achievement live on state TV	Broadcast live on national TV	3
12			Beijing wants the world to marvel		1
13			Zhai feels well		1
14			Zhai greets people		1
15			Zhai son of snack-seller		1
16			Zhai unveiled a small Chinese flag	To wave a Chinese flag	2
17			Colleague Boming helps Zhai and briefly pops head out of capsule		1
18				Two fellow astronauts stayed in the spacecraft	1
Score	12 (3 not shown)	16	20	20	

Table 4-2: Master Template for SCUs on topic historic Chinese spacewalk
SCUs for four Gold Standard three Sentence Lead Summaries are shown: These summaries are taken from the 5 single document summaries in Chapter 1. Only one Reuters summary is used in this illustration since the two Reuters summaries are nearly identical.

[5] There were 5 SCU's in the Daily Mail covered the U.S. situation. We only display SCUs 5 and 6 from Sentence 2 here. The three SCUs in Sentence 3 are not displayed since this information was only in this Daily mail lead summary.

Pyramid Tier	Weight	SCU Numbers from Table 4-2	Number with Weight
4	4	2	1
3	3	7, 9, 11	3
2	2	3, 8, 16	3
1	1	1, 4, 5, 6, 10, 12, 13, 14, 15, 17, 18	11 (14 with 3 not displayed SCUs)

Table 4-3: Pyramid Composition

Suppose we now score the first three sentences of the People's Daily Online lead sentence summary from Chapter 1 against the master template. These sentences are shown below:

> [1] Chinese taikonaut Zhai Zhigang slipped out of the orbital module of Shenzhou-7 Saturday afternoon, starting China's first spacewalk or extravehicular activity (EVA) in the outer space.
> [2] Donning a 4-million-U.S.dollar homemade Feitian space suit, Zhai waved to the camera mounted on the service module after pulling himself out of the capsule in a head-out-first position at 4:43 p.m. (0843 GMT), video monitor at the Beijing Aerospace Control Center (BACC) showed.
> [3] "Shenzhou-7 is now outside the spacecraft.

Sentence 1 matches SCU #2, 3 and 9 in the master template (Table 4-2.) We count Sentence #2 as a match to SCU #14, since it conveys that Zhai waved to the camera – which we loosely interpret as a greeting to the people - although this is not explicitly stated. We will not count the information in Sentence #3 as a match to any SCUs in Table 4-2. Note the reader of this thesis is welcome to disagree with these assessments – this only shows the subjectivity of these type of evaluations.

The using Table 4-3, the total SCU weight of the lead sentence People's Daily,

Total SCU weight D = (1*1) +(1*2) + (1*3)+(1*4) = 10

(SCU #14 has a weight of 1, SCU #3 a weight of 2, SCU #9 a weight of 4 and SCU #2 has a weight of 4).

Using the recall oriented version of Max, the average number of SCUs in the model summaries, $Max = \frac{12+16+20+20}{4} = 17$.

Then the recall oriented Modified Pyramid Score for the People's Daily lead summary $P = \frac{10}{17} = 0.59$

As illustrated in this example, there are some difficulties when creating master templates in the Pyramid method. These include (1) determining the granularity of a SCU and (2) determining whether a portion of text is "equivalent" to another text. For example in Table 4-2, even in the gold standard summaries, should the SCU be "41 year old jet pilot" as mentioned in the text, or split into 2 SCUs, "41 year old" and "jet pilot"? Does an age of "42" match "41" and does a "jet pilot" match a "fighter pilot"? In the DUC pyramid

scoring, both of these items are probably counted as a match. In current evaluations, summaries are not judged on the *correctness* of the information nor the selection of the best sources. It is worth noting that in the factoid method [Teufel and van Halteren 2004], where the text is decomposed to its atomic elements, 41 year old would probably be separated from 42 year old, as well as jet pilot from fighter pilot.

Since SCUs are coarse, even within SCUs there can be *more informative* SCUs. In our example, the BBC and the 3News lead sentence summary contained the information than the astronaut was outside the capsule for 15 minutes (SCU #9). Is this fact important and important enough to be separated in the Pyramid Method? We did not form another SCU for this information but included in SCU #9. Furthermore, the BBC summary did not mention that the capsule was a Shenzhou VII, which the Reuters and 3News summaries did. Are there *vital pieces of information* that *must be present* in a summary? In our opinion, The 3News summary phrase "15 minute EVA from Shenzhou VII spacecraft" is the best SCU to describe this event with the exception that EVA is not defined for the reader unfamiliar with this terminology (EVA is extra vehicular activity).

Another important item about the pyramid method is that the judgment of *importance* is designed solely on how many gold standard annotators place that item in the summary. The BBC lead sentence summary was the only one to mention that there were 2 other astronauts and thus had a very low weight of 1. However, this might be a more important fact to some than the fact that this event was broadcast live on TV (which has a weight of 3). Again – what is the purpose of the summary? If the focus is on collecting social network information, then the details of the other astronauts accompanying Zhai are important as well as his age and the fact that he is the son of a snack-seller.

However, by using this approach to "cluster" semantically equivalent SCU, summaries can be measured on how many of the most important SCUs they provide and an overall score can be computed that gives an indication of the summary that matched the most important SCUs as determined by the gold standard summaries.

The modified Pyramid Method has been shown to correlate fairly well to human judgments from the DUC data [Passoneau et. al 2005]. However, it is very expensive (time and cost) to mark the number of human model summaries with SCUs as well as each system summary. When working to improve a summarization system using this methodology, since each system summary produced must be marked with SCUs to use this scoring method, there is a high human overhead is continually required – leading one to prefer this methodology more for formal (one time) evaluations.

The research on the pyramid method suggested that 5 humans summaries with SCU identification are sufficient, a fact presented at DUC 2005 [Passoneau et. al. 2005] and presented in later research [Nenkova et. al 2007]. Teufel and van Halteren's research suggests that 20-30 human summaries (using single document summaries) may be needed for evaluation stability [Teufel and van Halteren 2004]. Their factoids are at a finer granularity than the SCUs and Nenkova suggests that might be the reason [Nenkova et. al. 2007]. However, if that is the case, one ought to examine the information content

loss and effect of summary quality by grouping less informative items and more informative items into one SCU as we have discussed in this section.

Hori's research [Hori 2003] also indicated that using multiple references would improve evaluation stability if a metric took into account consensus among 25 manual summaries for each of 50 utterances in Japanese TV broadcast news.

4.5 Other Metrics

In this section, we discuss other evaluation methods related to summarization.

4.5.1 BLEU

BLEU was proposed for automatic evaluation machine translation in 2002 [Papineni et. al.] because human evaluation approaches are expensive and can takes weeks or months to finish. Developers of systems need to be able to monitor the effects of daily changes to their systems to test out new ideas. Accordingly, an automated metric would allow for a significant increase in system development. We present this metric because it was the first metric that let to the development of the other automated metrics for text evaluations. It has also been used for evaluating summarization systems [Lin and Hovy 2003, Hori et. al. 2003]

For machine translation, the systems are aiming to get very close to the reference translation(s). Thus BLEU is based on precision and a modified unigram count to deal with the issue of word overgeneration. To compute this, one counts the maximum number of times a word occurs in any single reference translation. Then, one clips the total count of each candidate word by its maximum reference count, sums the clipped counts and divides by the total (unclipped) number of candidate words. Thus if a candidate translation has 7 of the same word and is of length 7, but only 2 appear in one of reference translations, then the modified unigram precision is 2/7.

The modified n-gram precision is then given as;

Equation 4-15: n-gram Precision

$$\text{n-gram precision } p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Countclip(n-gram)}{\sum_{C' \in \{Candidates\}} \sum_{n-gram' \in C} Count(n-gram')}$$

The Brevity penalty is designed to penalize translations shorter than any of the reference translations.

Equation 4-16: Brevity Penalty

$$\text{Brevity penalty BP} = \begin{cases} 1 & if\ c > r \\ e^{(1-r/c)} & if\ c \le r \end{cases}$$

Where c is the length of the candidate translation and r is the effective reference corpus length.

Equation 4-17: BLEU

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

Where N is the length of the n-grams.

It is mentioned that the ranking behavior is more apparent in the log domain and the authors suggest for their baseline N=4 and uniform weights of $w_n=1/N$.

The authors note that BLEU only needs to match human judgments when averaged over a test corpus to eliminate the varying that happens on scores for individual sentences. For example, a system that produces the phrase "the United States economy" is penalized if all the references read the "economy of the United States".

BLEU is focused on precision rather than recall since reference translations can be so different – the idea is not to reward a candidate translation that does not make much sense but it is composed of words from different reference translations. Such a situation rarely occurs in summarization due to fact that systems compose summaries from text that already exists.

BLEU was tested for summarization by Chin-Yew Lin, who determined that ROUGE, a recall oriented metric, was more suitable for summarization, since it suffers less from the problem mentioned above. Hori also tried the BLEU metric for summarization evaluation of speech utterances [Hori et. al. 2003] and found that the BLEU score varied according to the number of manual summaries, unlike Hori's proposed metric WSummACCY.

4.5.2 WSummACCY

In this section, we briefly describe WSumACCY [Hori et. al. 2003], the weighted summarization accuracy SumACY.

The summarization accuracy, SumACY is calculated by using the word network generated through merging manual summarization. This is used to compact sentences.

WSumACCY obtains a reliability that reflects that majority of the human's selections by weighting the summarization accuracy by a posterior probability of the manual summarization network. The reliability of the extracted sentence from the network is defined as production of the ratio of number of subjects who select each word to the total number of subjects.

Hori used this method for summarization of Japanese utterances. Early work used SumACY for English TV Broadcast news [Hori et. al. 2002]. For further details, please

refer to Hori's paper [Hori et. al 2003]. This method has worked well for high compression ratios of 40% and short utterances of one sentence each.

4.5.3 METEOR

The metric was designed for machine translation (MT) evaluations, but could be adapted for summarization systems. Research [Lavie et. al. 2004] showed that the harmonic mean of precision and recall, F1, outperformed the BLEU and NIST metrics for Machine Translation evaluations in terms of correlation with human judgments of translation quality (adequacy and fluency). Lavie and his colleagues presented evidence that better correlations with human judgments of adequacy and fluency for MT are achieved by putting more weight on recall (as summarization evaluations do) rather than precision as MT systems do. The correlations are highest with human judgments when almost all of the weight is assigned to recall.

METEOR creates a word alignment between two strings and thus could be used to do so for two summaries. The alignment is produced by a word mapping module, which identifies possible word matches using porter stemming and WordNet similarity using the synsets (synonym sets available through WordNet). Once an alignment is produced the METEOR score for a pairing is computed. The number of mapped unigrams between the two strings is m, the total number of unigrams is t and the total number of unigrams in the reference is r. Unigram precision $P = m/t$ and unigram recall $R = m/r$. F1 is as in Equation X, which can also be written as:

or $F_\alpha = \frac{PR}{\alpha P + (1-\alpha)R}$. $\alpha = 0.5$ is equivalent to $\beta = 1$ in Equation X.

METEOR computes a penalty for an alignment. The matched unigrams between the two items are divided into the fewest number of "chunks"so that matched unigrams are adjacent and in identical word order. A fragmentation fraction is computed: $frag = ch/m$. The penalty is computed as $Pen = \gamma \cdot frag^{\beta}$. The value of γ (between 0 and 1) determines the maximum penalty. The value of β determines the relation between fragment and penalty.

Equation 4-18: METEOR

$$\text{METEOR} = (1 - Pen) \cdot F_\alpha = (1 - \gamma \cdot frag^{\beta}) \frac{PR}{\alpha P + (1-\alpha)R}$$

METEOR depends on three parameters, α, β and γ. A higher value of $\alpha = 0.5$ equally favors precision and recall. $\alpha > 0.5$ favors recall over precision. For MT, METEOR parameters based on early experiments by Lavie [Lavie et. al. 2004] were initially set to $\alpha = 0.9$, $\beta = 3.0$ and $\gamma = 0.5$. In recent experimentation for adequacy and fluency [Lavie and Agrawal 2007] suggests $\alpha = 0.8$ for both MT measure adequacy and fluency and , $\beta = 1.0$ for adequacy and 0.8 for fluency, to 1.0 and $\gamma = 0.2$ for adequacy and 0.4 for fluency. Note that the correlations reported in this paper for the Pearson correlation are approximately 0.62 and 0.44, which are very low compared to summarization metric

correlations with human judgments. For ROUGE and PYRAMID correlations, all correlations have been over 0.8.

As mentioned, METEOR could easily be adapted for summarization and compared to another fully automated metric such as ROUGE-BE.

4.5.4 NUGGET PYRAMIDS and POURPRE

Another metric worth discussing is Nugget Pyramids [Lin and Demmer-Fushman 2006] and POUPRE [Lin and Demmer-Fushman 2005]. Nuggets Pryamids are a scoring methodology for the "other" questions in the Question Answering Track [Dang and Lin 2007] or for definition questions. An "Other" question can be best paraphrased as "Tell me interesting things about X that I haven't already explicitly asked about." The Other questions come after a series of questions. The series of questions typically consist of basic facts, such as "when", "who", "where", "what", "how many", etc. Since other questions are more "free-form", they suffer from many of the issues that summarization systems face. Assessors assign a value to an information nuggets (answers to the "other" questions or definition questions) are assigned a judgment by assessors as "vital/okay". The assessors differ in their notions of importance and even a single assessor over time can be inconsistent [Dang and Lin 2007].

Researchers observed that the TREC evaluations suffered from the fact that the median score for many questions turned out to be zero, since only vital nuggets affected nugget recall. Scores were easily affected by assessor errors. A score that was heavily skewed towards zero made the meta-analyses of evaluation stability difficult to perform (Vorhees, 2005).

Lin and Demmer-Fushman proposed a method of soliciting vital/okay judgments from multiple assessors after the list of nuggets had been produced by a primary assessor [Lin and Demmer-Fushman 2006]. This was based on the same intuition as the Pyramid Method [Nenkova and Passonneau 2004] where the importance of a fact is directly related to its popularity – it's inclusion in a summary. The weight assigned to a summary content unit (SCU) is dependent on how many human reference summaries contained it.

For the nugget pyramids, the weight assigned to each nugget is simply equal to the number of different assessors that assigned it a value of "vital". An assigned value of "okay" means that this nugget does not count towards the score of the question answering system. The vital nuggets weights are then normalized per question so that the maximum possible weight is one (by dividing each nugget weight by the maximum weight of that particular question).

Precision remains the same as in TREC evaluations – a length allowance of 100 non-whitespace characters and longer answers are penalized for verbosity. Nugget recall is modified to take into account nugget weight.

Equation 4-19: Nugget Recall

$$R = \frac{\sum_{m \in A} w_m}{\sum_{n \in V} w_n}$$

Where A is the set of reference nuggets that matched within a system's response, V is the set of all reference nuggets, w_m and w_n are the weights of nuggets *m* and *n* respectively. Thus, all nuggets now factor into the calculation of recall subjected to a weight, where vital nuggets get a score of one and okay nuggets zero.

In the NIST TREC-QA, nine sets of judgments were elicited from eight judges (the primary assessor who created the nuggets annotated them twice).

Jimmy Lin [Lin and Demmer Fushman 2005, Lin 2006] created an automated scoring system POURPRE which can be used for improving system performance based on scoring the "other" questions with based TREC QA data. POUPRE calculates a count based, stemmed, unigram similarity between each nugget description and each candidate system response. If the similarity passes a threshold, then this similarity is used to assign a partial value for recall and a partial length allowance. The ranking of systems is close to the official ranking. POUPRE is automated and designed for QA systems, similar to BLEU and METEOR for MT, and ROUGE for summarization.

4.5.5 NUGGETEER

The last metric we discuss is Nuggeteer. Nuggeteer was developed by Marton and Raduhl for the TREC Definition and Relationship questions – the "other" questions [Marton and Radul 2006]. Nuggeteer offers four improvements over POURPRE, (1) interpretability of the scores, (2) use of known judgments for exact information about some responses, (3) information about individual nuggets for detailed error analysis and (4) support for improving score accuracy through additional annotation. The system allows developers to add their own judgments and have the responses scored.

Nuggeteer builds one binary classifier per nugget per each question based on n-grams up to trigrams and calculates an idf-weight based on counts from the AQUAINT corpus. The authors' experiments showed that stemming hurt POURPRE slightly at peak performances, but Nuggeteer was stable with respect to stopwords, various weights to terms and stemming. The authors found that using bigrams yield slightly better results (POURPRE is based on unigrams), which they hypothesize is based on the fact that bigrams sometimes capture named entity and grammatical order features. It is interesting to note that ROUGE-2 (bigrams) has been shown to have one of the best correlations with human judgments. For further details about Nuggeteer, the reader is referred to Marton and Radul's paper [Marton and Radul 2006].

4.6 Issues with Gold Standards Summaries

As alluded to in this section, in order to effectively evaluate summary content (or responsiveness), there must be human gold standard summaries created. Researchers often gather students or others to build those gold standard summaries [Jing et. al. 1998, Mani et. al. 1998, Goldstein et. al 1999, Goldstein et. al. 2000, DUC 2001-2007, NTCIR 2001-2004, Radev et. al. 2002, Radev and Tam 2003, Teufel and van Halteren 2003, Nenkova and Passonneau 2004, Goldstein et. al. 2007]. However, such summarizers are usually novices, not professionals. One question arises if the summarizers were professionals would this change the results that are obtained?

There are some interesting ideas to consider in regard to this question. First, Nenkova mentions that for the Pyramid Method, the decision to weight SCUs by how many humans mention this item, "assumes that the summary writers are equally capable, and good at the summarization task." [Nenkova et. al. 2007]. Endres-Niggemeyer states that professional summarizers are provided with explicit methods, some of which are taken from standards, textbooks, and guidelines [Endres-Niggemeyer 2007]. When designing the types of genre-oriented summaries (such as personal and professional for movies), for our research, we decided to use three English students with some training in "abstracting" through studies in journalistic and/or technical writing. Some of their ideas of what should be included in a summary based on their training might be considered novel to a person non trained in such methods. The most surprising was that these three people all thought questions from an interviewer should not be contained in a summary – that a good summary sentence answer should have all the relevant pieces.

The same concept was supported at SUMMAC with an intelligence analyst. In an experiential exercise on summary creation, an intelligence analyst strongly suggested that summary sentences for "background" information should not be included. In her opinion, such sentences conveyed information that should already be known.

Such a response points to the fact that creating summaries is very task specific. The summaries must be designed to meet a user' information seeking needs. Many instructions given to the summarizers who create gold standards are not that specific. Should one include background or not? What level of knowledge does one assume on the part of the reader? What is the summary trying to accomplish? What differentiates a good piece of information vs. a mediocre - such as the vital and okay distinction for answers in the question answering track of the NIST Text REtrieval Conference?

In many fields, there are formal and/or informal guidelines to writing certain types of articles – such as newswire, press releases, technical articles, and reviews [Endres-Briggemeyer `1998. We have discussed some of these in the section on "what is a good summary", Section 2.2.2. The English students with whom we worked to create the guidelines for forming summaries had clear ideas as to what types of genre oriented summaries should exist and what the composition of the genre oriented summaries should be. Perhaps the summarization field would benefit from adopting more stringent guidelines to summarization creation. We are not aware if the NIST assessors have

detailed guidelines and standards as to what constitutes a "good summary" when creating their multi-document summaries.

We do not attempt to answer these questions in this thesis. Our point is only that even in evaluating summaries and creating gold standard summaries, there are issues that need to be further explored in the research community. In the next chapter, we give an overview of the formal summarization evaluations.

Chapter 5 Formal Summarization Evaluations

"However beautiful the strategy, you should occasionally look at the results."
Winston Churchill

In this chapter, we give an overview of the government sponsored evaluations that have taken place for summarization. These evaluations are designed to advance scientific research and are open to the worldwide community. In the U.S., evaluations included the Defense Advanced Research Projects Agency (DARPA) TIPSTER SUMMAC in 1999, the National Institute of Standards and Technology (NIST) Document Understanding Conferences (DUC) from 2001-2007, the Multilingual Summarization Evaluations in 2005 and 2006, and the upcoming NIST Text Analysis Conference (TAC) to be held this November 2008. In Japan, these included the Japanese NTCIR Text Summarization Challenge (NTCIR 2-4, TSC-1-3) from 2000-2003 and the NTCIR 6 pilot for Multimodal Summarization of Trends (MuST) from 2006-2007 and the upcoming NTCIR MuST evaluation to be held this December 2008. We will not cover the MuST evaluation as our focus is solely on written text documents.

We compare the sponsored evaluations for English (SUMMAC, DUC, GALE) both in the types of summarization covered as well as the data and languages they addressed. This is then followed by a discussion of the limitations of these evaluations, in particular, are they addressing the needs of the user focused applications?

We also present the results of the content evaluation of the Multilingual Summarization Evaluation, modeled after the DUC evaluations. These results showed that judging the quality of summaries without taking into account the original documents can result in inflated judgments of summary quality. This result was also observed at SUMMAC as a result of one of the evaluations. We conclude this chapter with a discussion on how evaluations might be improved based on our analysis of the properties of "good" summaries.

5.1 Overview of SUMMAC, DUC, MSE, TAC, NTCIR, GALE

An overview of the summarization evaluations are provided in Table A. This section discusses the SUMMAC, DUC, MSE, TAC and NTCIR evaluations. The MSE evaluation we performed is discussed in detail Section 5.3 as it indicates some pitfalls of human assessment of summaries.

5.1.1 SUMMAC

A focus on summarization started in 1997 with the TIPSTER SUMMAC evaluation. TIPSTER was a Defense Advanced Research Project Agency (DARPA) initiative with a focus on text processing., started in 1992. In the third phase of TIPSTER, DARPA

expanded the program to include text summarization and to hold an evaluation for automatic systems.

In October 1997, six TIPSTER Phase III participants participated in a dry run, including (1) Carnegie Group Inc. and Carnegie Mellon University, (2) Cornell University and SaBIR Research Inc., (3) GE Research and Development, New (4) New Mexico State University, (5) The University of Pennsylvania and (6) The University of Southern California Information Sciences Institute.

The formal evaluation, which was open to the community took place in October 1998. 16 systems participated [Mani et. al 1998]. The goals were to judge individual summarization systems in terms of the usefulness in summarization tasks (an extrinsic evaluation) and to gain a better understanding of the issues involved in building and evaluating summarization systems. One subgoal was to examine the effect of compression so participants were asked to provide summaries of both a fixed compression ratio of 10% of the character length of the document and a variable length summary of their choosing. The topics were taken from the NIST TREC (Text REtrieval Conferences) data and for each topic there was a set of to documents. There were three tasks:

(1) Categorization task for routing and filtering of information (extrinsic tasks): generic (overview) single document summaries, 10 topics, 100 document per topic. Summaries are used for judging document relevancy for both fixed and variable compression.
(2) Determine document relevance for a topic quickly and accurately (extrinsic task): adhoc (topic related) single document summaries used for judging document relevance for both fixed compression and variable, 20 topics and 50 documents per cluster
(3) Question Answering (intrinsic task) for 3 topics, 30 documents per cluster: Each summary is assessed as to whether it provides answers that must be satisfied for the document to be judged relevant. The set of questions was not made available to the participants. Systems were allowed to provide summaries at a variable length for the documents in the topic related cluster.

Due to the nature of extrinsic tasks, the time spent reading summaries was also measured. For a full discussion of the evaluation, refer to Mani's report on the results [Mani et. al. 1998].

The Carnegie Group/Carnegie Mellon summarizer was developed at Carnegie Mellon University and was the first summarizer to incorporate both novelty and anti-redundancy. Although anti-redundancy is not as prevalent for single document summarization in the genre of newswire, it is our opinion that addressing this topic assisted our summarizer, which was the top scoring summarizer in the adhoc (topic-focused) summarization evaluations for both fixed compression and variable length as well as had the most answer recall in the question answering task. The anti-redundancy measure we used is maximal marginal relevance (MMR) [Carbonell and Goldstein 1998] and is discussed in Section 3.2. The MMR metric is used by many summarization systems in the DUC evaluations [DUC] and researchers at the National University of Singapore have adapted

MMR for better multi-document summarization results by using WordNet and concept similarity to choose sentences [Ye et. al. 2005]. The number of content units in each sentence influences the selection. Note that this is a similar concept to our multi-document summarization algorithm [Goldstein et. al. 2000], which takes into account the number of passage clusters in which a sentence participates in sentence selection. The National University of Sigapore team [Ye et. al. 2005] using their modified MMR algorithm was able to achieve the top automated ROUGE-2 and ROUGE-SU4 scores among the participants in DUC 2005 (we did not participate).

5.1.2 DUC, MSE and TAC

Following SUMMAC, the Document Understanding Conferences (DUC) started in 2000 [DUC] and the Question Answering track in the Text Retrieval Conference started [TREC-QA]. The discussion of these evaluations started with DUC 2000 and the development of a roadmap [DUC-ROADMAP] in which summarization and question answering would eventually merge. This has come to fruition in the past year with the advent of the NIST Text Analysis Conference (TAC) bringing together question answering and summarization [TAC].

The first DUC evaluation took place in 2001 and the last in 2007. The summarization evaluations focused on intrinsic evaluations in the newswire domain for single and multiple summaries. There were some specialized tasks in the later years, covering multilingual summarization (Arabic and English), a focus on summaries answering the "who is PERSON" question and some topic focused summaries. These evaluations are summarized in Table A.

In DUC 2004, a multilingual summarization was explored, where systems were to create English summaries of newswire documents both in English and in Arabic. Machine Translation output was provided for the Arabic newswire documents by ISI's and IBM's systems. The decision was made not to continue this evaluation. Under the DARPA Translingual Information Detection, Extraction and Summarization (TIDES) program [TIDES], Columbia University had prepared 50 topic clusters consisting of Arabic and English documents and the Linguistic Data Consortion (LDC) had created 4 human summaries for each cluster. Columbia University made the data available, and two multilingual summarization evaluations (MSE) were conducted in 2005 and 2006 in conjunction with ACL summarization workshops. For MSE 2006, a content evaluation was performed – these results are discussed in Section 5.3. At least two interesting results emerged from this evaluation:

1. CLASSY, the highest performing system in 2005 and 2006, only used sentences from the English documents in the output summaries [Conroy et. al 2006b, Schlesinger et. al 2008].
2. After reading the original documents, human assessors judged system summary quality lower than their assessments of summary quality prior to reading the original documents. This was not true of the human assessments of human summary quality.

The most recent DUC evaluation in 2007 explored the update multi-document summarization task – providing an update summary based on a new cluster of a documents given one prior cluster or two prior clusters. This requires a focus on novel information. At DUC 2007, it was decided to move summarization to the Text Analysis Conference (TAC). TAC has three tracks, Question Answering, Recognizing Textual Entailment and Summarization [TAC]. It grew out of DUC (DUC) and the Question Answering Track of the TREC. The Recognizing Textual Entailment (RTE) track's purpose is to develop systems that recognize when one piece of text entails another. This task has been previously explored three PASCAL RTE Challenge Workshops 2005-2007 [PASCAL-RTE].

TAC 2008 [TAC] is continuing the update summarization task. It also is piloting a new task – that of writing summaries of opinions from blogs. Participants will produce short summaries of answers to questions, given the questions from the TAC QA Track and the text snippets output by QA systems. Summaries will be assessed for readability and content and evaluation of content will be based on the Nugget Pyramid Method. [Dang and Lin 2007] used for evaluating answers to questions.

5.1.3 NTCIR

Summarization evaluations were also conducted by NTCIR for multi-document summarization from Japanese newswire text – the Mainichi newspaper [NTCIR]. These lasted three years Text Summarization Challenge (TSC 1-3) from the 2nd NTCIR Workshop (2000/2001) to the 4th NTCIR (2004). Question answering was covered from the 3rd NTCIR to the 6th NTCIR. The 6th NTCIR (2006-2007) had a pilot on Multimodal Summarization of Trend Information (MuST) using Japanese documents only from the Mainichi Newspaper. The focus of MuST is how to exploit "trends" as an abstract of summary of information to be accessed using linguistic and visual information together where such information co-exists. At the 7th NTCIR (2007-2008), again Japanese newswire data, the tasks focused on specific themes in compilation of textual and numerical information for information access and use of visual information. These include text to numerical data conversion, text generation from numerical data and alignment of textual information and time series data.

5.1.4 GALE

In 2006, the Defense Advanced Research Projects Agency DARPA Global Autonomous Language Exploitation (GALE) program has introduced the concept of "distillation" [GALE]. In response to a query template, an information distiller seeks to capture all the relevant information nuggets in a set of documents which can consist of translated/transcribed data from speech input as well as foreign text with machine translation in the languages Arabic and Chinese. These nuggets are produced from input text in foreign languages as well as speech data that has been transcribed and translated. Nuggets that mean the same thing are grouped into sets, which are called *nugs*. Nuggets can come from any type of data, which includes text and audio, internet chat rooms & TV talk shows as well as newswire and radio/TV news shows. A thorough distiller would

extract all relevant information [White et. al 2007]. For our purposes we can think of the output of a distiller as potential items to input to a summarizer, which would want to organize and present the information in a usable manner. This concept is similar to the pilot opinion task in TAC, which is attempting to take the output of question answering systems in regard to opinions in blogs and form summaries. GALE has introduced a new metric *proficiency*, which measures how much information a distiller provides relative to all of the information provided by a collection of distillers working on a common query and corpora.

5.2 Discussion of Summarization Evaluations

Table 5-1 shows the major English summarization evaluations from the years 1998-present. The evaluations can be described in terms of the types of evaluation intrinsic, extrinsic as well as other parameters that define the evaluation, including languages, summary length and metric. SUMMAC has been the only evaluation that focused on extrinsic evaluations (marked in Table A) –how well a summary would allow users to perform a particular task, the rest of the formal evaluations have been intrinsic, focused on summarization quality. A natural question is - are intrinsic evaluations an effective measure of a good summary?

The Summary Purpose can be Generic, Adhoc or Query Relevant (Topic focused) or Question Answering focused. Adhoc can be referred to as query relevant. Summaries from such queries focus on a particular topic, e.g., in the general category of plane crashes, the focus could be on the topic of mechanical failures that cause crashes. The Question Answering type can be thought of as a subtype of Query Relevant. If the topic has specific questions, these focus the summary on items relevant to these specific goals. All these items can be thought of as subcategories of goal-focused summaries. In addition, the questions could be designed for what we term "genre oriented summaries". For example, if the user is interested in the movie genre, at a broad level, such a genre would have three types of useful summaries – overview, plot and opinion. Thus focused on a question "did people have positive opinions of the movie?" would lead to a genre oriented goal focused (question answering focus) summary based on movie opinions.

Other summary purposes could be "Update" – a focus on updating information, "Opinion" – a focus on opinions. It is obvious that there could be many types of categories in this section, some of which could be presented hierarchically as mentioned above for question answering and query relevant.

The Summary Role is to produce an informative or indicative summary. Informative summaries are designed to answer specific questions or items in the topic. Indicative summaries do not have to necessarily contain such content relevant information. They are more task specific – such as to indicate document relevance or provide an overall one line headline summary of the document.

Evaluation	Year	Data	Lang.	Data Set	Summ. Purpose	Summary length	Summary Role	Task	Metrics	Human Eval
SUMMAC Extrinsic	1998	Newswire	English	Single	Generic	Fixed & variable compression	Informative	Categorization of Doc	Time, Precision & Recall;	Quality, Correctness, Acceptable
SUMMAC Extrinsic	1998	Newswire	English	Single	Adhoc	Fixed & variable compression	Indicative	Relevance of Doc	Time, Prec. & Recall; Summary	Quality, Correctness
SUMMAC	1998	Newswire	English	Single	Adhoc-Topics	Variable	Informative	Provide Answers	Answer Recall	
DUC	2001	Newswire	English	Single	Generic	100 words				
DUC	2001	Newswire	English	Multi	Generic	400,200,100 50 words				
DUC	2002	Newswire	English	Single	Generic	100 words				Quality, Coverage
DUC	2002	Newswire	English	Multi	Generic	200,100,50 words				Quality, Coverage
DUC	2002	Newswire	English	Multi	Generic	10 words	Indicative - headline			Quality, Coverage
DUC Task 1	2003	Newswire TDT,TREC	English	Single	Generic	10 words	Indicative - headline			Quality, Coverage
DUC Task 2	2003	Newswire TDT	English	Multi	Generic	100 words				Quality , Coverage
DUC Task 3	2003	Newswire TREC	English	Multi	View-point	100 words	Informative	Express views		Quality, Coverage
DUC Task 4	2003	Newswire TREC	English	Multi	Novelty	100 words	Informative	Novel Info		Quality, Coverage
DUC Task 1	2004	Newswire TDT corpus	English	Single	Generic	75 bytes	Indicative - headline		ROUGE	Quality, Coverage
DUC Task 2	2004	Newswire TDT corpus	English	Multi	Generic	665 bytes			ROUGE	Quality, Coverage
DUC Task 3	2004	Newswire Arabic & MT output	English & Arabic	Single	Generic	75 bytes	Indicative - headline		ROUGE	Quality, Coverage
DUC Task 4)	2004	Newswire, and MT	English & Arabic	Multi	Generic	665 bytes			ROUGE	Quality, Coverage
DUC Task 5	2004	TREC	English	Multi	Answer "who is X?"	665 bytes	Informative	Answer question		Quality & Coverage
NTCIR TSC 1-3	2000-2004	Newswire	Japanese	Single Multi	Generic	Variable 10-50%			Recall, Prec, F1	Content & Readability
DUC	2005	Newswire	English	Multi	Adhoc Topic	250 words	Informative	Answer info need	ROUGE-2, -SU4	Quality,Cov, Responsive
MSE	2005 2006	Newswire and MT	English & Arabic	Multi	Adhoc	665 bytes			ROUGE	Responsive
DUC	2006	Newswire AQUAINT	English	Multi	Adhoc-Topics	250 words	Informative	Answer info need in topic	ROUGE-2,SU4, BE; Pyramid	Quality, Coverage, Responsive
DUC	2007	Newswire AQUAINT	English	Multi	Adhoc-Topics	250 words	Informative	Provide Answers	ROUGE-2,SU4,BE,	Quality, Pyramid
DUC	2007	Newswire AQUAINT	English	Multi	Update	100 words	Informative		ROUGE-2,SU4,BE,	Quality, Pyramid
NIST QA	1999-2007	Newswire	English	Doc Sets	Short Answer			Answer Question	MRR	Answers, Nugget Pyramid
GALE	2007	Variety	English, Ariabic, Chinese	Doc Sets	Distillation				P, R, Proficiency	
TAC	2008	Newswire	English	Multi	Update	100 word	Informative			Pyramid
TAC	2008	Blogs	English	Multi	Opinion					Nugget Pyramid

Table 5-1: Brief Details and Comparison of Summarization Evaluations
Cov = Coverage, Responsive = Responsiveness

In the past few years of DUC, the focus has shifted from indicative summaries to informative summaries. A new focus since 2007 is on update summaries – summaries that contain new information from what is assumed to be known from previous clusters. A new task in the first TAC to be held in November 2008 is to create summaries by using output of question answering systems. Although this distinction was made in the early days of summarization, it appears that all summaries should really be informative but be able to be used for indicative purposes as well.

Most of the American evaluations have focused on English with the exception of DUC 2004 and the two follow-up Multilingual Summarization Evaluation conferences in 2005 and 2006 which contained Arabic documents.

Initial evaluations focused on single document and multi-document generic summaries and then moved to multi-document topic based summaries in 2004. In addition, most of the evaluations have focused solely on the newswire genre. TAC 2008 is focusing on a new genre – blogs.

The Metrics column lists briefly the metrics used in the evaluation. The various metrics are discussed in detail in the previous chapter as well as their evolution as evaluations progressed. The previous chapter on metrics (Chapter 4) discusses these in detail that were as well as covers the evolution of human judgments within the DUC conferences including the development of the Pyramid Method [Nenkova et. al 2004] and Nugget Pyramids [Lin and Fushman-Demmer 2006] for in depth human evaluation of summarization and question answering respectively.

The column "task" in Table 5-1 attempts to summarize the type of task if specified in the description. This "task" description is vague or non-existent in most of the evaluations. DUC 2005 attempted to address this on a rough level by specifying whether the type of information provided by the summary should be specific or general. However, this did not make much difference in the evaluation and was dropped in future evaluations.

In the last section of this chapter, we discuss "what makes a good summary". One of the key components is that a summary needs to be task oriented and should answer what fits the users' needs and be tailored to the users' background. Most of the evaluations are too general to ascertain this information. Summarization might better be treated like a presentation – who is the intended audience as well as what is the intended purpose? The newswire is designed as a summary so summarizing newswire without a purpose is essentially creating a summary of a summary without a particular goal.

5.3 Multilingual Summarization Evaluation (MSE) Results

In this section, we discuss the results for the Multilingual Summarization Evaluation (MSE). MSE was modeled after the DUC evaluations and the official scoring metrics were ROUGE-2 and ROUGE-SU4. In the first MSE evaluation (2005), Columbia University performed a Pyramid Method evaluation for summary content; no other human judgments of summary quality were performed. In 2006, Columbia University

was unable to undertake such a time consuming event. Since only 8 systems participated in MSE 2006, we decided (without announcement to the participants) to design and perform a content evaluation based on the methodology used at the NIST DUC evaluations. The purpose of this evaluation was to determine whether or not reading the content of the documents prior to reading the summaries affected human summary assessment. Accordingly, a two phase evaluation was performed. Phase 1 was identical to a DUC NIST. In Phase 2, the human assessors read the content of all the documents in the cluster first and then read all the system and human summaries to assess quality. The details of the two MSE evaluations and the results of the MSE 2006 content evaluation are discussed in the next two sections.

5.3.1 MSE Overview

In DUC 2004, multi-lingual summarization was evaluated on clusters composed of Arabic and English. The decision was then made not to continue that evaluation thread in DUC. However, under the DARPA Translation Information Detection, Extraction and Summarization (TIDES) program [TIDES], the precursor to GALE, Columbia University had created 50 additional clusters of Arabic &English TDT newswire data using their clustering algorithms. The clusters had an average of 12.5 {5-25} documents per cluster, an average of 6.4 {2-13} Arabic documents per cluster and 6.1 {3-12} English documents per cluster. The Arabic data was translated by University of Southern California Information Sciences Institute's machine translation system. For each cluster, the Linguistic Data Consortium (LDC) prepared four human multi-document model summaries of 100 words.

It was decided to hold two Multilingual Summarization Evaluations (MSE) - MSE 2005 and MSE 2006. For MSE 2005, 25 clusters were used and for MSE 2006, 24 clusters. The remaining cluster was provided to participants in MSE 2005 as an example. The task was to create fluent 100 word answers to questions in the topic statement for a cluster of document. The topics requested "wh" information such as who, what, when, where, etc. Some of the topics had no narratives.

In 2005, there were 27 submissions from 10 different research groups. ROUGE-2 and ROUGE SU4 were the official evaluation metrics. Using the metric, ROUGE SU4, CLASSY significantly performed the other runs (0.186 compared to 0.169), using only sentences from the English newswire documents to produce summaries [Schlesinger et. al 2008]. (All other runs were between 0.169 and 0.157). A content evaluation was also performed using the Pyramid Method and one priority run was used from each participant. For this, the CLASSY run scored second, indicating that ROUGE SU4 and the pyramid method may not be as highly correlated as some of the results report.

In MSE 2006, there were 8 participants and 15 runs. ROUGE-2, ROUGE-SU4 and ROUGE-BE were used as official evaluation metrics. A content based evaluation modeled partially on the NIST evaluations for DUC was performed for the first submitted run from each participant. The participants were not aware that a content evaluation would be performed.

The results of this manual evaluation are discussed in the next section.

5.3.2 MSE Content Evaluation: Summaries Can Be Misleading

In MSE 2006, we conducted a manual evaluation of one run from each of the eight participants [Goldstein et. al. 2006b], using two NIST DUC measures: content responsiveness and overall responsiveness.

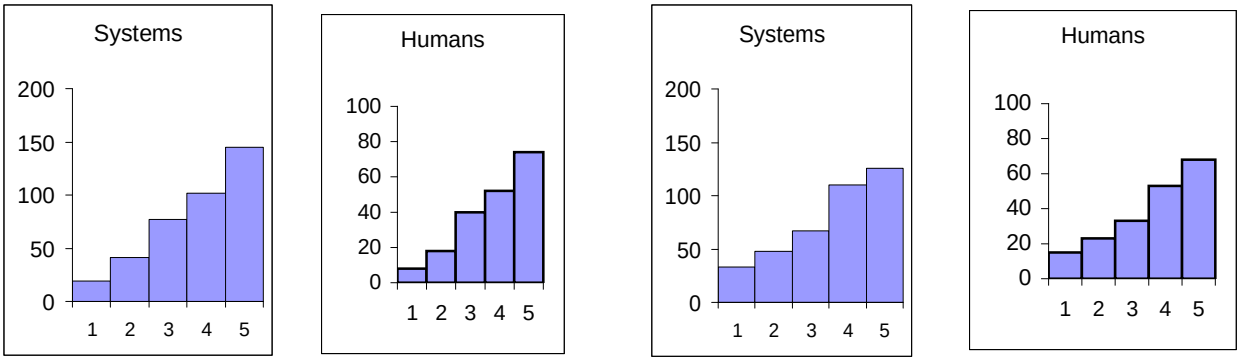
- *Content responsiveness* is measured based on the amount of information in the summary that contributes to meeting the information need expressed in the topic.
- *Overall responsiveness* is based on both the information content in the summary and the readability of the summary.

Two human assessors judged each human summary and system summary. In Phase I, for each document cluster, the assessors would read the topic and judge the system and human summaries for both content responsiveness and overall responsiveness. The summaries were judged on the same five point scale as DUC (1-Very Poor, 2 Poor, 3-Barely acceptable, 4-Good, 5-Very Good). The number of summaries with each of these ranks are shown in Figure 5-1.

In Phase 1, for content responsiveness and overall responsiveness, it turns out that systems and humans had somewhat similar overall distributions (Figure 5-1), although the individual systems and humans had quite different distributions; for these details please refer to the MSE presentation [Goldstein et. al 2006b].

For Phase 2, we wanted to assess summaries based on prior knowledge of the content of the documents. For each cluster, the same assessors were asked to read the topic (again) as well as all the documents, which is what we hypothesized a human might do when preparing a summary. Based on this knowledge, the assessors evaluated content responsiveness for each system summary again using the same scale as in Phase 1. The results are shown in Figure 5-2 .

Overall responsiveness (see definition above) is based on the content responsiveness and readability. Since the only other factor in the overall responsiveness as compared to the content responsiveness is the readability, we decided to modify this evaluation for Phase 2. We asked the assessors to form five equivalence classes of system summaries and human summaries for each cluster using a slightly different scale – 1 – Not Acceptable, 2 - Somewhat acceptable, 3 - Acceptable, 4 - Good 5 - Excellent. Note that this scale is slightly harsher than the NIST scale, since level 3 in the NIST scale is barely acceptable and level 3 in this scale is acceptable. Level 4 is the same in both scales "Good". This was partially motivated by the fact that the majority of NIST human summaries achieve a score of "5-Very Good" and the majority of system summaries in DUC did not achieve scores of "1-Very Poor". We thought by stretching the scale, we might learn more information about the quality of human and system summaries. These results are shown in Figure 5-3.



Content Responsiveness: Systems & Humans **Overall Responsiveness: Systems & Humans**

Figure 5-1: Phase 1 - Content Responsiveness and Overall Responsiveness
(8 Systems, 4 Reference Summaries). The vertical axis shows the number of summaries assigned that rank. The horizontal axis shows the rank assigned by the human assessor: Scale: (1-Very Poor, 2 Poor, 3-Barely acceptable, 4-Good, 5-Very Good).

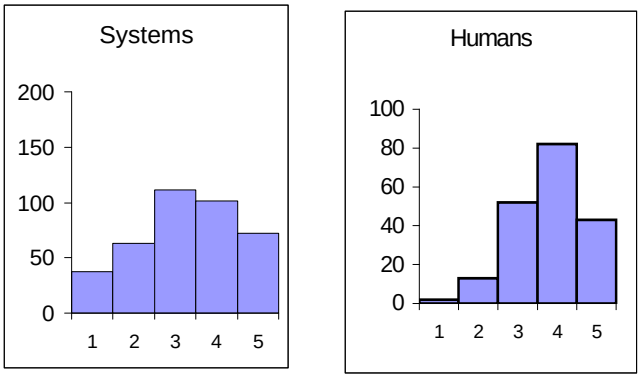


Figure 5-2: Phase 2 - Content Responsiveness (Summary assessment after reading original documents)
(8 Systems, 4 Reference Summaries). The vertical axis shows the number of summaries assigned that rank. The horizontal axis shows the rank assigned by the human assessor: Scale: (1-Very Poor, 2 Poor, 3-Barely acceptable, 4-Good, 5-Very Good).

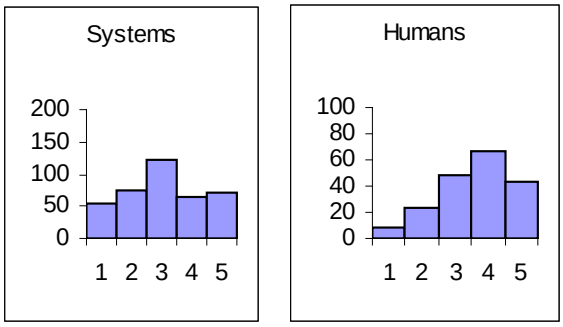


Figure 5-3: Phase 2 - Overall Responsiveness Equivalence Classes
(8 Systems, 4 Reference Summaries). The vertical axis shows the number of summaries assigned that rank. The horizontal axis shows the equivalence class assigned by the human assessor. Scale: (1-Acceptable, 2 Somewhat Acceptable, 3-Acceptable, 4-Good, 5-Excellent).

For both Phase 2 evaluations, the content responsiveness results show a different distribution than in Phase 1. Two significant differences emerge for the Phase 2 content responsiveness:

1. Lower Scores for Summary Quality for both Systems and Humans: The human assessors assigned both human and system summaries lower scores than in Phase 1. In Phase 1, both humans and assessors had more rank 5 summaries than any other rank level. In Phase 2, humans had the most summaries of rank 4 and the systems had the most summaries of rank 3.
2. Different Distribution of Rankings between Humans and Systems: The assessment of system summary quality is lower than that of the human quality. Unlike Phase 1, where the overall distribution is quite similar (Figure 5-1), Phase 2 shows that the human summaries are weighted more towards the top half of the distribution (rank 3-5) and system summaries are weighted more towards the middle (rank 3).
3. More of a Difference between Humans and Systems. The rankings after the assessors read the documents in the cluster, shows that there is more of a distinction between the assessment of quality of human and system summaries as compared to Phase 1. A higher percentage of system summaries, as compared to human summaries, received lower scores when reassessed.

Figure 5-4 shows the comparison of Phase 1 and Phase 2 for content responsiveness. The red line indicates the stability in the summary quality between the two evaluations. Ideal summaries should still have good content responsiveness even after the documents are read. Only one system run was close – 23 (IIT Hyderabad). One human summarizer of the four gold standard summary creators, labeled as D in Figure 5-4, had summaries that were judged by the assessors to not be of similar responsiveness after the full documents were read.

The difference between Phase 1 and Phase 2 shows that for the content responsiveness the evaluators, without prior knowledge of the content in the document cluster, thought the summaries were more relevant to the topic than they did after reading the documents in the cluster. In other words, summaries on their own can be misleading.

In terms of measurements using Rouge Recall (the data is not shown here), one system was grouped with the humans summaries for both content responsiveness and overall responsiveness, IDA-CCS's system CLASSY, which only used sentences from the English newswire documents in summaries to mitigate the effects of machine translation. Both the translated Arabic and English documents were used to create signature terms which were used to select summary sentences. This run (#20) ROUGE scores were better than three of the humans for ROUGE-2 and ROUGE-SU4. Since the other CLASSY run which allowed MT sentences, scored significantly less, we think that this shows that humans do not "like" sentences that are not well formed as are many that are produced using machine translation. Readers interested in the details are referred to the system description [Conroy et. al. 2006, Schlesinger et. al. 2008].

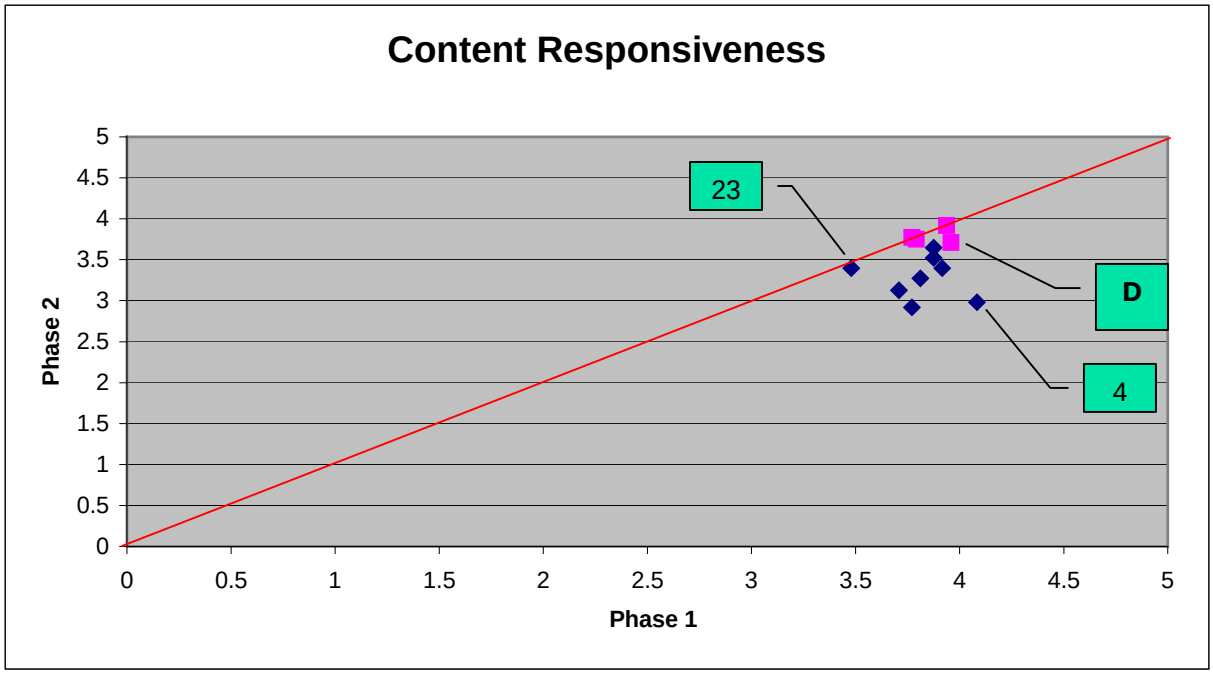


Figure 5-4: Difference in Content Responsiveness between Phase 1 and Phase 2
Humans are in pink Systems are in blue. (Same scale for both phases).
Scale for both: 1-Very Poor, 2 Poor, 3-Barely acceptable, 4-Good, 5-Very Good.

For Overall Responsiveness, we cannot directly compare Phase 1 and Phase 2. The evaluation procedure of using Equivalence Classes as well as the scale in Phase 2 are different from Phase 1.

Figure 5-5 and Figure 5-6 show the tabulated rankings between systems summaries and human summaries for the overall responsiveness equivalence classes in Phase 1 and equivalence classes/overall responsiveness in Phase 2 respectively. In Phase 1, the system and human scores were very similar 3.6 vs. 3.7 in Figure 5-5. Figure 5-6 shows that when the full documents are read and the summaries are put in equivalence classes, the humans now outperform the system summaries with a 0.5 score difference – a score of 3.6 vs. 3.1. The system summaries are now clearly less suitable than the human summaries, or humans are now producing better summaries than the systems. Whereas the system summaries just make "acceptable" (Rank 3), the human summaries are between "acceptable" (Rank 3) and "good" (Rank 4).

Number of System Summaries by Rank						
	1	2	3	4	5	AVG
1	5	6	11	10	16	3.5
10	2	9	11	11	15	3.6
20	3	3	8	14	20	3.9
23	6	5	13	12	12	3.4
4	2	2	5	22	17	4.0
42	8	4	8	14	14	3.5
46	2	9	6	12	19	3.8
6	5	10	5	15	13	3.4
AVG	4.1	6.0	8.4	13.8	15.8	3.6

Number of Human Summaries by Rank						
	1	2	3	4	5	AVG
A	4	7	11	10	16	3.6
B	1	5	8	17	17	3.9
C	7	5	7	14	15	3.5
D	3	6	7	12	20	3.8
AVG	3.8	5.8	8.3	13.3	17.0	3.7

Figure 5-5: Phase 1 - Overall Responsiveness in System Summaries and Human Summaries
1 - Very Poor, 2 -Poor, 3 – Barely Acceptable, 4 – Good, 5 – Very Good

Number of System Summaries by Ranks						
ID	1	2	3	4	5	AVG
1	4	5	23	7	9	3.3
10	3	10	17	10	8	3.2
20	1	12	9	12	14	3.5
23	4	7	18	9	10	3.3
4	11	12	8	4	13	2.9
42	9	9	14	9	7	2.9
46	13	9	15	6	5	2.6
6	8	10	18	7	5	2.8
AVG	6.6	9.3	15.3	8.0	8.9	3.1

Number of Human Summaries by Rank						
ID	1	2	3	4	5	AVG
A	5	4	8	23	8	3.5
B	1	5	12	15	15	3.8
C	1	9	15	13	10	3.5
D	2	5	14	16	11	3.6
AVG	2.3	5.8	12.3	16.8	11.0	3.6

Figure 5-6: Equivalence Class Overall Responsiveness in System Summaries and Human Summaries
1 -Unacceptable, 2 –Somewhat Acceptable, 3 - Acceptable, 4 – Good, 5 – Excellent

Figure 5-5 and Figure 5-6 show that the humans are also producing summaries that are judged unacceptable. In the NIST DUC evaluations [Dang 2006], only a few human summaries were judged as poor (rank 2) for both content responsiveness and overall responsiveness, but none of the summaries were judged as very poor as they were in Phase 1 of MSE. In our evaluation, 15 human summaries in Phase 1 were ranked as "very poor".

The distribution of the NIST human summaries [Dang 2006] shows that NIST human summaries receive a rank of 5 for approximately 75% of the human summaries and a

rank of 2-4 for the other 25%.[6] There are no scores of 1. There could be three possible reasons for this, since the rating of summaries in the NIST evaluation and the MSE use the same scale.

1. Our assessors are rating summaries more harshly than the NIST assessors.
2. The humans who created these summaries might not have created summaries with the same attention and care as the NIST assessors. These human summarizers were probably students as the Linguistic Data Consortium (LDC), who tends to hire students, performed this task.
3. Creating summaries based on English newswire and "noisy" machine translated Arabic newswire is a more difficult task for people. DUC 2004 in which this task was first evaluated had not yet adopted the responsiveness score so we are unable to compare NIST assessor scores of human summaries.

Table 5-2 shows the Cluster Statistics for the Equivalence Classes, the number of the same rankings between the different system summarizers and humans. Note that the summarizers tend to cluster. The average rankings between are as follows:
Human-Humans: 17.5
Systems-Systems: 12.9
Humans-Systems: 11.0

These rankings demonstrate that the system summaries still could use improvement to reach the levels of the human summaries, even in this evaluation where many human summaries were judged unacceptable.

Table 5-2: Cluster Statistics for Equivalence Classes

	Number of Same Rankings for Summarizers & Humans											
ID	1	10	20	23	4	42	46	6	A	B	C	D
1	48	16	12	16	15	17	12	18	6	12	17	13
10	16	48	14	13	10	13	13	12	14	8	11	14
20	12	14	48	18	9	12	10	9	15	21	14	18
23	16	13	18	48	9	14	10	12	9	17	12	17
4	15	10	9	9	48	14	14	14	6	9	12	7
42	17	13	12	14	14	48	15	11	5	13	9	7
46	12	13	10	10	14	15	48	10	7	6	10	7
6	18	12	9	12	14	11	10	48	6	10	16	5
A	6	14	15	9	6	5	7	6	48	14	15	22
B	12	8	21	17	9	13	6	10	14	48	17	20
C	17	11	14	12	12	9	10	16	15	17	48	17
D	13	14	18	17	7	7	7	5	22	20	17	48

[6] These scores were not for the multilingual summaries. Multilingual summarization was evaluated in DUC 2005, but DUC had not yet adopted the responsiveness score.

This study has shown that reading summaries alone is not sufficient to judge them. When the documents were read first there were a difference in scoring between humans and systems, which was not present when the documents were not read first. The adage about not judging a book by its cover could loosely translate into "be wary in judging a document by its automated summary." These results were also demonstrated in SUMMAC [conversation at conference], where one groups' summaries were pulling all the relevant extracts from a document leading the evaluator to believe the document was relevant, when in reality TREC assessors had judged it not relevant to the topic.

In summary, we must be careful when evaluating summaries. The summaries might contain an inherent bias that is reflected when assessed for content responsiveness, which changes when the content of the original documents is known. This issue is not present when measuring overlap between human abstracts and system summaries, which is how we focused our evaluations. As previously mentioned in Chapter 4, the difficulty with evaluations that utilize this type of content overlap matching, is the need for semantic equivalence. An example is the equivalence amount "a man who walks in space" (which implies the man must have traveled to space and is thus an astronaut), "an astronaut" and "a taikonaut". Accordingly, there is a need for sufficient numbers of gold standard human summaries to assess system summaries. These gold standard summaries may or may not be written to address the user's information seeking goals. In contrast, genre-oriented summaries, although they also have these issues, are by nature and by composition focused on a clearly defined user information seeking goal.

In this section we have noted some issues with summary evaluation without knowledge of the original documents. In the next section, we discuss factors that ought to be accounted for when performing evaluations.

5.4 Discussion – Good Summaries and Evaluations

In this chapter, we have discussed, metrics for summarization, the summarization evaluations that have occurred, and summary length. But what factors are necessary for a good summary? It is our view that, at a minimum, a good summary:

1. addresses the *information seeking goals of the user/reader* – the summary should be sufficient for the user's needs
2. has an *appropriate summary length* – how much information the person is willing to read or has time to read.
3. contains effective *content* – users want novel relevant information that doesn't repeat what they already know, including minimal or no redundancy in the information. They also want summaries that are readable (in terms of grammatically), coherent and do not contain information that leads to false implications. Furthermore, they want *goal focused* summaries, summaries that provide the information that they are seeking.
4. is *genre oriented*– a summary needs to reflect the genre for which it is created. A book review or movie review summary has a different nature than a newswire summary. In a newswire summary, it may be necessary to give background and

update the summary as events unfold, whereas a movie review typically does not change over time. The movie review would ideally contain specific information for that genre such as the length of the movie and the rating if available.

With the exception of SUMMAC, most summarization evaluations have focused on intrinsic evaluations without a specific task in mind (generic summarization), or in the most recent years, without an extrinsic task evaluation. SUMMAC used a fixed compression length per document [Mani et. al. 1998]. As shown in Table 5-1, DUC tended to use a fixed byte count, or a word count. The choice of byte count or word count has not been clearly motivated. Perhaps summary length might become more of a factor if extrinsic evaluations were performed.

In our opinion, if summaries were evaluated with regard to a task, the evaluations would most likely produce different results. In the MSE content evaluation, the summaries were supposed to fulfill the information need provided in the topic description. Although not statistically significant due to the small number of clusters and assessors, the MSE evaluation gives some indication that task based evaluation might be important. After the MSE evaluators were aware of the content of the full documents, their opinion of the content and overall responsiveness of the summaries to the topic focus was approximately 14% less than when they judged the summaries without reading the documents [Goldstein et. al. 2006b].

This indicates that perhaps there is a need for the "vital/okay" distinction made in the TREC Question Answering evaluation of other nuggets. When the humans first judged the system summaries (they did not know which were humans and which were system), they may have thought the summaries were extracting and presenting reasonable information nuggets. After reading the full documents, they probably thought many system summaries were missing vital pieces of information or the summaries did not clearly present the vital information.

We have also discussed in this chapter and in the previous one on metrics, that not only is it difficult to judge what pieces of information belong in a summary, but it is difficult to determine semantic equivalence of summaries, even when the summary might be good. The current Pyramid Method [Nenkova and Passonneau 2004] is only comparing summaries at a rough granularity. Using the finer grain factoids proposed by van Halteren and Teufel [van Halteren and Teufel 2003] to compare summaries would be even more labor intensive than the Pyramid Method.

Our last point is that there appears to be clear guidelines for summaries: Brigitte Endres-Niggemeyer summarizes some of these in her book *Summarizing Information* [Endres-Niggemeyer 1998]. We have discussed these in Section 2.2.2. She has six items by which to measure relevance (fact, topic, purpose, relpositive, contrast and stress). There were four meaning reduction strategies that also could serve as a guide for eliminating content in summaries (noreason, novoid, nocomment, noexample) and also to assess the quality of the summary. For example, "Does the summary contain comments and explanations (not desired)?". She also provides seven ways that could be used to assess importance, relevance and interestingness in a summary, several of which could be used

to assess summarization quality – such as "Does the summary convey the author's intention and design?" She also suggest that a summary must have information value, innovation value and interestingness; information value is evaluated through content responsiveness, but innovation value and interestingness were not evaluated in DUC or NTCIR evaluations.

Endres-Niggemeyer's criteria is designed around the methodology that professional summarizers use and could serve as the basis for the design of more comprehensive summary quality metrics than are currently used.

The next two chapters discuss the genres in our work and genre oriented goal-focused single document summarization. We then examine the email genre.

Chapter 6 Genre Identification

"Subject analysis is the first, the most important and the most difficult part of all classification and indexing. No retrieval system can be better than the subject analysis on which it is based."

D.W. Langridge, *Subject Analysis: Principles and Practice* 1989.

We suggest that genres are key to effective goal-focused summaries (discussed in detail in Chapter 1). As motivated in Section 2.3, we are using the definition of a genre as "a pattern of communication that has purpose, form and content" and are using *genre-topic*, i.e., topical content, as a way to distinguish genres, such as a movie review vs. product review. As such, the genre contains key information that the summarization system can use to tailor its summary. For example, if the genre/genre-topic is that of movie reviews, then a genre oriented summarizer can extract the audience rating (PG, G, R), the rating of the movie as well as the running time and provide this information in a summary. If the genre is a product page or a product review, the product description and cost can be provided to a user. There are many such genre specific information nuggets that are highly relevant only to their own particular genre/genre-topic. As an example, the audience rating of movies is not relevant to a product press release or a biographical summary. In this chapter, for the purpose of genre- identification we refer to these genre-topics as genres.

In order to determine whether or not it is feasible to construct a genre oriented summarization system, we must first determine at what level genre identification can be performed and at what cost (quantity of documents that must be identified for genres). If the classification accuracy of genre identification is too low, the summarization system is not be able to effectively utilize the passed information to produce genre tailored summaries. We think an overall classification accuracy of at least 80% is important for the genre information to be used in a summarization system.

In this chapter, we motivate the concept of genre identification for both categorization and downstream processes such as summarization. We discuss the performance of document genre identification on our data sets to its suitability for use in producing tailored summaries – *genre oriented summarization*. In order to test the feasibility of this concept, we create a test corpora to determine genre classification accuracy for 9 and 16 genres, using subsets of 119 features which consist of layout features, character features, infrastructure features, lexical and topical content based features. To our knowledge, this is the largest and most diverse web corpora available for genre studies. We also examine performance on various amounts of training data for machine learning algorithms - Random Forests, SVM light and Naïve Bayes. We end this chapter by discussing the numbers of documents that might be required for effective performance as well as the effects of adding random web pages on performance.

6.1 Introduction

With the continuing rapid growth of the Web, it is becoming essential to find ways of categorizing and filtering information to be able to quickly locate and access particular items of interest. The genre of a document can be used for such categorization purposes. Crowston and Kwasnik define an "information-access system" as having three components [Crowston and Kwasnik 2003]:

(1) the users who have information seeking needs, often contextually-based,
(2) the store of information
(3) an intermediating mechanism of connecting needs and the information, which may be a search algorithm, a browsing environment, a summarizer, or a person, among others.

For the internet, this mechanism has focused around search algorithms and short summaries intended to assist the user in choosing relevant web pages to view. Some engines include clustering processes, such as the Search Engine Clusty (previously Vivisimo), which are usually organized around topically based keywords or concepts. However, topic alone may be insufficient for categorizing documents/web pages. For any given product, one user may be interested in a product press release, another user may be interested in reviews about the same product and a third might be interested in the stores that are offering the product for the cheapest price. A user may decide what page they wish to view based on this genre.

Thus, genre which takes into account topic (*genre-topic*) (a product review page vs. a movie review page) may be a preferable way to organize and label documents. Genre information can be determined from the web page and used as categorical metadata about the document as a means of indexing and retrieving documents. Roussinov and colleagues, reported that the genre of the document was used by web searchers to assess relevance, value, quality and usefulness [Roussinov et. a. 1991].

Besides being used as a high level filtering and relevance assessment tool, the genre of a document can be used to craft summaries and include information that might be otherwise missed. If the genre is recognized as a biography, the system can focus on pulling out personal and professional facts that people typically like to have in such summaries. For movie reviews, the specialized item of audience rating and running time can be extracted. The availability of the genre information allows the summarization system to focus extraction efforts and summary creation efforts on items that result in a more effective user-tailored summary.

In the next section, we describe our selection of genres and preparation of the data set, which conforms to this type of expectation. The genres and genre-topics were chosen with the purpose of examining the effects the results, including misclassification, would have on summarization. We discuss our feature extractor, and the classification experiments we performed using Random Forests, SVM and Naïve Bayes.

6.2 Related Work

Genre classification has been studied since the early 1990s, where Karlgren and Cutting used discriminant analysis to classify four categories in the Brown Corpus (ranging from high level categories press, misc., non-fiction and fiction) using features derived from part of speech analysis [Kalgren and Cutting 2004]. They reported 27% error. Kessler and colleagues also used such structural cues as well, but also used lexical cues (such as terms of address), character-level cues (main frequency of punctuations) and derivative cues (ratios and variation measures derived from measure of lexical and character-level features, such as average word length and average sentence length) [Kessler et. al. 1997]. They used logistic regression as a classifier based on initial pilot studies showing that it gave better results than linear discrimination and linear regression. Their results indicate that there is only a marginal advantage to using the structural cues, which does not justify the additional computational cost. Stamatatos and colleagues used the Wall Street Journal for their approach and attempted to classify press genres using most frequent words and character level cues [Stamatos et. al. 2000]. Finn and Kushmerick classified subjectivity and positive or negative reviews with the C4.5 learning algorithm with features including the most frequent words, the structure cues (such as part of speech statistics) and the derivative cues on different topic domains (one set of three and one set of two). [Finn and Kushmerick 2003].

Dewdney and colleagues used seven genres and three classifiers Naïve Bayes, C4.5, and SVM-light [Dewdney et. al. 2001]. They used 89 features consisting of word frequency features (based on information gain [Yang and Pederson 1997] and derivative features. These combined features led to better results than that of just the word frequency or derivative alone and resulted in precision and recall scores above 82% (with an unknown category allowed in the classification results). Carvalho and Cohen recently used the Linear SVM classifier to classify email speech acts into 6 categories and by introducing contextual information through n-grams, was able to achieve at least 80% accuracy for each category, a relative error rate drop over their previous work ranging from 9-30% depending on category [Carvalho and Cohen 2006].

Our research differs from previous work in that we are examining the potential use of genre identification in a system and as such are using a greater number of categories (up to 42) than prior researchers. The largest data collection of which we are aware of consists of 7 distinctive web genres (200 blogs, 200 eshops, 200 FAQs, 200 online newspaper front pages, 200 listings, 200 personal pages and 200 search pages) [Santini 2006]. Unlike other research efforts, we have also chosen some of our categories with a high degree of topical overlap to be able to investigate the effects of genre identification in such potential confusable conditions: e.g., editorials and articles with a topic of President Bush, as well as product overlap from product press releases, product reviews and product store pages. Some of our high level categories are topically subdivided as well, this allows us to examine the effects of a small number of documents on classification accuracy.

We also examine the machine learning algorithm of Random Forests. To our knowledge, this classifier has not been compared with SVM for genre classification or text categorization.

6.3 Genre Identification Data

In this section, we describe the collection and preparation of the genre specific document data sets for genre classification. The goal is to examine the effects of genre classification output on summarization, accordingly we collected data that reflected the summaries that we wanted to produce.

We selected nine categories for our genre-id experiments (9-SUM-GENRES), seven of which were targeted towards summarization experiments. We had several students collect data on the web aiming for a total of 1000 documents per every genre. A maximum of 10 data items were collected from any site on any particular topic/product with the exception of editorials-and articles. The genres are biographies, interviews, movies, articles-politics, editorials-politics, product-press-releases, product-reviews, search-results, and store-products. The number of documents collected are shown in Table 6-1. The data collected resulted in approximately a total of 1000 documents per category since some files just contained links and no content and were deleted in a data clean-up pass. A maximum of 10 data items were collected from any web site. We added 1000 randomly selected documents from 7 additional categories (7-CMU) that were collected by CMU to form a total of 16 genres (16-GENRES) (Table 6-1). Although many of the 7-CMU genres had more than 1000 documents, for our initial studies, we wanted to keep the amounts of documents per genre fairly even so we only used at most 1000.

Genres were chosen with the objective of performing specific experiments that reflect real world considerations. Accordingly, several categories were chosen to see how well we could classify the results. Biographies were split into 8 subtopic areas: actors, artists, authors, directors, leaders of historical empires, musicians, scientists, sports, so that we could examine categorization results at this level (Table 6-2). The size of each category varied from 30-175 with an average of 120 per category. Store products consisted of 20 different topics of approximately 50 items each (Table 6-3.)

The topic of politics-articles and politics-editorials contained only articles about President Bush, in order to ascertain how well the classifier could split editorials and articles about a single person.

Table 6-1: Genre Document Collection – 16 genres

Genre - SUM	# of Sub-categories	# of docs	Genre - CMU	# of docs used	# original docs
Biographies	8	959	Advertisements	1000	1091
Interviews		1025	Bulletin Board	998	998
Movie Reviews		968	FAQs	1000	1063
Editorials-politics		1006	Message Board	1000	1106
Articles-politics		1092	Radio News	1000	2000
Product Press Releases		992	Reuters	1000	1836
Product Reviews		930	TV News	1000	1462
Store Products	20	954			
Search Results		956			

Table 6-2: Number of Documents in the 8 Biographies 8 Subcategories

Biographies Genre Subcategory	# of Doc
Actors	161
Artists	185
Authors	175
Directors	30
Historical Empire Leaders	52
Musicians	167
Scientists	97
Sports Figures	92

Table 6-3: Number of Documents in the 20 Products Subcategories

Products Genre Subcategory	# of Documents
DVD player	50
MP3 players	50
PDAS	50
Perfume	50
Running Shoes	36
Televisions	50
Watches	48
Camcorders	50
Cars	41
Cases, bags	50
Cellular accessories	50
Fitness	38
Game Consoles	50
Guitars	50
Notebooks	50
Pagers	50
Radar Detectors	49
Remote Controls	50
Synthesizers	51
Video Games	50

6.4 Classifiers and Features

Two different classifiers were used for our experiments: Support Vector Machines and Random Forests. We used SVM light with a radial basis function and the default settings. SVM light builds binary models, so the time needed to build models in cases of many categories can be quite significant.

In contrast, Random Forests grows many classification trees and each tree gives a vote for a particular class. The forest chooses the classification having the most votes. We use 100 trees in our experiments. The training time is significantly less than that of SVM-light.

Three different classifiers were used for our experiments. WEKA's implementation of Naïve Bayes [NaiveBayes], Joachims' implementation of Support Vector Machines [Joachims 1998] and Random Forests [Breiman 2001]. We used the default settings for Naïve Bayes. SVM has been shown to perform well for text categorization [Yang and Liu 1999]. We used SVM light [SVM_LIGHT] with a radial basis function and the default settings. SVM-light builds binary models, so in our cases, where we have 8-42 classes, a model must be produced for each class, which can be quite time consuming.

In contrast, Random Forests [Breiman 2004] grows many classification trees and each tree gives a vote for a particular class. We use 100 trees in our experiments. An integer *mtry* is set by the user (we use the square root of the number of features as suggested by Leo Breiman). This number represents the number of variables selected at random at each node. The node is split on the best split among the selected mtry and the tree is grown to its maximal depth. In regression, as a test vector x is put down each tree, it is assigned the average values of the y values at each node it stops at. The average of these over all trees in the forest is the predicted value for x. The predicted value for classification is the class with the majority of the forest votes. Random Forests has the characteristic that the training time is far less than of SVM light, especially for large numbers of genres (categories). For the 16 genres, SVM took approximately twice as long as RF due to the fact that SVM trains binary classifiers for each genre. Joachims has now developed a version of SVM using multi-classes [SVM-Multiclass 2007, Crammer and Singer 2001]. Joachims states that this version is "orders of magnitiude faster". For chunking tasks multiclass SVMs have been reported to be 100 times faster that state-of-the-art SVM methods [Wu et. al. 2008].

Similar to previous research (except Cohen's latest work using n-grams [Carvalho and Cohen 2006]) and expanding on Dewdney's research [Dewdney et. al. 2001], we use 119 features in our feature extractor, a combination of content word frequency features and lexical and derivative cues. We did not modify our feature vector for these experiments – some of our features are designed towards web document identification in general. The derivative cues consist of three types of layout features (13), character features (23), infrastructure features (30), grammatical features including part of speech features (26) and genre structural content words and topical content words (27). The baseline of 66 items includes the layout features, character features and infrastructural features (e.g., number of white space lines, number of quotes, counts of words appearing in headings)

plus derivative cues (e.g., readability measures). The layout features, character features, and structure features compose our baseline of 66 features. We can think of this as out of the box classification, as no genre specific content words are used in the classification.

- *Layout features* attempt to determine the format of the document by examining items such as lines containing only white space or only a series of separators, or blocks of contiguous non-blank lines.
- *Character features* examine the proportion of various characters in the document, which can be indicative of particular document types such as chat or interviews (which might have many colons), as well as the Flesch readability score.
- *Structure features* are designed towards particular genres such as forums and newsgroups, which have items such as "To:" and "From:". This group also contains items such as words per line, words per sentence, the word size and the proportion of alphanumerics, all of which can be calculated independent of any specific language.
- *Grammatical/Semantic feature* are designed towards identifying certain classes of words. These include the proportion of first person, second person, third person, and word lists such as temporal words, colors, human familial words and pronouns.
- *Content based features* contain lists of words that are geared towards identification of certain specific genre categories such as forums, newsgroups and news (which contain category specific words such as post, subscribe, join) as well as topic content based features. As a first pass, the topic content features were extracted by analyzing a small independent sample (50 documents of the same type) and extracting high frequency words (excluding stop words) that occurred in more than 20 documents and more than 30 times.

6.5 Experimental Results

The evaluation results for the three classifiers on the 9 and 16 genres using 10 fold cross validation are shown in Table 6-4. We calculate our results using precision, recall and F_1 with two types of averaging: the micro-average, where each relevant document is a point in the average, and macro-average, where each genre is a point in the average. For cases where the number of documents are similar in all genres, the micro-average score is very similar to the macro-average score. Table 6-4 F_1 (micro-average) shows that Random Forests achieved a better performance than SVM, a result that occurred consistently across the data. This result was most pronounced with no topical features (our baseline of 66 features) and the two topic confusable categories, articles-politics and editorials-politics, which entirely consisted of stories that discussed President Bush (refer to confusion matrices in Table 6-7 and Table 6-8).

Table 6-4 also shows that performance is better for all algorithms on the 16 genres, perhaps due to the addition of more training data. Surprisingly, Random Forests (RF) performs extremely well just using our baseline 66 general features (no lexical items or part of speech items and no topic based content words). The 66 general features are designed to be language content independent. The version of Naïve Bayes that we used

did not come close to the scores for RF and SVM. For the 9 SUM Genres and 16 genres, Naïve Bayes scores ranged from 0.48-0.71, whereas RF scores ranged from 0.81-0.90 and SVM from 0.67-0.85. Since Naïve Bayes scores were remarkably less than the other two classifiers, we chose not to experiment with it in the rest of our explorations.

Using RF with baseline features, the addition of the extra 53 grammatical, content and topical features only accounted for an approximate 0.06 increase in performance for the 9 SUM Genres and 0.04 for the 16 Genres. For SVM, there was a 0.14 increase in perform for the 9 SUM Genres and 0.09 for the 16 Genres. We need to further explore which features are contributing to an increase in the scores. In the case of RF, the 06 increase in performance does not seem worth the overhead of computing 53 features.

Table 6-4: Classification Results for 3 sets of Features for 9 and 16 genres
10 fold cross validation. Random Forests (RF), Support Vector Machine (SVM) and Naïve Bayes (NB).

	9 SUM Genres			16 Genres		
Features \ Classifier	RF	SVM	NB	RF	SVM	NB
Baseline (basic) (66)	0.81	0.67	0.48	0.86	0.76	0.63
Baseline+grammatical (92)	0.85	0.78	0.58	0.88	0.83	0.69
Baseline+grammatical+content (119)	0.87	0.81	0.63	0.90	0.85	0.71

The detailed precision and recall per category for RF and SVM for 9 genres (Table 6-5) and 16 (Table 6-6) genres respectively. RF and SVM have similar scores in many genres. However, SVM is not able to perform as well as RF in the topically confusable categories such as articles and editorials with the topic of President Bush and product press releases and product reviews. For SVM, movie reviews are often also confused with interviews, this could possibly because the interview category contain interviews with directors and actors/actresses.

Table 6-5: Overall within genre statistics for RF and SVM, 9 genres, 119 features.

	Random Forests			SVM		
Genre	P	R	**F_1**	P	R	**F_1**
biographies	0.83	0.89	**0.86**	0.80	0.79	**0.79**
interviews	0.87	0.94	**0.90**	0.79	0.89	**0.84**
movie reviews	0.86	0.86	**0.86**	0.80	0.73	**0.76**
editorials (topic Pres. Bush)	0.91	0.78	**0.84**	0.77	0.72	**0.75**
articles (topic Pres. Bush)	0.82	0.94	**0.88**	0.78	0.88	**0.83**
product press releases	0.81	0.89	**0.85**	0.79	0.85	**0.82**
product reviews	0.83	0.64	**0.72**	0.76	0.57	**0.65**
store products	0.93	0.95	**0.94**	0.87	0.93	**0.90**
search results	0.94	0.88	**0.91**	0.92	0.90	**0.91**

Table 6-6: Overall within genre statistics for RF and SVM, 16 genres, 119 features.

	Random Forests			SVM		
Genre	P	R	**F_1**	P	R	**F_1**
advertisements	0.86	0.92	**0.89**	0.83	0.86	**0.85**
biographies	0.82	0.89	**0.85**	0.82	0.77	**0.79**
bulletin-board	0.90	0.80	**0.85**	0.86	0.80	**0.83**
FAQ	0.91	0.99	**0.95**	0.95	0.99	**0.97**
interviews	0.87	0.94	**0.90**	0.79	0.89	**0.84**
message board	0.99	0.99	**0.99**	0.98	0.99	**0.99**
movie reviews	0.88	0.85	**0.86**	0.81	0.73	**0.76**
editorials (topic Pres. Bush)	0.92	0.78	**0.84**	0.76	0.73	**0.75**
articles (topic Pres. Bush)	0.82	0.94	**0.88**	0.78	0.89	**0.83**
product press releases	0.82	0.89	**0.85**	0.79	0.85	**0.82**
product reviews	0.84	0.65	**0.73**	0.76	0.56	**0.64**
radio news	0.98	0.86	**0.91**	0.94	0.73	**0.82**
Reuters	0.99	0.99	**0.99**	0.98	0.98	**0.98**
store products	0.94	0.89	**0.91**	0.87	0.92	**0.90**
search results	0.93	0.96	**0.95**	0.92	0.90	**0.91**
TV news	0.90	0.99	**0.94**	0.80	1.0	**0.89**

Table 6-8 show the confusion matrices for Random Forest and SVM results for the 9 SUM Genres. As we might have expected, the Editorials are confused with the Articles on the same topics. For Random Forests, these two categories are primarily confused with each other (Table 6-7). For SVM, there is some slight confusion with other categories such as biographies, movie reviews and product press releases.

The category with the worse performance is Product Reviews. They are highly confusable with Product Press Releases as well as other genres. Random Forests still does better than SVM at distinguishing these topically overlapping categories for both precision and recall (F_1 of 0.73 compared to 0.64). We need to investigate developing features that might improve their categorization.

Some categories, in particular message boards and Reuters received very high scores - 0.99 for RF, and 0.99 and 0.98 for SVM. It is interesting that Reuters was not confused with the articles since both come from the newswire genre. Even with the baseline of 66 features, there was no confusion between Reuters and the articles. In future work, we will explore what features are contributing to the successful identification of these two genres.

Next we analyze the classification accuracy of smaller numbers of documents within genres. As mentioned, some of the genres were designed to have topical subcategories, such as the biography genre (Table 6-2) and the store product genre (Table 6-3).

Table 6-7: Confusion matrix (%) for Random Forest, 9 SUM genres, 119 features.

Confusion Matrix RF									
Genre	**Bio**	**Int**	**Mov**	**Ed-P**	**Art-P**	**PPR**	**PR**	**Srch**	**SP**
Bios	**89**	3	4	0	0	2	0	0	0
Interview	2	**94**	3	0	1	0	0	1	0
Movie Reviews	6	5	**86**	0	0	1	1	1	0
Ed-Politics	0	0	0	**79**	29	0	0	0	0
Art-Politics	0	0	0	6	**94**	0	0	0	0
Product Press Releases	3	0	1	0	0	**89**	6	0	1
Product Reviews	5	6	4	1	1	13	**65**	3	3
Search	1	0	1	0	0	1	1	**96**	1
Store Products	1	1	0	0	0	5	4	1	**88**

Table 6-8: Confusion matrix (%) for SVM, 9 SUM genres, 119 features

Confusion Matrix SVM									
Genre	**Bio**	**Int**	**Mov**	**Ed-P**	**Art-P**	**PPR**	**PR**	**Srch**	**SP**
Bios	**79**	6	6	1	1	4	1	2	0
Interview	2	**90**	4	1	0	0	1	2	0
Movie Reviews	8	10	**73**	2	0	2	3	2	1
Ed-Politics	1	0	1	**73**	24	1	0	0	0
Art-Politics	0	0	0	10	**89**	0	0	0	0
Product Press Releases	2	1	1	1	0	**86**	7	1	1
Product Reviews	4	5	5	6	1	12	**58**	5	4
Search	1	2	1	0	0	1	1	**93**	1
Store Products	1	1	1	0	0	3	3	1	**90**

Most of the F1 scores for the 8 subcategories in the biography genres ranged between 0.62 and 0.88 (Table 6-9) with the exception of subcategory of Directors, which had a low F_1 score of .027. The confusion matrix in Table 6-10 shows a strong confusion with actors, authors, and musicians. However, the Leader category, which had 53 documents achieved an F_1 score of 0.88, showing that it may not be just the number of documents that caused the poor performance for Directors. However, when we add 75 more documents to this category, the F_1 scores increases to 0.78 from 0.27 (Table 6-9). Table 6-11 shows that the confusion among directors and the other categories has greatly decreased. There is little change in the overall performance of the other subcategories of the biography genres that were not confused with the Directors (Table 6-9).

The next group of subcategories for the biography genre that are highly confused with each other are Actors, Authors and Musicians (Table 6-11) with "low" F_1 scores (0.67, 0.67 and 0.68 respectively). There is also some confusion with Artists. Scientists have

some confusion with Authors but Authors only have a little confusion with Authors. In future work, we will explore whether adding topically related words to distinguish these subcategories improves results.

Table 6-9: Classification Results for Biography Genre before & after adding documents 75 documents to Directors, RF 119 features

	30 Director Documents			105 Director Documents		
Genre	P	R	**F_1**	P	R	**F_1**
Actors	0.70	0.64	**0.67**	0.69	0.65	**0.67**
Artists	0.81	0.84	**0.82**	0.80	0.86	**0.83**
Authors	0.60	0.66	**0.62**	0.62	0.72	**0.67**
Directors	0.83	0.16	**0.27**	0.94	0.67	**0.78**
Leaders -Historical Empires	0.95	0.83	**0.88**	0.97	0.83	**0.89**
Musicians	0.56	0.73	**0.63**	0.62	0.76	**0.68**
Scientists	0.82	0.77	**0.79**	0.85	0.76	**0.80**
Sports - Athletes	0.95	0.72	**0.82**	0.95	0.72	**0.82**

Table 6-10: Confusion Matrix for Eight Topical Categories of Biographies with 30 items in Director Category, RF 119 features

Genre	**Items**	**Act**	**Art**	**Aut**	**Dir**	**Lea**	**Mus**	**Sci**	**Spo**
Actors	161	**65**	3	11	0	0	19	1	2
Artists	185	0	**84**	7	0	0	8	2	0
Authors	175	7	10	**67**	0	1	13	3	0
Directors	30	20	0	30	**17**	0	30	3	0
Leaders -Historical Empires	52	2	0	8	0	**83**	0	8	0
Musicians	167	12	3	9	1	0	**73**	2	0
Scientists	97	3	5	10	0	1	5	**77**	0
Sports - Athletes	92	4	1	9	0	0	13	1	**72**

Table 6-11: Confusion Matrix for Eight Topical Categories of Biographies with 105 items in Director Category, RF 119 features

Genre	**Items**	**Act**	**Art**	**Aut**	**Dir**	**Lea**	**Mus**	**Sci**	**Spo**
Actors	161	**66**	4	13	1	0	14	1	2
Artists	185	0	**87**	7	0	0	5	2	0
Authors	175	6	8	**73**	0	1	11	2	0
Directors	105	13	1	10	**68**	0	8	1	0
Leaders -Historical Empires	52	0	4	6	0	**83**	2	6	0
Musicians	167	10	4	8	0	0	**76**	1	0
Scientists	97	5	6	8	0	0	5	**77**	0
Sports - Athletes	92	4	2	8	1	0	13	0	**72**

We can also observe a similar decrease in performance for the 954 store product pages subcategorized into 20 store products. There are 36 to 50 documents per category and F_1 scores varying from 0.27 to 0.84 (Table 6-12). Some categories such as cars and running shoes had very few misclassification errors, whereas some categories such as MP3 players and televisions had many errors with the other categories. We hypothesize that adding documents to these categories would also increase classification accuracy as it did in the case of the 8 topical categories of the biography genre. The confusion matrix for is shown in Table 6-13.

Notebooks (laptops) had the lowest F_1 score (0.27). This category was most confused with PDAs, MP3 players and DVD players, although there were several other categories as well in which notebooks were categorized. MP3 players (F_1 score of 0.31) were frequently confused with PDAs, Synthesizers and Guitars.

The macro-average and micro-average for these store products are shown in Table 6-14.

Table 6-12: Classification Results for Store Products Genre, RF 119 features

Products Genre Subcategory	# of Documents	Precision	Recall	F_1
DVD player	50	0.56	0.6	**0.58**
MP3 players	50	0.34	0.3	**0.31**
PDAS	50	0.34	0.38	**0.36**
Perfume	50	0.61	0.64	**0.62**
Running Shoes	36	0.81	0.83	**0.82**
Televisions	50	0.46	0.28	**0.35**
Watches	48	0.62	0.70	**0.66**
Camcorders	50	0.4	0.36	**0.37**
Cars	41	0.78	0.73	**0.75**
Cases, bags	50	0.60	0.64	**0.62**
Cellular accessories	50	0.65	0.58	**0.61**
Fitness	38	0.88	0.81	**0.84**
Game Consoles	50	0.59	0.5	**0.54**
Guitars	50	0.48	0.72	**0.57**
Notebooks	50	0.36	0.22	**0.27**
Pagers	50	0.59	0.58	**0.58**
Radar Detectors	49	0.43	0.59	**0.50**
Remote Controls	50	0.37	0.38	**0.38**
Synthesizers	51	0.39	0.45	**0.42**
Video Games	50	0.75	0.78	**0.76**

Table 6-13: Confusion matrix (%) for 20 Categories of Store Products, RF, 119
20 categories: DVD, MP3 Players, PDAs, Perfume, Running Shoes, TVs, Watches, Camcorders, Cars, Cases, Cellular Accessories, Fitness, Game Consoles, Guitar, Notebooks, Pagers, Radar Laser Detectors, Remote Controls, Synthesizers, Video Games

Cat.	#	D	MP	PD	PE	RU	TV	W	CA	CR	CS	CL	FI	GA	GU	N	P	RL	RC	S	SP
DVD	**50**	**60**	2	4	2		12	2	4		2			2		4		4		2	
MP3	50	4	**30**	14			2		4	2		2		2	10	6	2	6	4	12	
PDA	50	2	8	**38**			2		20		2			4	2	6	2	4	2	4	4
Perf	50	2		2	**64**			16		4					8					4	
Run	36					**83**	3						11		3						
TV	50	20	6	10	2		**28**	2	2	2				2	4	2	2	4	2	10	2
Wat	48		2		15			**71**			4				8						
Cam	50	2	8	14					**36**		4			6		6	8	6	6	2	2
Car	41			2	5			5		**73**					10		5				
Case	50								4		**64**	4		2			2	14	8	2	
Cell	50	2	4				2		4		6	**58**				4	2	4	10	4	
Fit	38					18							**82**								
Game	50	2	2	8					4			2		**50**				8	2	6	16
Guit	50				6			8		6					**72**					4	2
Note	50	8	10	14	6			2	6			2		4	6	**22**	6	2	6	4	
Page	50			4	2		4	2		2		2			12	2	**58**	8		4	
RLD	49	2	6				2		2		8	4		4				**59**	8	4	
RemC	50		8					4	2		14	10		2		4	6	8	**38**	4	
Synth	51	2	2				2		2		2				18	2	2	6	14	**45**	
VG	50						4					4		6		2	4		2		**78**

Table 6-14 shows the micro-average and macro-average precision, recall and F1 for various experiments. The range of documents per categories of genres are shown as well as their scores. The number of genres as well as subcategories within genres and the number of documents per genres clearly makes a difference in performance. As we have shown from the subcategories in the bio genre and in the store product genre, some of these topical subcategories can be easily confused with each other. The largest numbers of categories for which we ran experiments is 42: the 20 subcategories of store products, the 8 subcategories of biographies, the remaining 7 basic SUM genres, and the 7 CMU genres.

Table 6-14: Classification results for various combinations
10 fold cross validation, RandomForests,, 119 features.

Set	Num Classes	Data	Micro-average			Macro-average		
			P	R	F_1	P	R	F_1
9 SUM genres	9	~1000 per genre	0.87	0.87	0.87	0.87	0.86	0.86
7 CMU genres	7	~1000 per genre	0.95	0.95	0.95	0.95	0.94	0.94
7 CMU genres	7	Full data (Table 6-1)	0.96	0.96	0.96	0.94	0.95	0.94
9 SUM + 7 CMU	16	~ 1000 per genre	0.90	0.90	0.90	0.90	0.89	0.89
8 bio categories	8	53-195 per category	0.75	0.75	0.75	0.80	0.75	0.77
8 bios + 8 SUM	16	53-1092 per category	0.84	0.84	0.84	0.83	0.71	0.75
20 store products	20	38-51 per category	0.55	0.55	0.55	0.55	0.55	0.54
20 store + 8 SUM	28	38 – 195 per category	0.82	0.82	0.82	0.70	0.56	0.60
20 store+8 bio+7 SUM	35	38 – 1092 per category	0.80	0.80	0.80	0.71	0.56	0.60
20 store + 8 bio + 7 SUM + 7 CMU	42	38-2000 per category	0.86	0.86	0.86	0.75	0.61	0.64

Next we examine the overall effects on training/test corpus on results for both our baseline 66 features and our full set of 119 features (Figure 6-1). Note that for Random Forest, after 250 documents, the increase in collection effort for the extra documents might not be worth the performance gain.

As a last experiment, we wanted to determine the effects of unknown pages, not necessarily from our set of genres, on system performance. To approximate this condition, we downloaded 930 random web pages, no more than 10 per website and included this "genre" in our 10-fold cross validation experiments. Since it is a small sample of the variety of web pages that can exist, it is not necessarily representative of what could occur. However, it is a first approximation.

The results for RF and SVM are shown in Figure 6-2. They are compared to the original 16 genres. The addition of the random files affected SVM's performance more than RF's.

In conclusion, in this chapter, we have shown that we can get excellent overall genre identification performance of approximately 0.9 F_1 (Figure 6-1) with approximately 750 documents per genre. An increased number of genres appear to assist performance scores. 100-250 documents with some carefully chosen lexical and topical features seem to result in performance scores of about 0.8 F_1. These scores seem adequate for use in a genre oriented summarization system, which is discussed in the next chapter.

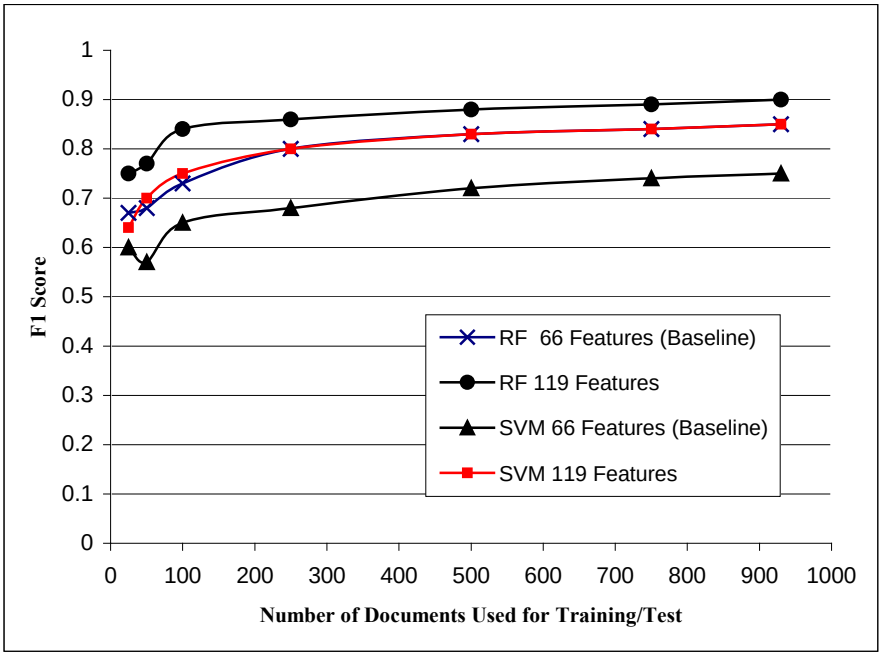


Figure 6-1: Classification Performance Effects based on Number of Documents Random Forests (RF) and SVM, 16 genres, 66 features (baseline) and 119 features (10 fold cross validation).

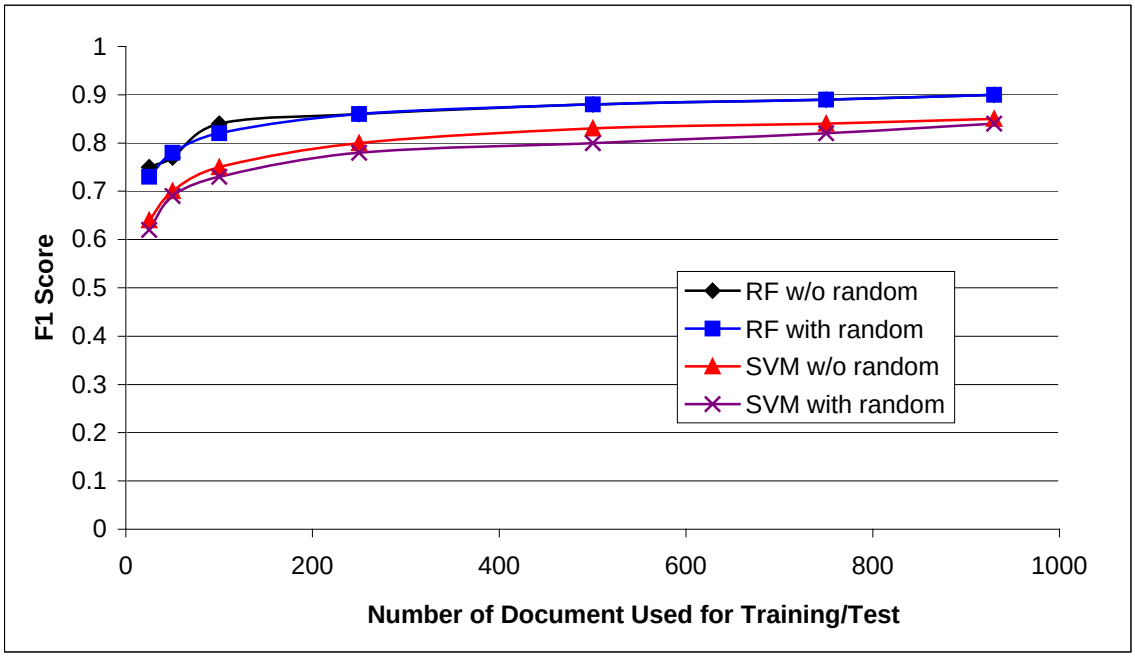


Figure 6-2: Classification Performance Effects based on Number of Documents Random Forests (RF) and SVM, 16 genres, 119 features with the additional random files genre (17 genres) and without the random files (16 genres from Figure 6-1) 10 fold cross validation.

Chapter 7 Genre Oriented Goal-Focused Summaries

"Knowing how people will use something is essential."
Donald Norman.

As previously mentioned, summarization systems have focused on two types of summaries: the *generic* or *overview* summary, which gives an overall sense of the document's content, and a *query-based* summary, which presents the content that is most closely related to the initial search query. Most summarization research has focused on the news article genre and the scientific article genres [Mani et. al. 1998, Teufel and Moens 1997, DUC] for which these two types of summaries work well. Consider, however, the genre of movie reviews. A user may want a variety of information: an overview of the review (generic), a specific answer to question (query-based), details of the plot, opinions on the quality of the movie and/or acting, or details on what audience for which it is suitable and the reasons. Accordingly, the movie review genre, as do others, require a new class of summary - that of the *genre oriented goal-focused* summary.

This chapter discusses genre oriented summarization: the data preparation, the experiments performed and the results. These results are also discussed in light of the results of the previous chapter, which suggested that the error rates obtained for genre identification, which can be on average 14% for 42 categories, are probably sufficient to successfully inform a downstream summarization system. Such a system could use the identified genre information to construct summaries that more effectively address a user's information seeking goals.

The evaluation of any system requires data sets for either an automatic evaluation or one based on a ``gold standard'' using human summaries or judgments. In Chapter 4, we discussed the difficulties with evaluating summarization systems, namely the lack of adequate automatic evaluation techniques, the sparsity of human judgment data and the variance in human opinions. Nevertheless, since scoring of systems is a necessity in measuring the quality of the system, we used three human summarizers to form an evaluation corpus.

The collection process for the gold standard corpus is described in this chapter. This includes the process of identifying goal-focused summaries for a set of genres and determining their composition. We describe our summarization algorithm for extracting sentences from the genres.

We end by comparing our summaries to the gold standard summary formed by concatenating the human selected sentences.

7.1 Genre Oriented Document Data Set Description

In this section, we describe the collection and preparation of the genre specific document data sets for summarization. As discussed in Section 2.3 and the previous chapter, genres are defined as a particular class of documents, with form, purpose and content. There can also be a topic separation (genre-topic). Our overall category is web pages, which is composed of several genres such as movie reviews, store product pages, and biographies.

The goal of genre specific document summarization is to be able to apply ``rules'' tailored to the particular genre, (e.g., movie, biography) in order to provide a more useful summary for the users' needs. For example, one might want to have a summary of plot or opinion from a movie review summary, in contrast to the most important points of a newswire article.

Accordingly, we selected seven genres: movie reviews, product reviews, product press releases, interviews, biographies, editorials and news articles. We collected 30-45 html samples from each class from various websites on the internet.

For each genre, we wanted to determine what types of summaries a user might want to view - these types of summaries would provide the basis for our *goal-focused genre oriented summaries*. Since we are not focusing on natural language generation, we selected the sentence as the summarization unit.

We interviewed (over 10 students) and selected three senior students (with above a 3.5 GPA) from the Carnegie Mellon University English departments for the genre evaluation and summary creation team. During the interview process, students were asked to pick 5 opinion sentences for a movie summary. Any student who did not select five or picked (on purpose) sentences other than opinion sentences were automatically eliminated from consideration for the team.

In the first phase of this project, we met to determine genre appropriate summaries. Three documents were chosen from the seven genres and the students read the documents and then discussed and decided what type of goal focused summaries might be appropriate for this genre,

We then generated summarization guidelines for each type of summary. For example, a generic movie summary must have at least one sentence summarizing the plot. An appropriate number of sentences were also decided for each genre summary. Interviews, which tend to be long, were allowed more sentences in their summaries. The overall results without the details of each summary type are shown in Table 7-1.

Instruction sheets for each genre were created and the students created a ranked list of sentences according to the guidelines as well as a most readable list of sentences – reordering the ranked sentences into a “most readable” order. Such an order took into account coherency of the sentences as a group, intelligibility, flow and summary theme development.

Each type of goal-focused summary had specific guidelines for picking summary sentences:

1. Articles – the main points of the articles.
2. Editorials 1-2 on the thesis of the editorial, 2-3 sentences on the reason for writing the editorial, 1-2 on supporting evidence.
3. Movie Reviews:
    1. Plot: like movie trailer.
    2. Opinion: Reviewers or Others' opinions.
    3. Overview: 1 plot sentence, 1 opinion sentence, 1 cast sentence, 1 movie genre sentence, 1 other supporting sentence.
4. Product Reviews: 1-2 specifics, 1 price range, 1-2 overall opinion, 1 critique.
5. Product Press Releases: Overview of product and purpose
6. Biographies: Professional: 1 training sentence, 1 about job, 1 on career breakthrough or career start, 1-2 achievements. Personal: 1 birth date, 1-2 marital, 1 on children/family, 2-3 sentences on characteristics, unique or interesting information. Overview: 1 birth date, 1 about job, 1 on career, 1-2 sentences on achievements.
7. Interviews: Thematic – Area of expertise and knowledge of subject area. Opinion – Interviewee's character and perspective on life. Articles were collected for author interviews (7 thematic summary sentences, 7 opinion), politician interviews (9 thematic, 5 opinion) and entertainer overviews (5 thematic, 5 opinion). The number of summary sentences was chosen based on the nature of summary and length of document (Table 7-2).

The *interviews* genre was the most difficult genre as well as having the most number of sentences of all the genres. In early phases of the project, before working with the "expert" team with some useful training, we had discussed this genre and decided on two types of summaries: an overview and focused summary. The overview summary would contain three segments spanning the most important topics contained in the interview. Each segment would be composed of the interviewer's question and 3 sentences from the particular answer (resulting in a summary of approximately 12 sentences). The focused summary would have the same composition of three segments, but contain what was considered the most important point(s) of the interviewee and as such might reflect a small section of the interview.

The team of "experts" thought that it was not worthwhile to include the questions in the summary (at least for these three genre-topics of authors, politicians and entertainers). They thought that a good summary sentence should only contain sentence extracts from the interviewee's answer, which should automatically contain the relevant portions of the question. They also thought that the number of sentences needed to vary according to the genre-topic as the occupation of the person being interviewed had a major bearing on what was being discussed. Depending on the class of interviewees, this would determine how many salient sentences could be found, e.g., for entertainers this tends to be low and for politicians this tends to be higher.

Table 7-1: Goal-focused summary defining characteristics as a result of Genre Evaluation Phase

Genre	# of Docs	Type of Goal Focused Summary	# of Summary Sentences	Composition
Movie Reviews	30	Plot	5	Plot sentences
		Opinion	5	Opinion sentences
		Overview	5	Plot, opinion & general sentences
Product Reviews	30	Overview	5	Opnion, comments, specifics of products, price/price range, critique
Product Press Releases	30	Overview	3	Overview of produce and purpose; differentiation from others
Biographies	45	Professional	5	Training, job, breakthrough or start of career, achievements
		Personal	5	Birth information, marital status, children/family, characteristics, unique information, interesting facts
		Overview	5	Birth information , job summary, breakthrough or start of career, achievements
Interviews - Authors	15	Thematic	7	Expertise and knowledge of subject area
		Opinion	7	Interviewee's character and perspective on life
Interviews – Politicians	15	Thematic	9	Expertise and knowledge of subject area
		Opinion	5	Interviewee's character and perspective on life
Interviews - Entertainers	15	Thematic	5	Expertise and knowledge of subject area
		Opinion	5	Interviewee's character and perspective on life
Editorials	30	Overview	5	Thesis (writer's opinion), causal event (situation), supporting details
Articles	30	Overview	5	Main points

Table 7-2: Corpus Description.
# of documents collected, # of sentences in summary and # of summary types and types for that genre. O=Overview, Pl=Plot, Op=Opinion, Pe=Personal, Pr=Professional, T=Thematic.

Genre	# Docs	N_G: # Summary Sentences	Types	Min Sent	Max Sent	Mean Sent	% Human Summary Sentences > Lead
Articles	30	5	O	27	82	47	53%
Editorials	30	5	O	8	37	37	78%
Movie Reviews	30	5	O, Pl, Op	17	163	41	88%
Product Reviews	30	5	O	15	216	60	87%
Product Press Releases	30	3	O	6	116	34	48%
Biographies	45	5	O, Pe, Pr	13	351	89	79%
Interviews	45	5-9	T, Op	37	1040	176	97%

For all three genre-topics of interviews, the students attempted to select 9 sentences, but had difficulty in the author and entertainers category. The students also decided that 9 sentences were definitely sufficient for the politicians category.

The team also thought that there were two types of summaries that were important for the interview genre:

1. **thematic**: Focus on the interviewee's area of expertise: the interview is usually precipitated by some object, or some specific experience, on which the interviewee is an expert. Pick out sentences based on the interviewee's knowledge of their subject area.
2. **opinion**: This type of summary is less focused. The interviewee often expresses opinions on a broad range of topics. Our objective is to capture something about the interviewee's character and perspective on life.

In the summary creation phase, one student did not complete the summaries and so we interviewed and selected a graduate student from the University of Pittsburgh with a background in English to perform the task according to the created instruction sheets for forming summaries.

7.2 Summary System

For each genre, we need different sentence selection methods than that used for newswire event articles, an idea noted by Marcu in 1997 for FAQs [15]. For newswire, since news

is a summary, lead sentences tend to create a good overview summary since they occur in summaries 70% of the time [Goldstein et. al. 1999]. We use lead sentences as a baseline.

The summarization system needs to take the class of a genre (i.e., news event, movie review, etc.) and act accordingly in accordance with the type of goal-focused summary that the user desires. The construction of genre oriented goal-focused summarization systems is a new area of research. We present overall results for the current state of our system and give an example of how this process works for movie reviews.

Based on the criteria by which the humans choose summary sentences (Section 7.1), we designed summary sentence extraction algorithms for the genres and goal-focused summary types in the summarization corpus. Note that an analysis of some of the summary corpus data indicates that humans did not always choose summaries according to these criteria. We empirically determined the weights by experimentation.

The summarization algorithms currently used are listed below and fit into the framework discussed in Section 3.1. Some of these algorithms count the entities, e.g., number of people, locations and organizations in the sentences. In these cases, the entities were identified using Alias-I's Lingpipe (www.alias-i.com/lingpipe).

In order to determine sentences that match a summary sentence, we often use lists. One such list for opinions is Wilson's list of subjectivity and sentiment clues [Wilson et. al. 2005]. Other word lists are hand crafted based on the genre. For example, for movie reviews, movie genre lists were created from genres listed on the web.

For a similarity match score, we computed a score using Lucene [LUCENE]. Each list, including high frequency terms, was treated as a document for the purpose of creating a "*match*" score with a document. Each component score, e.g., the score of the top 15 high frequency terms, is normalized on a scale of 0-1 before being used in computing the final score. The normalization is based on the highest scoring sentence in the document for that component.

Algorithms:

1. Newswire Articles –
   - If the first sentence is over 10 words in length, use it[7].
   - The remaining sentences are the highest scoring sentences, where
     Score-HFT-NPLO = (0.7*Score of Top 15 High Frequency Terms in Document) +(0.3*Number of Persons, Locations and Organizations (PLO)).

   Before the score is calculated the individual score components are normalized to 1 based on the highest scoring sentences in the document.
2. Editorials:
   - If the first sentence is over 10 words in length, use it.

[7] This first sentence score was found to work well in our single document experiments and in our top scoring summarization system in SUMMAC. It eliminates catchy initial sentences used at the beginning of newswire articles to attract the readers interest.

- The remaining sentences are the highest scoring sentences, where sentences are chosen by

  Score- = (0.7*NHLT-NLPO) + (0.3*Number of Matches with Opinion Word List).

3. Movie Plot: Only consider sentences from the first 2/3 of the document. Select sentences using

   Score = (0.3* Number of Capital letters) + (0.3 * Number of PLO) + (0.1 * Plot Keyword List Matches) + (0.3 * Consecutive Sentence Match).

4. Movie Opinion:.
5. Movie Overview:
   - Choose the best sentence with match from cast. Cast summary is the highest scoring single sentence based on:

     Score = (0.2 * 1st sentence) + (0.3 * cast keyword match) + (0.2 * Number of persons) + (0.3 * Number of capital letters in parentheses).

   - Choose the best sentence from with match from genre + director (combined keyword). If the sentence is the same as the cast summary sentence, no genre+director sentence is used.
   - For the remaining sentences, choose sentences using Score-HFT-NPLO.
6. Bio Personal:
   - Choose the best 3 sentences from a match with (a) birth keyword terms, (b) family, keyword terms, and marriage keywords terms. Duplicate sentences are ignored.
   - Choose the next best sentence using

     Score = (0.6 * marriage keyword match) + (0.3 * family keyword match) + (0.1 * number of dates).

   - For the remaining sentences, select the highest ranking sentences from:

     Score = (0.5 * High Frequency word match) + (0.5 * Number of Capital Letters).

7. Bio Professional: Use only sentences in the first 2/3 of the document.
   - Choose the best sentence from the match with awards keywords.
   - Choose the sentence with the highest number of organizations.
   - Choose the sentence that scores highest with the high frequency keywords.
   - For the remaining sentences select the highest ranking sentences from:

     Score = .(0.2 * Number of PLO) + (0.4 * Number of Persons and Orgs) + (0.1 * number of Dates) + (0.2 * High Frequency Keywords) + (0.1 * Award Keyword list match).

8. Bio Overview:
   - Choose the best sentence from the match with (a) birth keywords (b) ,award keywords, and (c) the score-HLT-NPLO.
   - For the remaining sentences,

     Score = 0.2 * Number of persons, organizations, and locations) + (0.05 * Number of persons and organizations) + (0.1 * Number of dates) + (0.6 * High frequency keyword match+title) + (0.05 * Award keyword match)).

9. Interview – Opinion:

   Score = (0.5 * High Frequency Term Match) + (0.2 * Number of PLO) + (0.3 * Opinion Keyword Match).

10. Interview – Thematic

    Use Score-HLT-NPLO for all sentences.

11. Product Press Release:
    - If the first sentence is over 10 words in length, use it[8].
    - Choose any sentence in the first 3 sentences that has a nonzero product press keyword search result.
    - The remaining sentences are chosen from the highest scoring sentences:

      Score = (0.6 * high frequency term match) + (0.4 * (if sentence is in first 5 sentences)).

12. Product Review:
    - Choose the best cost sentence, if available. This is chosen by examining all sentences that contain a number, and of these sentences, selecting the highest ranked sentence from matching with the cost/price keyword list.
    - Choose the highest scoring sentences based on:

      Score = 0.5 * (high frequency term match) + (0.2 * num caps) + (0.3 * opinion keyword match).

Table 7-3 shows the results of evaluating these algorithms using lenient scoring - a match with any human selected sentence is used in the score. Table 7-3 demonstrates that only 50% of our genre specific algorithms are the top scoring algorithms for their particular genre (although some are close to the top scoring algorithms). We need to make further improvement to the summarization algorithms, which we plan to do in the future by considering new extraction algorithms and optimizing weights. We also plan to use the algorithmic components as features and perform sentence selection as a classification task. However, Table 7-3, also shows that many algorithms were able to outperform the lead algorithm (shown in blue). Indeed, ideally a genre oriented summary should be able to do better than the lead sentence algorithm, with the possible exception of product press releases. Product press releases tend to cover the types of sentences desired for summary sentences in the first few sentences.

[8] This first sentence score was found to work well in our single document experiments and in our top scoring summarization system in SUMMAC. It eliminates catchy initial sentences used at the beginning of newswire articles to attract the readers interest.

Table 7-3: Results of various summarization algorithms on Summary Data. Algorithms: Articles (Newswire), Bio-Overview, Bio-Personal, Bio-Professional, Editorials, Interviews-Thematic, Interviews-Opinion, Movie Reviews-Opinion, Movie Reviews-Plot, Movie Reviews -Overview, Product-Press-Release, Product, Review, Leading Sentence, Number of Initial Capital Letters, Number of Persons (using Alias-I Lingpipe), Num of Named Entities (Person, Location, and Organizations using Alias-I Lingpipe), High Frequency Terms from Doc. These values are computing using lenient scoring – a system summary sentence is counted as a match if it matches any summary sentence produced by a human. Maximum score is 1.0.

	ALGORITHMS																
Summaries	NEWS	BIO-O	BIO-PE	BIO-PR	EDIT	INT-OP	INT-TH	MOV-OP	MOV-PL	MOV-OV	PROD-PR	PROD-RE	**Baseline** LEAD	Num Caps	Num Pers	Num NE	HF
Articles	0.45	0.43	0.35	0.48	0.46	0.42	0.41	0.21	0.37	0.39	0.75	0.44	**0.76**	0.24	0.16	0.24	0.44
Bios - Overview	0.34	**0.40**	0.32	0.36	0.35	0.30	0.31	0.24	0.35	0.31	**0.40**	**0.40**	**0.40**	0.26	0.21	0.28	0.32
Bios - Personal	0.25	**0.31**	**0.39**	0.31	0.25	0.21	0.21	0.18	0.26	0.20	0.33	0.27	**0.34**	0.22	0.23	0.21	0.18
Bios - Professional	0.30	0.30	0.23	**0.40**	0.32	0.28	0.29	0.23	0.34	0.31	0.36	0.32	**0.37**	0.26	0.21	0.25	0.32
Editorials	0.43	0.47	0.41	**0.48**	0.44	0.41	0.45	0.32	0.44	0.41	0.47	0.40	**0.47**	0.38	0.38	0.37	0.44
Interviews - Opinion	0.07	0.10	0.11	0.06	0.08	0.08	0.09	**0.12**	0.05	0.10	0.05	0.09	**0.05**	0.09	0.09	0.06	0.09
Interviews - - Thematic	0.14	0.14	0.13	0.14	0.13	0.13	0.15	0.06	0.12	0.14	0.12	**0.16**	**0.12**	**0.16**	0.12	0.09	0.13
Movie Reviews - Opinion	0.30	0.37	0.37	0.21	0.34	0.35	0.31	**0.40**	0.21	0.32	0.29	0.37	**0.29**	0.25	0.26	0.25	0.33
Movie Reviews - Plot	0.41	0.36	0.36	0.43	0.36	0.41	0.49	0.28	**0.51**	0.44	0.25	0.38	**0.23**	0.39	0.36	0.41	0.43
Movie Reviews - Overview	0.49	0.49	0.43	0.37	0.48	0.51	**0.53**	0.28	0.42	**0.51**	0.37	0.49	**0.36**	0.39	0.37	0.39	0.50
Product Press Release	0.36	0.38	0.50	0.56	0.34	0.46	0.43	0.32	0.40	0.29	**0.79**	0.51	**0.73**	0.31	0.22	0.27	0.37
Product Reviews	0.32	0.31	0.27	0.31	0.34	0.36	0.32	0.29	0.34	0.30	0.31	**0.38**	**0.31**	0.25	0.17	0.22	0.34

Table 7-3 shows that many genre specific algorithms are performing better than the lead baseline summaries, indicating the utility of genre oriented summarization. For our purposes, we also treat the newswire algorithm as a baseline as there has been much research in single document summarization to produce summaries that outperform the lead summaries [Mani et. al. 1998].

- Biographies: Some biography summarization algorithms (SA) outperform the lead sentence SA – the *Professional* SA 0.40 to 0.37 and the *Personal* SA 0.39 to 0.34. Both biography SAs outperform the newswire baseline SA as well (Overview 0.44 to 0.34 and Personal 0.39 to 0.24).
- Product Press Releases: The SA outperforms the lead SA 0.79 to 0.73 and greatly outperforms the newswire baseline SA, 0.79 to 0.36.
- Product Reviews: The SA outperforms the lead SA 0.38 to 0.31 and the newswire SA, 0.38 to 0.32.
- Movie Review Categories: There is a big difference in the as compared to the lead SA – for the *Opinion* SA 0.40 compared to 0.29, the *Plot* SA, 0.51 to 0.23 and the *Overview* SA, 0.71 as compared to 0.36. There was a less of a difference for as compared to the newswire baseline SA, but still a fair difference for the first Opinion and Plot, but little difference for Overview. The *Opinion* SA score 0.40 as compared to 0.30, the *Plot* SA scores 0.51 as compared to 0.41 and the *Overview* SA scores 0.71 as compared to 0.49.

For the other three genres, the summaries from the newswire and editorial genre did not outperform the lead summaries. The summarization algorithms performed the worst in the interview genre – all scores are less than 0.20 (Table 7-4), although in most cases, the scores are higher than the lead sentence SA. The low performances of the interview SAs indicate that the interview algorithms need to be improved. We suggest that the low performance is due to the diversity and length of the genre. Since a particular topic in any interview was not selected for focused summarization, the human judges (and the summarizers) could choose a variety of topics to be included in the summary. In this case, the summaries may need to be evaluated in a different manner, such as by quality or in task based scenario. A detailed analysis of this genre may indicate additional methods for how to extract good summary sentences.

Even if we cannot demonstrate in all cases from Table 7-3 that the genre oriented summarization algorithm is the best scoring algorithm, genre oriented goal-focused summaries are still important due to the fact that from the genre tag, they include genre related information for that genre, such as the audience ratings, overall ratings and running time for the movie genre and the product cost for the store product page genre.

Table 7-4: Sentence match (overlap) results for subcategories of interview data.

	ALGORITHMS																
Summaries	NEWS	BIO-O	BIO-PE	BIO-PR	EDIT	INT-OP	INT-TH	MOV-OP	MOV-PL	MOV-OV	PROD-PR	PROD-RE	**Baseline** LEAD	Num Caps	Num Pers	Num NE	HF
Interviews - Author - Opinion	0.11	0.15	0.15	0.07	0.11	0.12	**0.17**	0.15	0.04	0.13	0.07	0.13	**0.07**	0.09	0.13	0.09	0.11
Interviews - Author - Thematic	0.16	0.11	0.17	0.15	**0.19**	**0.19**	**0.19**	0.09	0.13	0.16	**0.19**	**0.19**	**0.19**	**0.19**	0.11	0.13	**0.19**
Interviews - Entertainers - Op	0.07	**0.15**	0.05	0.05	0.08	0.09	0.07	0.12	0.04	0.08	0.04	0.08	**0.04**	0.09	0.11	0.05	0.08
Interviews - Entertainers - Th	0.19	0.15	0.11	0.17	0.12	0.12	0.17	0.08	0.13	0.17	0.08	0.15	**0.08**	**0.20**	0.16	0.16	0.13
Interviews - Politicians - Op	0.04	0.08	**0.12**	0.07	0.04	0.04	0.04	0.09	0.08	0.08	0.05	0.05	**0.05**	0.07	0.04	0.04	0.07
Interviews - Politicians - Th	0.07	**0.15**	0.11	0.09	0.07	0.08	0.09	0.01	0.09	0.09	0.09	**0.15**	**0.09**	0.08	0.09	0.09	0.08

Table 7-5 shows the recall results of running ROUGE-BE on the same summaries as in Table 7-3. Note that the performance of some algorithms is quite different, for example the news article algorithm or the news article data as well as the product review algorithm on the biography overview genre data. In the sentence match scoring algorithm, system summary sentences that don't match a human sentence do not contribute to the score. With ROUGE scoring, all sentences have the potential to contribute. Tuning the summary algorithms using to optimize ROUGE scores would result in different weighting parameters.

Again, as in Table 7-3, the interview genre had the lowest scores – which we believe is due to the length of these document and the diversity of items discussed which makes it difficult to create agreement in gold standard and system summaries that target the same focal areas.

Table 7-6 shows the significance results of the ROUGE-BE recall scores using the ROUGE generated confidence scores. 5 of the genre oriented algorithms (including bios and movie reviews) performed better than the two baselines (news articles and leading sentences) at a 85% confidence interval and 2 above 95% (movie reviews plot and movie reviews overview).

Table 7-5: ROUGE-BE Recall results for summarization algorithms on Summary Data.
Algorithms: Articles (Newswire), Bio-Overview, Bio-Personal, Bio-Professional, Editorials, Interviews-Thematic, Interviews-Opinion, Movie Reviews-Opinion, Movie Reviews-Plot, Movie Reviews -Overview, Product-Press-Release, Product, Review, Leading Sentence, Number of Initial Capital Letters, Number of Persons (using Alias-I Lingpipe), Num of Named Entities (Person, Location, and Organizations using Alias-I Lingpipe), High Frequency Terms from Doc.
Generated using "ROUGE-1.5.5.pl –3 HM –SIMPLE

	ALGORITHMS																
Summaries	NEWS	BIO-O	BIO-PE	BIO-PR	EDIT	INT-OP	INT-TH	MOV-OP	MOV-PL	MOV-OV	PROD-PR	PROD-REV	**Baseline** LEAD	Num Caps	Num Pers	Num NE	HF
News Articles	**0.57**	0.37	0.30	0.40	0.36	0.34	0.30	0.11	0.26	0.29	0.56	0.36	**0.57**	0.21	0.12	0.17	0.33
Bios - Overview	0.22	0.22	0.20	0.24	0.25	0.22	0.20	0.16	0.23	0.21	0.22	**0.29**	**0.21**	0.19	0.14	0.17	0.22
Bios - Personal	0.19	0.18	**0.28**	0.21	0.14	0.12	0.16	0.09	0.18	0.11	0.19	0.16	**0.19**	0.17	0.17	0.15	0.10
Bios - Professional	0.19	0.16	0.14	**0.25**	0.21	0.20	0.18	0.14	0.24	0.19	0.19	0.24	**0.18**	0.26	0.13	0.17	0.20
Editorials	0.28	0.29	0.24	**0.29**	0.26	0.25	0.28	0.18	0.29	0.27	0.28	0.24	**0.28**	0.22	0.22	0.24	0.25
Interviews - Opinion	0.02	0.05	**0.06**	0.04	0.04	0.04	0.05	0.05	0.03	0.05	0.02	0.05	**0.02**	0.05	0.05	0.04	0.05
Interviews - Thematic	0.05	0.07	0.07	0.07	0.06	0.06	0.08	0.04	0.07	0.07	0.05	0.08	0.05	**0.09**	0.06	0.07	0.06
Movie Rev. - Opinion	0.16	0.22	**0.23**	0.11	0.20	0.22	0.18	**0.23**	0.13	0.21	0.16	**0.23**	**0.15**	0.17	0.15	0.15	0.19
Movie Reviews - Plot	0.16	0.27	0.27	0.35	0.30	0.33	0.38	0.12	**0.41**	0.37	0.19	0.33	**0.15**	0.36	0.36	0.35	0.33
Mov Rev. - Overview	0.21	0.35	0.33	0.27	0.33	0.35	0.34	0.14	0.30	**0.37**	0.23	0.35	**0.19**	0.31	0.28	0.29	0.35
Product Press Release	0.69	0.38	0.41	0.47	0.49	0.49	0.36	0.34	0.34	0.40	0.69	0.50	**0.70**	0.33	0.18	0.28	0.48
Product Reviews	0.17	0.18	0.13	0.17	0.21	**0.23**	0.17	0.16	0.21	0.19	0.17	**0.23**	**0.17**	0.16	0.11	0.14	**0.23**

Table 7-6: Significance of ROUGE-BE recall results for genre focused summarization algorithms.
Algorithms: Articles (Newswire), Bio-Overview, Bio-Personal, Bio-Professional, Editorials, Interviews-Thematic, Interviews-Opinion, Movie Reviews-Opinion, Movie Reviews-Plot, Movie Reviews -Overview, Product-Press-Release, Product, Review, Leading Sentence, Number of Initial Capital Letters, Number of Persons (using Alias-I Lingpipe), Num of Named Entities (Person, Location, and Organizations using Alias-I Lingpipe), High Frequency Terms from Doc.
Generated using "ROUGE-1.5.5.pl –3 HM –SIMPLE

Genre	Genre Summarizer Score	Baseline News Article	Baseline Leading Sentence	Significance using ROUGE Scoring
News Articles	0.57	-	0.57	No significant difference.
Bios - Overview	0.22	0.22	0.21	No significant difference. Product Review algorithm score is 0.29, differs from NewsArticle & Lead at 85% Confidence Interval
.Bios - Personal	0.28	0.19	0.19	Differs from NewsArticle & Lead at 85% Confidence Interval
Bios - Professional	0.25	0.19	0.18	Differs from NewsArticle at 83% Confidence Interval & Lead at 85% Confidence Interval
Editorials	0.26	0.28	0.28	No significant difference.
Interviews - Opinion	0.04	0.02	0.02	No significant difference.
Interviews - Thematic	0.08	0.05	0.05	No significant difference.
Movie Rev. - Opinion	0.23	0.16	0.15	Differs from NewsArticle & Lead at 85% Confidence Interval
Movie Reviews - Plot	0.41	0.16	0.15	Differs from NewsArticle & Lead at 99% Confidence Interval
Movie Review - Overview	0.37	0.21	0.19	Differs from NewsArticle at 95% Confidence Interval & Lead at 97% Confidence Interval.
Product Press Release	0.69	0.69	0.70	No significant difference.
Product Reviews	0.23	0.17	0.17	No significant difference.

Of the seven basic genres (newswire, editorials, product press releases, interviews, product reviews, biographies and movie reviews), two genres show the benefits of genre oriented summarization. Both product reviews and interviews – would benefit from the additional information that can be extracted for this genre, such as product name and cost for product reviews and interviewer and interviewee for interviews. If the scoring took

into account this information, we hypothesize that these genres would also show significant differences with the two baselines. We expected that it would be difficult for the other three genres, newsarticles, editorials (often very similar to news) and product press release algorithms to outperform the baseline of lead sentence summaries, since the initial sentences of a document are known to provide excellent summaries in the case of news [Goldstein et. al. 1999] and product press releases are written in a style where the key information is presented in the initial sentences.

7.3 Movie Reviews

To further motivate the use of goal focused summaries, let us examine the movie-reviews genre (MRG) in detail. In this genre, users may want to have a overview summary (Figure 1-11) or goal-focused summaries such as plot (Figure 1-12) or the reviewers opinion (Figure 1-13). The unique sentences for a particular summarization algorithm (SA) as compared to another summarization algorithm are shown in green. There is one unique sentence for the overview algorithm, one for plot and three for the opinion algorithm.

- overview summary using the MRG overview SA:
    - one sentence (3) overlaps with the summary created using the lead sentence SA (Figure 1-9) and MRG plot SA
    - three (14, 17, 24) overlap with a movie review summary created using the newswire genre SA (Figure 1-10). Sentence 24 also overlaps with the opinion summary using the MRG opinion SA. We expect at least one sentence to overlap with the opinion summary using the MRG opinion SA, since the overview summary requires an opinion sentence (Table 7-1).
- plot summary using MRG SA: three sentence (3, 4, 5) overlap with the lead sentence SA.
- Opinion summary use MRG SA: one sentence (24) overlaps with the overview summary using the MRG overview SA.

In the individual algorithms, the MRG opinion SA will give a higher weight to sentences containing sentiment words since the goal is to extract the reviewer's opinion of the movie. For the MRG plot SA, sentences that have a high count of named entities are given a high score since named entities (e.g., people's names and place names) frequently appear in plot sentences. Consecutive sentences to a highly ranked chosen sentence are also given a boost to their score.

The results of comparing five sentence summaries for the movie-reviews to the baseline lead sentence summary and newswire genre summary is shown in Table 7-7. The baseline summary created just using high frequency terms in the document is also shown. This table shows that there is a significant difference in sentence selection, judged by summary score and sentence overlap, for different genre and goal-focused summary mechanisms.

Table 7-7: Percent of Overlapping Sentences among System Produced Summaries and Human

Baseline algorithms include: High Freq Terms - Summary sentences chosen by High Frequency Terms, News – Newswire Generic Summarization Algorithm, and Lead – First five sentences of the document

System Summary Sentence Overlap for Movie Summaries						
System Summarizer	**High Freq. Terms**	**News**	**Lead**	**Overview**	**Opinion**	**Plot**
High Frequency Terms	**100**	70	21	65	9	47
News article	70	**100**	31	76	6	53
Leading Sentence	21	31	**100**	26	21	23
Movie Overview	65	76	26	**100**	9	57
Movie Opinion 2	9	6	21	9	**100**	11
Movie Plot 3	47	53	23	57	11	**100**

The agreement between human summarizers in shown in Table 7-8. Note that there is little agreement between humans for plot vs. opinion, indicating the need for goal-focused summarization. There is more overlap between both plot and overview summaries as well as opinion and overview summaries – both plot sentences and opinion sentences are part of the composition of the movie overview summary. The fact that two summaries for a given genre are very different, such as plot summaries compared to opinion summaries, and have little overlap in the human gold standards indicates the importance and need for goal-focused summarization.

Table 7-8: Percent sentence agreement between human summarizers (labeled 1, 2, and 3) for the movie review goal-focused summaries.

Human Summaries Sentence Overlap for Movie Summaries									
	Movie Plot			Movie Overview			Movie Opinion		
Human Summary	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
Plot 1	**100**	43	46	35	23	31	13	7	9
Plot 2	43	**100**	47	39	47	9	25	19	19
Plot 3	46	47	**100**	25	22	30	9	6	2
Overview 1	35	39	25	**100**	45	39	47	31	31
Overview 2	23	47	22	45	**100**	41	43	56	39
Overview 3	31	39	30	39	41	**100**	41	33	44
Opinion 1	13	25	9	47	43	41	**100**	47	49
Opinion 2	7	19	6	31	56	33	47	**100**	49
Opinion 3	9	19	2	31	39	44	49	49	**100**

As previously mentioned, it is important to realize that the genre tag contributes to a successful goal focused summary in two ways:

- Due to the goal-focused summarization algorithms, genre misclassification can result in a decrease in the overall summary score or value (Table 7-3).
- Incorrect genre identification or lack of any genre identification can result in the omission of useful information that the summarizer has extracted from the document based on the genre tag. Such information may be important to the user's information seeking goals. One example is the display of the movie rating and running time, shown in red in Figure 1-11 for the movie review genre overview summarization algorithm. This information is not extracted and displayed in Figure 1-10, which shows a movie review summary created from the news article summarization algorithm.

Thus if the genre classifier can pass the accurate tag for movie reviews, $F_1 = 0.86$ (Table 6-6) and for all 9 genres $F_1 = 0.87$ (Table 6-14), the summarization system can present a summary that more effectively addresses a user's information seeking goals.

7.4 Scientific Articles

As another example of a genre that could show the benefit of genre oriented summarization, we decided to examine the scientific article genre. The scientific articles are part of the Computation and Language Corpus and were annotated in various ways, including sentences in the abstract and any corresponding sentences in the article [Teufel and Moens 1997, Teufel and Moens 2002].

We examined this data set for abstracts of 5 sentences in order to correlate with the majority of sentence summarize sizes in our genre oriented data set described previously. 14 such articles were found. Of those, a few had abstracts that contained sentences without a corresponding sentence in the article. In these cases, we selected the best sentence from the article. This resulted in one gold standard summary per article –a human produced sentence extract summary corresponding to the abstract as close as possible.

The algorithm use two components, each component's range of values are normalized to a scale of 0-1 before score is assigned to the sentence. The score is based on 0.6 * Score HFT-NPLO + 0.4 * (Sentence is a member of initial 10 sentences or last 10 sentences) where Score-HFT-NPLO = (0.7*Score of Top 15 High Frequency Terms in Document) +(0.3*Number of Persons, Locations and Organizations (PLO)).

The results for the exact sentence match and running ROUGE-BE are shown in Table 7-9 and Table 7-10 respectively. This Scientific Article Algorithm outperforms the baselines NewsArticle and Lead at the confidence interval as computed by ROUGE.

Table 7-9: Sentence match results for summarization algorithms for Scientific Article Data.
Maximum score is 1.0.

	ALGORITHMS																
Summaries	NEWS	BIO-O	BIO-PE	BIO-PR	EDIT	INT-OP	INT-TH	MOV-OP	MOV-PL	MOV-OV	PROD-PR	PROD-REV	SCIART	**Baseline** LEAD	Num Caps	Num NE	HF
ScientificArticles	0.13	0.13	0.10	0.09	0.11	0.06	0.06	0.07	0.10	0.06	0.16	0.13	**0.24**	0.13	0.09	0.04	0.09

Table 7-10: ROUGE-BE recall results for summarization algorithms on Summary Data.

	ALGORITHMS																
Summaries	NEWS	BIO-O	BIO-PE	BIO-PR	EDIT	INT-OP	INT-TH	MOV-OP	MOV-PL	MOV-OV	PROD-PR	PROD-REV	SCIART	**Baseline** LEAD	Num Caps	Num NE	HF
ScientificArticles	0.13	0.20	0.13	0.14	0.16	0.11	0.11	0.13	0.16	0.11	0.17	0.19	**0.31**	0.13	0.09	0.06	0.14

7.5 Discussion

We have discussed the importance of genre identification and the use of genre specific information for creating *genre-oriented goal-focused* summaries that are more beneficial for a user's information seeking goals. We have shown that for many of our genres, we can create genre oriented goal focused summaries which obtain better scores than the baseline summarization algorithms in an evaluation using our gold standard corpus. We have also demonstrated that summarizing based on the inappropriate genre can provide sub-optimal summaries both through their evaluation performance and from the lack of extracting the pertinent information related to the genre for the user.

The genre tag therefore allows us to utilize such goal-focused algorithms to generate goal-focused summaries for a genre. The user can choose the desired goal focused summary, which includes the genre specific items in the summary and/or summary header.

In the next chapter, we will discuss the genre of email and how we can create summaries that are beneficial in this domain.

Chapter 8 Email Summarization

"A great productivity enhancer? Ha! E-mail can be a tremendous waste of time unless you know how to tame the savage beast."
Stever Robbins

In this chapter, we examine the summarization of email. Emails are a communicative genre reflecting the purpose of the writer. Often this is to forward information, request information or set up meetings. In this chapter we examine such intents.

Emails are often very short, reflecting an item in a conversational thread or a response to a request for information. As such, the summary of an email could possibility consist of the overall purpose of the email as well as a short summary, perhaps either the subject line or a "sentence" from the text body of the last sender. We explore the suitability of such approaches to summarization in this chapter. One question that we are interested in determining is, "how often the first line of an email makes a good summary sentence?"

We define sub-genres of email as well as a subset of "email speech acts" relevant to email enhanced for email specific discourse. After creating a ground truth set of emails based on these email acts, we compare the performance of two classifiers (Random Forests and SVM-light) in identifying the primary communicative intent of the email and its corresponding sub-genres. We experiment with using feature sets derived from two verb lexicons as well as a feature set containing selected characteristics of email.

We end by examining a subset of the Enron Email corpus to determine the suitability of an enriched email summary consisting not only of the subject of the email (as entered by the author), but the subject of the email and one sentence from the most recent email sent to the author.

8.1 The Email Genre

Today the Internet is used on a daily basis by over 70% of adult Americans and almost 60% use email on a typical day [Pew 2005]. These figures represent a dramatic increase over usage levels of only five years ago. Recently, several events have focused public attention on email. On both the national and local level, email communication is being combed for information that may be used for legal or other purposes [Olesker 2005]. The most prominent of these events, the collapse of the Enron Corporation, resulted in the courts making available in 2004 to public access a corpus of over 500,000 corporate Enron emails. The presence of this corpus and the need to develop tools for email processing has stimulated interest in research into email.

Understanding the structure and functions of email will aid in the development of much needed tools for the categorization and summarization of email. There has been a great deal of work in email filtering and spam detection [Schneider 2003]. The more

demanding task of categorizing email is receiving more attention [Klimint and Yang 2004]. And with the availability of large corpora, there is growing interest in using email analysis to determine the structure of social networks [Diesner and Carley 2005].

Historically, email structure and function has been of interest to a variety of research communities. Among them are linguists, social scientists studying communication and organizational behavior, and those interested in genre studies. With the availability of the Enron and other large email corpora, utilizing the analytical techniques from both the computational linguistics and computer science communities is more feasible. A long-standing question involves the basic structure of email: is email a form of writing, a form or speech, or a new, hybrid genre [Baron 2000, Taylor 1992]? It is appealing to characterize email as a genre. Writing and, less-frequently, speech can be characterized by its genre:

> [A genre is] a patterning of communication created by a combination of the individual, social and technical forces implicit in a recurring communicative situation. A genre structures communication by creating shared expectations about the form and content of the interaction, thus easing the burden of production and interpretation. [Erikson 2000]

If, in fact, email is a distinctive genre that has emerged as the result of a new communicative medium [Shepherd and Watters 1999], researchers can be guided by an expected form and content. Myka argues that email is an amalgam of several genres [Myka 1997], a theory we support, and Crystal observes that the email genre is still evolving [Crystal 2001].

Email can be considered to be an amalgam of speech and writing. Biber used statistical techniques to analyze the linguistic features of twenty-three spoken and written genres [Biber 1998] and found that the relationship among these genres is complex and that there is no simple dichotomy between speech and writing. Collot and Belmore extended Biber's work to examine electronic messages posted to an electronic bulletin board [Collot and Belmore 1996]. They found that these messages most closely resembled interviews and letters, and, in the dimension measuring the level of interaction and personal affect (Biber's Dimension 1), they more closely resembled the spoken genres rather than the written ones. Baron cites linguistic features of much email: informality of style, psychological assumption that the medium is ephemeral, and a high level of candor, that are speech-like [Baron 2000].

Following some recent work of Cohen, we believe that email most closely resembles speech and look to analyze email in terms of speech acts [Cohen et. al. 2004]. Our research differs from Cohen's in that we focus on verbs and the combination of verbs with email specific features. The philosopher John Austin posited that a speaker in using words performs an act in making an utterance [Austin 1962]. He suggested that this act can be categorized as one of five "illocutionary" acts and identified a class of verbs with each. His approach has been amplified by others, [Vendler 1972, Bach and Harnish 1979] (Table 8-1).

Table 8-1: Comparison of Traditional Speech Act Categories by Author

SA Category	Austin	Vendler	Bach & Harnish	Description
Assertive	Expositive	Expositive	Constative	expound views, state, contend, insist, deny, remind, guess
Commissive	Commissive	Commissive	Commissive	commit the speaker: promise, guarantee, refuse, decline
Behabitive	Behabitive	Behabitive	Interpersonal	reaction to others: thank, congratulate, criticize
Interrogative		Interrogative	Directive/Query	ask, question
Exercitive	Exercitive	Exercitive	Directive/Request	exercise power, rights or influences: order, request, beg,
Verdictive	Verdictive	Verdictive & Operative		giving a verdict: rank, grade, define, call, analyze

An email might be considered to be a sequence of one or more utterances (a "soliloquy" of sorts) and thus a sequence of speech acts. Characterizing an email by its most important speech act and its genre could provide a way of categorizing email in terms of the intended action of the sender and expected action on the part of the recipient. Such information could be utilized to triage large volumes of incoming email to produce "to-do" lists for the recipient [Corston-Oliver et. al. 2004] or track responses to the user's requests for information or action.

8.2 Genres (Speech Acts) of Email

Computational linguists have used annotation schemes that label speech acts at the utterance level. Dialogue Act Markup in Several Layers (DAMSL) [Allen and Core 1997] and Dialogue Act Modeling [Stolcke et. al. 2000] are two such schemes. A set of detailed annotation guidelines based on DAMSL expanded the number of possible categories [Jurafsky et. al. 1997]. Table 8-2 indicates the relationship between the categories defined in these methodologies and traditional speech acts.

We extended these schema for emails, to allow for annotation of the primary communication intents of email, which we refer to as the "email act" (column 1 in Table 8-2). Table 8-2 shows the comparison of our email acts to categories of Dialogue Acts, DAMSL and SWBD-DAMSL and the corresponding traditional speech acts. In dialog analysis acts, there are two primary utterance characterizations: "*forward-communicative*" – describing the effect on the subsequent dialogue and interaction, and "backwards-communicative" – describing how the current utterance relates to previous discourse.

Email C	Category	Dialogue Act (DA) Modeling	DAMSL	Switchboard SWBD-DAMSL Annotation	Speech Acts Austin	Vendler	Bach&Harnish
	Forward-Communicative Function						
	Statements		Statement-Assert		Expositive	Expositive	Constative
			Statement-Reassert		"	"	"
			Statement-Other		"	"	"
A1		***Statement***		Statement-non-opinion	"	"	"
A2		***Opinion***		Statement-opinion	"	"	"
	Influencing-addressee-fut.-action						
I1 & I2	*For information*		***Info-request***		Exercitive	Interrog.	Directive-Query
I1 & I2		***Yes/No Question***		Yes/No Question	"	"	"
		Wh-question		Wh-question	"	"	"
		Declarative Wh-Ques			"	"	"
		Open-Question		Open-Question	"	"	"
				Or-Question	"	"	"
		Or-clause		Or-clause	"	"	"
		Declarative Yes-No-Ques		Declarative-Question	"	"	"
		Tag-Question		Tag-Question	"	"	"
		Rhetorical questions		Rhetorical questions			
	For action	Action-Directive			"	Exercitive	Directive-Request
D2			***Open-option***	Open-option	"	"	"
D1			***Action-directive***	Action-directive	"	"	"
	Committing-speaker-fut.-action						
C1			***Offer***	Offer	Commissive	Commissive	Commissive
C2			***Commit***	Commit	"	"	"
		Offers, Options, & Commits			"	"	"
	Other	Conv-opening	Conv-opening	Conv-opening	"	"	"
		Conv-closing	Conv-closing	Conv-closing	"		
V1			***Explicit-performance***		Verdictive	Verdictive	
B2		***Thanking***		Thanking/You're welcome	Behabitive	Behabitive	Interpersonal
B1		***Apology***		Apology	"	"	"
B3			***Exclamation***	Exclamation	"	"	"
O1		Other	***Other-forward-funct***	Other-forward-function			
	Backwards-Communicative Function						
R3, F3	***Agreement***				Assertive	Expositive	Constative
		Agreement/Accept	Accept	Accept	"	"	"
			Accept-part	Accept-part	"	"	"
		Maybe/Accept-Part	Maybe	Maybe	"	"	"
			Reject-part	Reject-part	"	"	"
		Reject	Reject	Reject	"	"	"
		Hold before answering/a	Hold before answerin	Hold before answering/agre	"	"	"
		Dispreffered answer			"	"	"
	Understanding						
R3, F3, R5, F5		Signal-non-understandin	***Signal-non-underst***	Signal-non-understanding	"	"	"
R3, F3, R5, F5			***Signal-understanding***		"	"	"
		Backchannel/Ack., Backchannel/Question		Acknowledge			
R5 & F5		***Response Ack.***		Ack-answer			
			Repeat-rephrase	Repeat-rephrase			
		Collaborative Complet.	Completion	Completion			
R3 & F3		***Summarize/reformulate***		Summarize/reformulate	Assertive	Expositive	Constative
		Appreciation		Appreciation	Behabitive	Behabitive	Interpersonal
				Sympathy	"	"	"
		Downplayer		Downplayer	"	"	"
			Correct-misspeaking	Correct-misspeaking	Assertive	Expositive	Constative
		Repeat phrase					
	Answer						
R1 & F1		Yes Answers	***Answer***	Yes-Answer	Assertive	Expositive	Constative
R1 & F1		No Answers	***"***	No-Answer	"	"	"
		Affirm non-yes answers		Affirm non-yes answers	"	"	"
		Neg non-no answers		Neg non-no answers	"	"	"
R2, F2, R4, F4		***Other answers***		Other answers	"	"	"
				No plus expansion	"	"	"
R5 & F5	***(Other backwards)***			Yes plus expansion	"	"	"
				Statement expanding y/n a	"	"	"
				Expansions of y/n ans	"	"	"
				Dispreferred ans	"	"	"

Table 8-2: Email Classes as Related to Shallow Discourse Annotation & Speech Acts

In emails, responses often contain a mixture of speech acts, e.g., answers and comments to a sender's email as well as additional questions for the sender. We found it necessary to expand the backward communicative function of speech acts to allow for this additional forward-communicative function. This resulted in the forward-backward communicative category (response with expectation of reply). We found when annotating emails for the overall intent, it is often very difficult to ascertain the primary intent of the email in the cases when there are both answers and questions, and adding this category improved inter-annotator reliability. If an email were allowed multiple categories or annotated section by section (or possibly paragraph by paragraph), there might not have been this need.

We also added three email specific categories: T (Transmissives) in which the sender's intent is to forward information to a recipient, S (Self) - sending email to oneself (reminders), and N (Non-personal) - emails such as newsletters or items from list servers.

We suggest that email, unique in its form, be categorized by several new genres (Table 8-3). Some of these genres correspond to recognized written & spoken genres [Biber 1988], e.g., email conversations to telephone and face-to-face conversations. Others are novel and peculiar to email, such as spam, which includes advertising and phishing. We believe that email conversations are a distinct genre and have two specific forms – one includes explicit threads and the other does not. Email conversations have content-based subgenres that can best be characterized by the primary intent of the email, i.e., the primary email "speech" act. Email conversations can be formal or informal, which might necessitate additional genres.

Table 8-3: The 12 main Email Acts and their corresponding genres.

Category	Example	Suggested genres
(S) Self	Emails to self	Email reminders/notes
(N) Non-Personal	Bulk Emails	Spam (Advertising, Phishing, etc.) ENewsletters
(T) Transmissives	Forwarding documents	Digital cover letter/memo with attachments
(R) Responses	Provide info to question.	Email Conversations
(F) Response w/ forward function	Provide info to question and ask questions	Email Conversations
(I) Info request	Asks for information	Email Conversations
(D) Directive	Ask someone to do something	Email Conversations
(C) Commits	Commit/offer to do something	Email Conversations
(A) Assertions	Make statements/state opinions	Email Conversations
(B) Behabitive	Express feelings	Email Conversations
(V) Verdictive	Statement accomplishments act, e.g., paper notifications	Official Digital Documents, Digital Letters
(O) Other	Hellos, introductions	Email Conversations

The flowchart of Figure 8-1 describes how to select the primary discourse intent of the email. Although this intent does not distinguish all the genres as listed in Table 8-3, we believe that additional features would allow us to distinguish between cases in which there are multiple suggested genres. Table 8-4 provides details of the annotation methodology, including precedence among the 30 subcategories of email acts, 23 speech email acts (S, N and T are not speech acts). The annotation was designed in a decision tree format in an attempt to maximize agreement among annotators.

One item to note about this flowchart is that there is no category specifically for meetings. Depending on the intent of the sender, such a message could be part of several categories:

- A1 – informing someone about a meeting with no expectation of reply
- D1 – meeting invitation with expectation that the person will attend, such as the required attendee in Microsoft Outlook's meeting invitation
- D2 – meeting invitation with recipient option, such as the optional attendee in Microsoft Outlook. Reply is still expected (or desired) on the part of the sender.
- C1 – accepting a meeting invitation

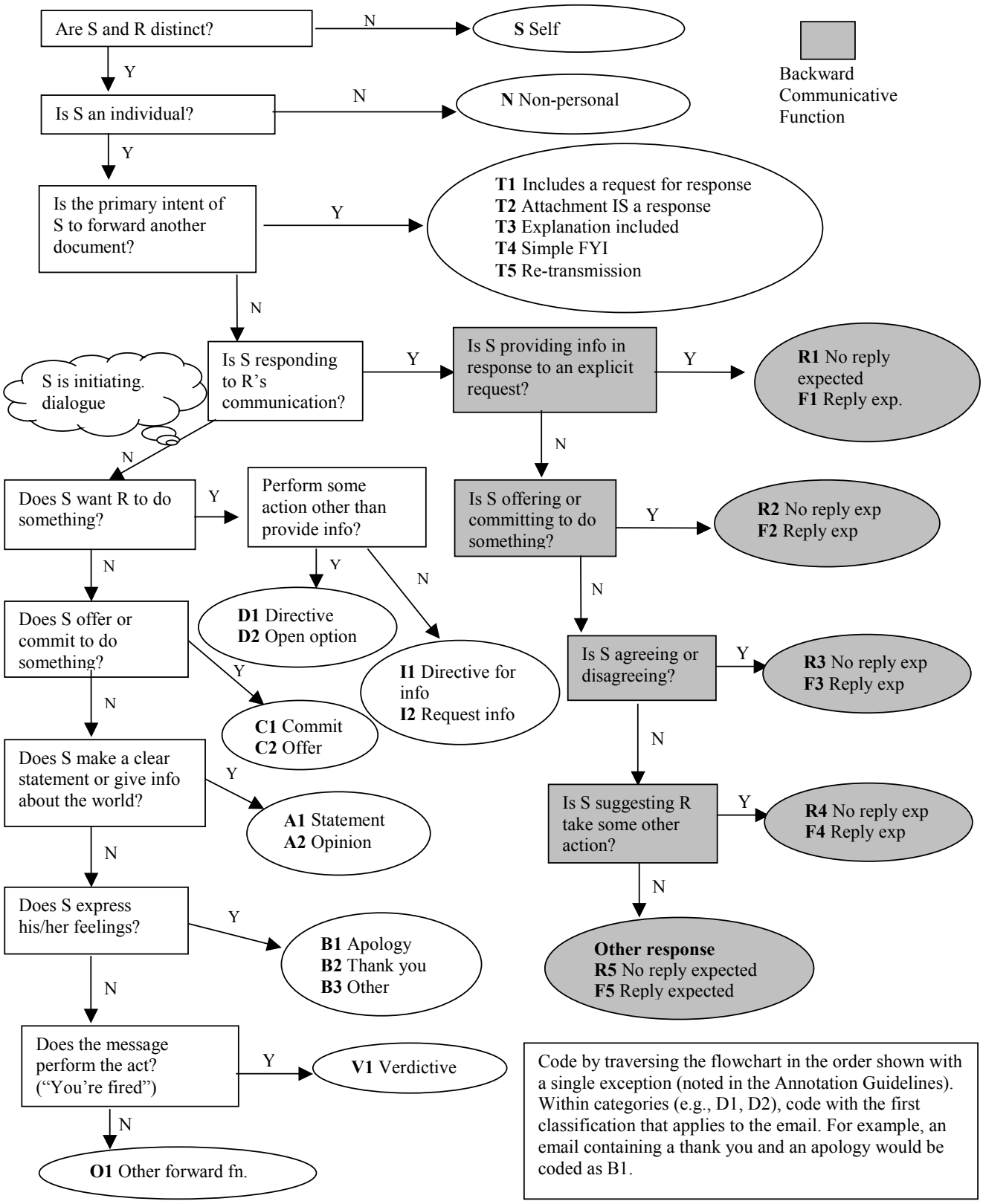


Figure 8-1: Email Speech Acts for Email. 12 Main Categories, 30 Subcategories consisting of 23 traditional speech acts and 7 email specific acts.
The gray color shading indicates a backward communicative function.

Table 8-4: Annotation Guidelines

Flowchart Question	Clarification on answering "Yes"	Annotate as	
Is this email from S to S?	A tickler or reminder with some information content. Attachments forwarded to oneself should be coded as T4.	S	
Is S an individual?	If S is a business, listserv, or some other institution.	N	
Is the primary intent of S to forward another document?	S may add some minimal explanation but not enough to warrant a separate email. The value is in the attachment(s).		
	S is sending info and asks for a response.	T1	
	S is sending info as a response to a request from R.	T2	
	S includes a short explanation (not just a list of topics).	T3	
	S is sending information "FYI".	T4	
	S is re-transmitting after a failed attempt ("Here it is again.") or with correction ("Here's the corrected version.").	T5	
To enable social network analysis, give precedence in annotating to backwards functions, i.e., if an Rn or Fn is encountered, code the entire email as Rn or Fn.			
Is S responding to R's communication?	R may have used email or another method originally. The presence of Re: in the subject *may* indicate a response. The request from R must be explicitly evident. "As you asked, …"		
In these "backward function" annotations, we differentiate between those cases where a response is expected, i.e. includes a forward function (Fn) and where none is expected (Rn)—a dead-end response. The forward function should be explicitly stated and may be a request for information, statement or directive directed at recipient, who is expected to reply. Open-ended invitations to reply ("Let me know if you need more") should be coded as Rn.			
Is S providing info in response to an explicit request?[9]	S is providing an answer to a question or providing info that R requested. "The sales figures are…"; consists of statements of fact; the refusal to provide info or declaration of lack of knowledge should be coded as R5.	R1 F1	
Is S offering or committing to do something?[1]	S is offering or committing to do something, including providing additional information or attending a meeting. "I'll get the data to you by Friday." Include here statements of S's already completed actions in response to R's initiative, "I've sent that report this AM." Include conditional commitments, "If you send the report, I'll go."	R2 F2	
Is S agreeing or disagreeing?	S may give a "yea,", "nay," or a mixed response or commentary to R's statement(s), idea(s), directive(s), or commitment(s); "on topic". Includes corrections.	R3 F3	
Is S suggesting R take some other action?	Include directions that R take some action (other than a commitment previously proposed by R). "See HR about that." "You need to send the report to…" Include suggestions that R provide additional info. "Please send me the data to which you're referring."	R4 F4	
Some other response?	Includes simple acknowledgement, thanks. Include emotional responses and "chit-chat."	R5 F5	
(S is initiating dialogue)			
Code as the forward speech act first encountered in the series of questions. This gives "precedence" to those speech acts that require action or response from the recipient.			
Does S want R to do something?	S wants the dialog to result in R's doing something, verbally or otherwise.		
Perform some action other than provide info?	S tells R to do something other than assembling and transmitting data. "Please send this to all committee members." "Please clean the labs."	D1	

[9] Exception to the chronology of annotation implicit in the flowchart: In cases where a response includes both information (R1 or F1) AND an offer or commitment (R2 or F2), code the entire email according to the act that appeared first in the email.

	S asks R to do something and R can reject the suggestion. "Would you send me the book?" "Can you come…" Verification from R is expected.	D2	
Request information?	A directive to provide information. The form of the information is assumed to be unknown to the sender: requests for specific reports or documents (containing desired information) should be coded as D1 or D2. "Send me the figures."	I1	
	Answer a simple question or provide other information, including feedback (comments). "Who should we invite?", "Can you tell me what is required?". "What do you think about this?"	I2	
Does S offer or commit to do something?	Offer—may or may not. "Would you like me to…", "I can…" Include conditional commitments, "If you send me the report, I'll go to the meeting."	C1	
	Commit—definitely will. "I'll send it tonight", "I'll attend."	C2	
Does S make a clear statement or give info about the world?	S making statements of fact. S is stating his/her opinions.	A1 A2	
Does S express his/her feelings?	"I'm sorry…" "Thanks so much." "I'm exhausted" "Welcome" "It's a shame."	B1 B2 B3	
Does the message perform the act?	The statement of the words actually accomplishes an act. "You are awarded the contract", "You're guilty". Probably very rare in email.	V1	
Another forward function	Includes general friendly hellos, introductions.	O1	

8.3 Personal Email Corpus

We prepared a set of approximately 280 randomly selected, redacted emails from the authors' personal email collections. The first 160 emails were used to develop and refine the annotation guidelines. One hundred emails were used to test inter-annotator agreement. Using the kappa statistics [Cohen 1960], we obtain a kappa of 0.89 for the 30 subcategories, indicating high agreement between the two annotators.

For such a small set of emails, it was difficult to obtain enough samples for each category. Accordingly, we merged all subcategories into main categories and eliminated some categories with very small samples, namely S, N, B, V, O. Commits were only found in responses and and were annotated as R2 or F2 (see Figure 1). D and I (Directives and Information requests) were merged into one category I&D. We then supplemented any deficient main categories by hand selecting samples in the resulting five categories. The result was 50-56 emails for each of the five main categories: transmissives (T), requests, either directives or information requests (I&D), assertions (A), responses (R), and responses with an expectation of reply (F) .

8.4 Features

To detect the overall intent of the email, each email message is represented by a set of features. The focus is only on the text that the sender has written; as such we have implemented methods to eliminate signature blocks and included text. We investigate two different types of features sets: one based on verbs and the other based on email specific characteristics.

8.4.1 Verbs

Levin shows the correlations for a large set of English verbs (about 3200) between the semantics of verbs and their syntactic behavior and interpretation of their arguments. From these correlations, she defines classes of verbs [Levin 1993]. Each verb class has subclasses; using the most general class results in 48 total classes. These general classes of verbs have been reformulated in the Lexical Conceptual Structure (LCS) Database by using alternations [Dorr and Jones 1996], resulting in 29 additional classes, for a total of 77 classes [LCS]. Each class is used as a single feature in the feature set. From the email text, we count the stemmed verbs using a simple string matching algorithm to the verbs in LCS, supplemented with the expansions of irregular verbs [Byrd 2005]. We use two methods of computing counts. One method matches stemmed words in the text to the verbs in the verb classes, the other uses the TnT Part of Speech tagger [Brants 2000] to identify that the part of speech class is indeed a verb (since noun roots can appear identical to verb roots). Membership in each of the LCS verb classes is one feature of the feature vector. We make a pass through the text, identifying verbs (by comparing a stemmed word in the text to a stemmed version of the verbs in LCS). If a verb is in multiple classes, each class is proportionally incremented so that each verb contributes one unit. The final counts are normalized by the word count of the sender's body of the email (that excludes punctuation, signature block, included text from another email, etc.).

Ballmer and Brennenstuhl (BB) present a classification of speech acts based on linguistic activities and aspects [Ballmer and Brennenstuhl 1981]. These were originally devised for German verbs and then translated into English, resulting in many multi-word "verbs" in the 600 categories and 24 classes. As a first pass, all multi-word entries in the BB classes were ignored. As with LCS, each BB class was supplemented by irregular tenses of verbs and each class is used as a single feature in the feature set, resulting in 24 features. Verb counts for the features were computed using the same process as for LCS.

8.4.2 Email Specific Features

Certain types of email contain features that indicate particular communicative intents. These features can be presentation oriented (which include features found in header information or in punctuation) or text oriented. For example, the presence of "Re:" in the subject line usually indicates a response, although sometimes writers shift topics or introduce new topics without changing the subject line. Similarly, "Fwd:" often signals a transmissive intent. The presence of interrogative sentences (detected by the presence of question marks in the body of the email) can identify an information request.

In addition, certain words can indicate various email speech acts. For example, "Thanks" and its variants can indicate an acknowledgement. "Attached", "Enclosed", "Here is" signal the presence of an attachment; the file name of the attachment may appear in the header or in the body of the text depending on the email software.

Table 8-5 summarizes an initial list of additional email features, based on form and content – to assist in genre identification (refer to Table 8-3). All features are either binary or were normalized over the document length (word count or sentence count). All

words/phrases that were used are listed in the table. In the future, we plan to do further analysis to expand this list

Table 8-5: List of 16 email speech act features

EMAIL CHARACTERISTICS FEATURES (EF)
Presence of Re:
Presence of Fwd:
Attachment signified in header info or by an insertion in text body
Fraction of interrogative sentences (sentences ending in '?'/total sent)
Fraction of "I" or "we" (count of words / total word count)
Fraction of "You" (count of words / total word count)
Attachment indicators such as "attached, here is, enclosed"
Apology indicators such as "sorry, apology, apologies"
Opinion indicators: "think, feel, believe, opinion, think, comment"
Politeness indicators such as "please"
Gratitude indicators such as "thank"
Action indicators such as "can you", "would you"
Commitment indicators such as "I can", "I will"
Information indicators such as "information", "info", "send"
Auto-reply indicators such as "out of the office", "away"
Email length

8.4.3 Classification

Two different classifiers were used for our experiments: Support Vector Machine [19] and Random Forests [9]. We used SVM-light [32] with a radial basis function and the default settings. SVM-light builds binary models, so in our cases, where we have five classes, a model must be produced for each class.

In contrast, Random Forests [28] grows many classification trees and each tree gives a vote for a particular class. The forest chooses the classification having the most votes. We use 100 trees in our experiments. Random Forests takes far less training time than SVM-light.

The experiments on the data described in Section 4 were run using the ten-fold cross-validation method. This splits the data into training and test sets with a 90% training, 10% testing portion. Experiments are repeated ten times, so that all the data is used both in training and testing but not all at the same time.

We compared classifier performance using the LCS verb features, the BB verb features (Section 4.1) and the email characteristics feature set EF (Table 8-5), as well as combinations of these feature sets. The results for the Random Forests classifier are presented in Table 8-6. Recall (percentage of emails correctly classified), Precision (percentage of classifications that were correct) and F1 (the harmonic mean of recall and precision, $F1 = 2*R*P/(R + P)$) are all equal.

Table 8-6: Results (precision) of Random Forests Classifier for identifying the five email act classes (T, I&D, A, R, F) without and with part-of-speech tagging TnT.

	No TnT	TnT
EF	.57	N/A
BB	.38	.30
BB+EF	.63	.52
LCS	.35	.34
LCS+EF	.57	.54

Table 8-6 shows that the email characteristics feature set (16 features) does very well (.57), outperforming both LCS and BB (.32 and .38 respectively). This is not a surprise since many of the features give clear indications of the appropriate category – such as Re: for a response. Adding the verbs features to the email characteristic feature set slightly increases performance, for both verb feature sets. BB+EF results in the best feature set combination with a precision of .63. TnT decreases performance when used in combination with BB, but not when combined with LCS. In the future we hope to determine what characteristics of the verb lexicons are contributing to such results.

We also compared Random Forests and SVM-light for the LCS feature set (Table 8-7). The results indicate that these two classifiers are often very close in performance. The one exception to this was LCS.

Table 8-7: Results (precision) of Random Forests compared to SVM-light for LCS on the five email act classes (T, I&D, A, R, F).

	SVM-light	Random Forests
LCS	.28	.35
LCS + TnT	.35	.36
LCS+EF	.59	.57
LCS+EF+TnT	.55	.54

The confusion matrix for Random Forests on the email characteristics feature set EF is displayed in Table 8-8. Table 8-9 and Table 8-10 show the matrix for the verb features sets BB and LCS respectively. The results in Table 8-11 clearly indicate the improved performance of the combination of BB and the email feature set EF.

From Table 8-8, Table 8-9 and Table 8-10, we can see that using BB alone and LCS alone results in higher classification accuracy for Directives (I&D) and Assertions (A) than that of just EF. This indicates that we would either need to expand our email characteristics set to include more distinguishing features for I&D and A or use verb classes of the type found in BB to assist in such characterizations. However, using the verb classes of BB combined with EF resulted in a decrease in performance for Responses with expectation of reply (Table 8-11).

Table 8-8: Confusion matrix for EF (Random Forests) – 16 features.

Class	# Items	T	R	F	I&D	A
T	55	**81%**	2%	6%	7%	3%
R	56	11%	**67%**	15%	2%	4%
F	56	3%	11%	**83%**	2%	2%
I&D	53	14%	6%	33%	**27%**	21%
A	50	29%	11%	11%	22%	**28%**

Table 8-9: Confusion matrix for BB only no TnT (Random Forests) – 24 features.

Class	# Items	T	R	F	I&D	A
T	55	**68%**	5%	10%	8%	8%
R	56	38%	**10%**	19%	17%	16%
F	56	14%	14%	**32%**	24%	16%
I&D	53	8%	14%	26%	**31%**	21%
A	50	9%	8%	18%	18%	**48%**

Table 8-10: Confusion matrix for LCS only no TnT (Random Forests) – 77 features.

Class	# Items	T	R	F	I&D	A
T	55	**62%**	13%	6%	12%	7%
R	56	27%	**19%**	18%	19%	16%
F	56	17%	22%	**18%**	25%	18%
I&D	53	16%	18%	24%	**32%**	10%
A	50	12%	13%	12%	18%	**44%**

Table 8-11: Confusion matrix for BB+EF no TnT (Random Forests) – 40 features.

Class	# Items	T	R	F	I&D	A
T	55	**82%**	2%	4%	8%	4%
R	56	12%	**65%**	14%	2%	8%
F	56	5%	14%	**71%**	7%	3%
I&D	53	8%	9%	18%	**43%**	22%
A	50	9%	11%	4%	23%	**53%**

Only for Assertions (A) do both the feature sets LCS and BB perform clearly better than EF. Both verb lexicons have difficulty with the response categories (both R and F). LCS outperformed BB only for responses R. It is the intent of LCS to classify verbs by analyzing the grammatical structure in which they appear. We did not grammatically parse the text of the email; this may have contributed to the poor functioning of LCS.

For the email feature set (EF), Responses with expectation of reply (F) are also often confused with Responses (R) and vice versa, a result we might expect since both classes involve backward communicative functions. We assume that the use of the question marks as a feature in EF helps to distinguish between these two classes as compared to LCS and BB. For the verbs features only (BB and LCS), Responses (R) are most often mis-identified as Transmissives (T) and Responses with an expectation of reply (F) are most often confused with I&D (requests and directives).

Overall, the Transmissive category (T) is the easiest to identify. We believe this is due to the fact that the form and content of this genre often clearly identifies this category. The remaining four categories consist of the email conversation genre.

We have shown that the combination of the feature sets of verb classes with email specific characteristics can give reasonable classification performance for five email act categories and two email genres (email conversations and transmissives). In the future, we want to investigate methods to improve our performance results as well as develop methods to distinguish the various genres listed for each email act category in Table 8-3. We would like to incorporate grammatical structure in identifying LCS class membership and fold in the multi-word "verbs" of BB and/or replace multi-word verbs with equivalent single words (e.g., "anger" to replace "make angry") to investigate whether this would improve results. We would also like to compare these results for verbs classification to that obtained by using VerbNet [37]. We also plan to collect and annotate more email in the specific categories to be able to test the classifiers for all the subcategories of Figure 1.

We believe our findings support the characterization of email as an amalgam of unique communicative genres, where the common genre – email conversations is most similar to spoken communication. The methods we employed, built upon methods used for analysis of dialog, are practical and produce meaningful categorization of email. We hope our research spurs additional research to understand the form and content, as well as the overall genre, of email.

8.5 Enron Email Corpus

The purpose of the Enron email annotation project was to create an annotated corpus that could be used for further email research.[10] In 2003, the Federal Energy Regulatory Commission (FERC) as a result of its investigation of Enron's energy trading practices [FERC] made available to the public the Enron email corpus. Portions of two subsets of this corpus were annotated: The BEAAP Partial Corpus and the Joint Corpus [Goldstein et. al. 2006a].

The BEAAP Partial Corpus (BEAAP-PC) is a subset of the UC Berkeley Enron Email Analysis Project (BEEAP) (http://bailando.sims.berkeley.edu/enron_email.html), which consists of approximately 1900 genre-labeled emails from the Enron corpus. Selected emails are primarily related to business and specifically to the California Energy Crisis; each is labeled with a category. The second was a subset of emails from the Voice Transcripts Email Correlated Corpora. We created this data set by selecting emails from authors for whom there are also available audio files and corresponding voice transcripts.

We filtered the BEEAP Enron subset, removing emails in which the forwarded information would not be interesting to annotate. This resulted in the removal of emails containing forwarded news articles, government and academic reports, press releases, pointers to urls, newsletters, and jokes (see Table 8-12).

[10] Due to some issues in the annotation of this corpus, it is not clear how useful it will be.

Table 8-12. BEEAP: Enron Coarse Genre Email Statistics

Coarse Genre	BEAAP	BEAAP-PC Corpus
Company Business	855	304
Purely Personal	49	33
Personal Professional	165	104
Logistics	533	355
Employment	96	68
Document Collaboration	176	111
Missing Attachment	25	21
No Sender Text Body	26	17
Total Emails	**1925**	**1013**

To create the Joint Corpus we identified Enron employees in the phone call from the Voice Transcripts and then selected all available emails for that employee. All the phone calls were saved by a system that recorded Enron's Western electric traders' operations. To identify the Enron employees in the calls we used a combination of heuristics and inference from objective evidence. In some cases it was possible to identify a party from the call's conversation flow. In other cases, we took advantage of the fact that calls from the same trader tend to appear in the same channel of the recording system (FERC Exhibit SNO-161) [FERC]. Using all this information we classified participants as "certain", "probable", or "unknown". The "certain" category was used for those Enron employees who are identified without doubt as participating in at least one call. The "probable" category was used for those Enron employees who may be participating in at least one call. This category occurred when we identified the first name of one of the call parties and there was a match between this name and the name of an employee who used the recording channel corresponding to the call. A person identified as both "certain" and "probable" was assigned the overall category "certain".

8.6 Annotation of Email Corpus

Besides the internal annotations of the email text body which consisted of tagging entities, some relations and the best summary sentence (marked with tag SUM) for the email not including the subject line of the email, annotators were supposed to tag four overall email document properties, listed below. All of these annotations apply only to the "top level" message, in cases where there are layers of replies/forwards.

1. Subject line alignment with content of the text body of the email (discussed in the next section).
2. Email speech act of the sender (annotated using the 30 subcategories in Figure 8-1 and Table 8-4
3. Mention of a face-to-face meeting in the past, present or future, marked yes, no, or unclear. The unclear case covers both possible future meetings and the case where there is ambiguity as to whether the meeting was face-to-face or via telephone.

4. Mention of a telephone conversation in the past, present or future. These were marked in the same manner as face-to-face meetings.

The annotation guidelines were given to the task leader with the instructions to ask any questions as needed, however, no clarification questions were asked. The task leader formed the annotation team, which consisted of at least five people:

For all items, inter-annotator agreement was supposedly to be performed and there were target guidelines for agreement of 80%. From analysis of the delivered corpus for email speech acts (item #2), it does not appear for the items that were doubly or triply annotated that the target inter-annotator agreement was met. Thus the quality of the overall data may be questionable.

Although the quality of this corpus is questionable, we can still analyze the data.

8.7 Email Summarization Results

In this section we analyze the human annotated Enron corpus in terms of the use of the subject line and one line summary sentence for summary purposes.

8.7.1 Email Subject Line and Content

Subject line alignment was supposed to determine whether the subject line accurately reflects the content of the email. This can be important because besides the sender, the subject line is what gives a clue to the reader of whether or not he/she wants to read the email and whether or not to prioritize this email in the queue (assuming it is not marked with a tag such as urgent). If there are a series of emails in an email *thread*, the subject line may no longer reflect the topic of the original email – we refer to this as *topic drift*.

For this task, we envisioned three possible values: (a) *content summary* – the subject line accurately describes the main purpose of the email: the reader can correctly surmise the intent of the sender without reading the email, (b) *connected* – may describe the purpose of the email, but provides little content, for example, "a question", "status", or "tablet pc again," or (c) *unconnected* – the subject is not relevant to the email topic, such as may result from topic drift or the sender not entering a subject.

For the subject line alignment task, the annotators did not mark them in the manner described. Furthermore, it is not clear if they had any training, or if any inter-annotator agreement was performed. The team used the following guidelines:

"In this task we mark how well the subject line of an email message matches the actual content of that message. Note that this overlaps partially with the use of the SUM tag [which is a tag used to mark a Summary sentence in the text body of the email], except that the SUM tag is used for only a single sentence and this task is concerned with the overall content. There will be cases where there is a good match between the subject line and the content, but no good match between the subject line and the SUM'd line, because

there may be no single line which fully captures the overall content of the message. This is a three-way decision:

1. Perfect. The subject line refers to the same content as the body of the email.
2. Partial. The subject line captures only part of the content of the body of the email, or captures it in a very imprecise way. The former indicates that a conversation is drifting away from the original topic, the latter that a sender doesn't provide very informative subject lines.
3. No Match. There is no relation between the subject line and the content of the body of the email. This can indicate either that a conversation has completely abandoned the original topic, or more frequently that a sender used only a very generic subject line like "FYI". Blank subject lines also fall in this category."

Some annotators used "good", which the task lead claimed was the same interpretation as partial, and some used "bad" instead of "no match".

The results are this are shown in Table 8-13.

Table 8-13: Subject Matches Content

Subject Line Evaluation	NumberMarked in Data	Percent
Perfect	1198	56%
Partial	508	24%
No Match	435	20%

There are 2141 total documents, although some might be duplicates. Table 8-13 shows that there is a fair amount of topic drift or blank subject lines in this corpus – 20%. It also shows that for the emails that are not in this category, the assessors thought 70% of the subject lines (of the 1706 that were tagged perfect or partial) are good content matches to the email text body.

8.7.2 One Email Text Body Sentence as a Summary

For emails, we were interested in whether the first sentence of the text body might be an effective summary sentence. For the newswire genre, we had previously determined that the first sentence is used in generic summaries approximately 70% of the time [Goldstein et. al. 1999].

In order to determine this, the annotators were supposed to choose and tag the best one sentence summary of the email text body (the portion of the email written by the user not including the subject line) of the last sender's email. The subject line was not to be included. For this task, it appeared that these instructions were not followed. There were cases where only the Subject line was marked as the Summary "sentence" and other cases where both a Subject line and a Summary sentence was marked. In addition, Summary sentences were not marked for the entire corpus – they only appear to be tagged in 1215 of the total 2045 annotated documents. The results are shown in Table 8-14.

If the cases where the subject lines are chosen as the summary sentence are excluded (column 4 in Table 8-14), then in 66% of the "useful" emails (725 of 1092), the first sentence would be an effective summary sentence for this corpus.

Table 8-14: Distribution of Summary Sentence Number in Email Text Body

Sentence Number	Subject line counted Number Selected (%)	Summary sent. counted Number Selected (%)	Summary sent. counted (excluding subjects) Number Selected (%)
Subject	254 (21%)	123 (10%)	
1	659 (54%)	725 (60%)	725 (66%)
2	176 (14%)	191 (16%)	191 (17%)
3	55 (4.5%)	79 (6.5)	79 (7.2%)
4	27 (2.2%)	43 (3.5%)	43 (3.9%)
5	17 (1.4%)	20 (1.6%)	20 1.8%)
6-10	20 (1.6%)	24 (2.0%)	24 (2.2%)
11-15	6 (0.5%)	7 (0.6%)	7 (0.6%)
20-28	1 (0%)	3 (.0.2%)	3 (0.3%)

8.7.3 One Email Text Body Sentence as a Summary

As a last experiment, we decided to rate the subject lines and summary sentences. For this task, we randomly chose 50 emails from the 1092 available emails with SUM tags in the email text body. Two of these files had to be eliminated because the taggers did not choose summary sentences from the last senders email. Introductory and closing salutations were not counted as sentences and eliminated from the sentence count. The corpus contained an average of 3.4 sentences per the last sender's email (Primary Email) and an average of 2 emails per thread – where a thread is defined as the "conversation" ongoing between multiple email senders and recipients, and must be contained within the last sender's email whether or not they are relevant to the current topic being discussed. Forwarded items or inline news articles do not count as part of the sentence count. The distribution of the number of sentences is shown in Table 8-15 and the summary sentences distribution is shown in Table 8-16.

Table 8-15: Distribution of Primary Email Length in Evaluation Corpus

# of Sentences in Primary Email	Count (%)
1	14 (29%)
2	9 (19%)
3	9 (19%)
4	6 (12%)
>5	10 (21%)

Table 8-16 shows that the distribution of summary sentences in this corpus is reasonably close to that in Table 8-14. However, it is important to note that 29% of this corpus (Table 8-15) only has one sentence. Accordingly, we decided additionally to tabulate and

report statistics for only the portion of the corpus that contains more than 2 sentences in the text summary body. This subset of the corpus has first sentences chosen as the summary sentences, 56% as compared to 69% (Table 8-16).

Table 8-16: Distribution of Selected Summary Sentence in Evaluation Corpus

Summary Sentence Number	48 doc Count (%)	25 doc Count (emails >= 3 sentences)
1	33 (69%)	14 (56%)
2	9 (19%)	5 (20%)
3	2 (4%)	2 (8%)
4	1 (2%)	1 (4%)
5	1 (2%)	1 (4%)
7	2 (4%)	2 (8%)

To test, whether or not subject lines and summary sentences make good summary sentences, we asked people to answer the following questions for the 50 emails:

Q1. Is the subject a good summary for indicating the content of the email
    1. Yes. 2. Partially. 3. No

Q2. Suppose that the subject line is going to be used to choose an appropriate action for this email (for example, read it immediately or save it for later), For making such a decision, is the subject line
    1. Not informative at all (missing, topic drift, etc.)
    2. Not informative enough
    3. The right level of information
    4. More information than necessary
    5. Far too informative

Q3. Is the one line sentence a good summary for indicating the content of the email
    1. Yes. 2. Partially. 3. No

Q4. Suppose that the summary is going to be used to choose an appropriate action for this email (for example, read it immediately or save it for later), In terms of summary length for this task, is the one sentence summary
    1. Far too short
    2. Too short
    3. Just right
    4. Too long
    5. Far too long

13 people participated in this survey and their responses as well as the averages are summarized in Table 8-17. The results of the survey tabulated for emails with 3 or more sentences are shown in Table 8-18. Note that the scores are pretty much the same as in Table 8-17.

Table 8-17: Results of Survey on Subject and Summary Line - 4 questions (48 emails)

The letter, A-M, is a label for one of the 13 raters.

	A	B	C	D	E	F	G	H	I	J	K	L	M	Avg
Q1	1.4	1.5	1.9	1.5	1.5	1.3	1.4	1.4	1.3	1.9	1.3	1.9	1.3	1.5
Q2	2.5	2.5	2.3	2.7	2.3	2.8	2.6	2.7	2.6	2.1	2.7	2.0	2.7	2.5
Q3	1.2	2.0	1.4	1.8	1.1	1.1	2.0	1.4	1.7	1.8	1.2	1.8	1.3	1.5
Q4	2.8	2.2	3.3	2.8	2.2	2.8	2.3	2.8	2.4	2.3	2.8	2.2	2.8	2.6
Q4 % of 3s	77%	23%	63%	56%	81%	88%	23%	77%	44%	25%	81%	38%	77%	58%
Q4 % of >=3	77%	27%	92%	67%	81%	88%	31%	80%	44%	25%	81%	38%	77%	62%
Q4-1	1	8					9	4	5	2		8		2.8
Q4-2	10	27	4	16	9	6	24	6	22	34	9	22	11	15
Q4-3	37	11	30	27	39	42	11	37	21	12	39	18	37	29
Q4-4		2	10	5			3	1						1.6
Q4-5			4				1							0.4

Table 8-18: Results of Survey on Subject and Summary Line - 4 questions (25 emails >= 3 sentences)

The letter, A-M, is a label for one of the 13 raters.

	A	B	C	D	E	F	G	H	I	J	K	L	M	Avg
Q1	1.4	1.5	1.8	1.6	1.4	1.3	1.3	1.2	1.2	1.9	1.4	1.9	1.3	1.5
Q2	2.6	2.5	2.4	2.6	2.4	2.8	2.8	2.9	2.6	2.1	2.6	2.1	2.7	2.5
Q3	1.2	1.9	1.4	1.8	1.1	1.1	2.0	1.4	1.6	1.7	1.3	1.9	1.3	1.5
Q4	2.6	2.1	3.4	2.9	2.8	2.8	2.1	2.7	2.4	2.3	2.8	2.2	2.8	2.6

From Table 8-17, the rater's selection of 3 (the right length of the summary in Q4), had a huge variance - from 23% to 88%. Upon questioning the raters, the answers were determined to have this distribution from various interpretations and assumptions about the survey:

- *As summarizing an entire email thread*: Some raters interpreted that for the survey they were looking for a summary sentence for the entire email thread rather than just the last sender's email
- *Factoring in author's writing and lack of context*: Some raters judged the writing of the author. If the writer of the email did not provide enough information for the rater to understand what the writer was discussing, the rater gave the subject line and/or summary sentence a low score. If this was one's personal email, ideally one would not have this problem – since the context would be known to the receiver and any information passed via telephone calls or in-person meetings would also be known.

- *Assuming subject line of email is known*: Some of the summary sentences require the subject line to interpret the sentence. The text in the email body refers to the subject line as if it is part of the document. Some raters judged the summary sentence independently (as they were supposed to), others ranked the summary sentence as if the subject line was known.

We therefore decided to conduct another survey. This time we informed the raters, that they were only judging the last email and not the entire thread, and that the thread was there for context. We had thought this would have been the interpretation from the question which stated that the rater should judge the email based on whether they could take some action on the email (see question 4), but apparently it wasn't. In addition, since the summary sentences often seemed dependent on the subject, we added three more questions to analyze the effects of subject and summary sentence combined.

To test, whether the subject line combined with the summary sentence forms a good summary, we asked people to answer the following questions for the 50 emails:

Q5. Is the combination of the subject and the one line sentence a good summary for indicating the content of the email

1. Yes. 2. Partially. 3. No

Q6. Suppose that the combination of the subject and the summary line is going to be used to choose an appropriate action for this email (for example, read it immediately or save it for later). For making such a decision, is the combination of these two

1. Not informative at all (missing, topic drift, etc.)
2. Not informative enough
3. The right level of information
4. More information than necessary
5. Far too informative

Q7. Suppose that the combination of the subject and the one line summary is going to be used to choose an appropriate action for this email (for example, read it immediately or save it for later). In terms of summary length for this task, is the combination of the subject and the one sentence summary

1. Far too short
2. Too short
3. The right length or about the right length
4. Too long
5. Far too long

We asked three of the raters with low scores to redo the survey and had 3 new people complete it. One of the other raters without new low scores also did the survey. The ratings for the seven are shown in Table 8-19 for the 48 emails and in Table 8-20 for the 25 emails that are longer than 3 sentences. The raters from the initial study are coded with their label letters from Table 8-17.

Table 8-19: Survey Results on Subject and Summary Line – 3 new questions (48 emails)

The letters {A, B, I, J} and the numbers {1,2,3} indicate one of the 7 raters

Question #	A	B	I	J	1	2	3	**AVG**
Q5	1.04	1.16	1.0	1.13	.13	1.19	1.63	**1.18**
Q5 - % of 1s	96%	88%	100%	88%	88%	85%	48%	**85%**
Q5 = % of 1s and 2s	100%	96%	100%	100%	100%	98%	90%	**98%**
Q6	2.94	2.88	3	2.69	2.85	2.83	2.79	**2.86**
Q6 - % of 3s	94%	76%	100%	73%	88%	85%	79%	**85%**
Q7	2.87	2.96	2.98	2.69	2.92	2.90	2.69	**2.87**
Q7 - % of 3s	98%	88%	98%	69%	90%	85%	69%	**85%**

Table 8-20: Survey Results on Subject and Summary Line – 3 new questions (25 emails >= 3 sentences)

The letters {A, B, I, J} and the numbers {1,2,3} indicate one of the 7 raters

Question #	A	B	I	J	1	2	3	**AVG**
Q5	1.08	1.33	1	1.2	1.16	1.28	1.6	**1.24**
Q6	2.92	2.89	3	2.64	2.84	2.76	2.76	**2.83**
Q7	2.96	3	3	2.64	2.88	2.8	2.56	**2.83**

Table 8-19 shows that approximately 85% of the time, people think that the subject and summary line are about the right length for a summary (Q7), provide the right level of information (Q6) and that it forms a good summary. They think this combination forms a good or partially good summary 98% of the time.

Comparing Table 8-20 to Table 8-19, the 25 emails longer than 3 sentences had slightly lower scores (0.04 to 0.06) for judging the effectiveness of the combination of subject line and summary.

8.8 Discussion

In this chapter we have shown that we can determine the communicative intent of an email (for a small corpus) approximately 60% of the time. Since this corpus had less than 60 samples for category, from our experiment with genre identification in Chapter 6, we think that if we double the number of samples, the performance will increase.

From a study of human annotations of the Enron corpus, we found that 66% of the time, the annotators thought that the first sentence makes the best summary sentence. Perhaps Google (gmail) concurs since they use their one line screen real estate to display first the subject line and then the beginning of the email.

We have also demonstrated through a survey, that 85% of the time, raters thought that the subject combined with the one line summary makes a summary that is both informative

and of approximately the right length for deciding whether or not to take some sort of action based on this summary. In addition, 85% of the time the raters thought that this was a good summary and 98% of the time that this was either a good summary or partially a good summary.

To summarize, for the email genre, we have motivated the use of a genre specific summary consisting of the subject line and one summary sentence from the text body for the purpose of deciding whether to take some action. The summary length is less than 2 sentences. For the purposes of summarizing an entire email thread (such as lawyers may want to do after subpoenaing email), this may not be an effective summary. Such a summary may require sentences from various emails in the thread. In addition, the email thread may lack information that has transpired through meetings or telephone calls. The summary of email threads is left for future work.

Chapter 9 Multi-document Summarization

"Do not say a little in many words but a great deal in a few."
Pythagoras

In this chapter, we present our multi-document summarization experiments. We describe the data collection process, the experiments performed and our results. We use the abbreviation MDS to refer both to multi-document summarization and multi-document summary/summaries.

9.1 Multi Document Evaluation Corpus Description

In Section 2.2.1, we presented the fact that users have various information seeking goals and in Section 2.5 we presented factors to consider when constructing multi-document summarization systems. An ideal multi-document summary must contain the relevant information to fulfill a user's information seeking goals, as well as eliminate irrelevant and redundant information. A first step in creating such summaries is to identify how well a multi-document text summarizer can extract what people perceive as key information and to evaluate types of data sets that reflect the user's information seeking goals for multi-document summarization. As motivated in Chapter 1, the standard information retrieval technique of using a query to extract relevant passages is no longer sufficient for multi-document summarization (MDS) due to the high levels of redundancy that occur in large data sets. In addition, constructing a query relevant summary using just the rank order of passages retrieved does not capture the temporal sequence of the information.

We constructed a gold standard corpus to study the effects of anti-redundancy and other features on MDS quality. Furthermore, these sets allow us to examine how humans perform the MDS task in a constrained setting.

We chose the newswire domain for our experiments due to its specific characteristics. In particular, we can construct topic clusters in this domain to contain articles with:

- daily redundancy due to reporting by multiple news sources (often even from the same original source),
- temporal redundancy due to the fact that events tend to occur over varying time windows and background information is repeated in the articles.

Furthermore, the newswire domain allows

- the ability for high compression ratios from the selection of multiple articles to summarize, and
- an inherent temporal dimension due to the nature of news. In addition, newswire articles have been extensively studied in single document summarization [Jing et. al. 1998, Mani et. al. 1998].

Specifically, we constructed a database of newswire articles consisting of 30 sets of 10 newswire articles collected from Yahoo, CNN and BBC. A topic was selected based on the Yahoo category and then 10 articles were selected from the Yahoo results or in some cases retrieved

from CNN or BBC based on the previous related articles listings on the web pages. The main factor for selection was the article's timestamp so that we could do a preliminary analysis of a snapshot of an event from single and multiple sources, as well as the unfolding of events. To this end, we did some small filtering of the articles to collect 10 relevant articles to the topic. Each set was arranged by its timestamp; the earliest article labeled as article 1 and the most recent as 10.

Our sets consists of the following four types of article clusters:

- **Multiple Sources, Snapshot**: (9 data sets) A snapshot of an event from multiple sources (e.g. the first report of an airline crash). This reflects what might happen if one did a search for articles on a certain date from a collection of various news sources, such as Yahoo.
- **Same Source, Snapshot**: (5 data sets) A snapshot from the same source (the first 10 articles from the same source on an airline crash possibly spanning a few days). This type of data would occur if one did a search on a news site such as BBC or CNN, which have frequent updates.
- **Event Unfoldment - short timespan**: (7 data sets) Articles from the same or different sources spanning a week to six weeks. This type tends to contain articles on the same event as well as articles on the same type of event occurring in different locations - such as the millennium flu bug affecting people worldwide.
- **Event Unfoldment - long timespan**: (9 data sets) Articles from the same or different sources spanning months to years. These sets also contain update sets on the same event from a later date, such as the last ten articles posted to Yahoo on the TWA 800 plane crash at the time the data was collected.

Our data set was designed to allow us to produce specific types of summaries for analysis:

- **Event Comparative Summary**: (2 data sets) of an event or events across geographic dimensions. One set is on the millennium flu bug (multiple sources snap shot) and the other on shootings in schools (event unfoldment long). These sets are specifically designed for "comparative summaries" - comparing related events.
- **Update Summary**: (4 data sets) updates of a previous set topic in a new time period. Three are on airline crashes and one is on the Discovery shuttle.

In order to analyze the effects of summary "pollution" from unrelated sets (not examined in this thesis) as well as the ability of our summarizer to produce comparative summaries, we also collected data sets from three particular topic groups: airline crashes (9 sets), space missions (5 sets), and shootings (3 sets). The airline update crash test set allows us to compare how well our summarizer can provide an update summary based on a previous summary for the set ("summary evolution summary"') to a new summary for that set ("update set summary").

To provide summaries as well as other information for the data set, we used three assessors who are college graduates but who, unfortunately did not have formal training in journalism, summary creating or analyzing data[11]. Although, three people are most likely insufficient for

[11] The multi-document summarization study was performed before the genre-oriented summarization study. We used the experience we had gained in both this study and the construction of the relevant sentence single document set (described in Goldstein et. al. 1999) to design what we hope is a better gold standard set for the genre-oriented summaries.

summarization task ([Mani et. al. 1998, Teufel and van Halteren 2004, Nenkova et. al 2007]), we used three people to allow for some of the variance that occurs within this task. The resultant human sentence extract summaries and subtopic scoring of sentences provide a means for evaluating our machine generated multi-document summaries that does not require repeated involvement of human judges. Ideally, at the least, analyses using this data can identify summarization approaches that are worth further pursuit.

We chose ten sentences as the summary length for a MDS – designed to represent a "short summary", which would approximately fit in one computer window display and provide a flavor of the set without requiring scrolling. We also created topics for each of the sets so that we could explore query relevant summaries as well as generic summaries.

For each set of 10 articles, the three assessors selected the following:

- *multi-document summaries*: The assessors selected the 10 most important sentences (in rank order) for the set of articles based on the generic or query relevant task. The assessors were then asked to arrange the sentences in a *readable order*, i.e., an order which flows well, is intelligible and is as coherent as possible from a group of disjointed sentences. The assessors formed two types of summaries:
  1. *sentence extract generic (overview) summary* covering the information they thought was most important for the topically related cluster.
  2. *sentence extract query relevant summary* covering the information that is most relevant to a specified request for certain information; e.g., the current theories as to why an airplane crashed.
- *the three most informative articles in a cluster*: The assessors selected the most informative article to give to a reader, if they could only provide one article. They then chose the second most and third most informative articles.
- *sentence subtopic assignment*: The assessors assigned each sentence to one or more provided human generated subtopics for the cluster.
- *most important subtopics in a cluster*: The assessors selected the most important subtopics of the provided subtopics.
- *generic single document summaries*: The assessors selected the three most informative sentences for each article.

The assessors marked their selections on paper and not on the computer. Since it could be difficult to view the constructed multi-document summaries, after they had made their selections, we gave them copies of the summaries (in readable order) and allowed them to change any sentences.[12]

These collected sets of 10 sentence summaries serve as our multi-document gold standard for system generated summaries to answer the questions – "Does our summarization system pick similar summary sentences as compared to humans?" and "Does the summarization system pick sentences covering the same subtopics as the human selected sentences?" Section 9.3 addresses these questions.

[12] If repeating this study, we would use a combination of paper and computer. The paper for reading the text that would contribute to the summary. The computer for constructing the summary, so the assessors could see what their summaries looked like as they created them

9.2 Data Sets Analysis

In this section we present an analysis of the properties of our collected data and in particular, compare the human selected multi-document sentence extract summaries to their single document counterparts.

The 30 data sets averaged 301 sentences in length, with the minimum set sentence length of 177, maximum of 516 and a median of 293. Thus, the human summaries are compressed to 3.3% of a multi-document cluster set (Table 9-1).

Table 9-1: Summary Data Comparison Multi-Doc and Single-Doc

Property	Human (10 sent) Query-Relevant Multi-Doc	Human (10 sent) Generic Multi-Doc	Human No-Limit Generic Multi-Doc	Human => Extracted Single-Doc
Document (Set) Features				
# of doc sets (docs)	30	30	3	
# of docs	300	300	30	2250
Avg. sent/doc set	293	293	293	
Avg. sent/doc	29	29	29	26
Summary Features				
% of doc set length	3%	3%	12%	
% of doc length				20%
Includes 1st Sentence	13%	14%	25%	69%
Summary Composition				
Non consecutive sentences	84%	88%	48%	-
2 consecutive sentences	13%	10%	26%	-
3 consecutive sentences	2%	2%	16%	-
>= 4 consecutive sentences	1%	0%	10%	

Single document newswire summaries include the first sentence of the article approximately 70% of the time; human-produced 10 sentence multi-document summaries include approximately 1.3 first sentences of the 10 possible initial sentences per summary. This of course varies by type of summary desired (refer to Table 9-2). This data indicates that in our summarizer, lower weight should be given for the option of including the first sentence in multi-document summaries ($w_{\text{position-genre}}$ in Equation 3-3) as compared to single document summaries.

For 10 sentence summaries human summarizers tend to use non-consecutive single sentences 84% of the time (Table 9-1) as compared to 48% of the time for no limit summaries. The no-limit summaries also tended to have more first sentences of documents, an average of 25% of the available first sentences of articles (2.5 sentences used in the summary) as compared to 14%.

Table 9-2: Lead Sentences in Generic Human 10 Sentence Summaries

Summary Type (number of sets)	Number of Article Initial Sentences
Multiple Sources, Snapshot (8)	1.0
Same Source, Snapshot (5)	1.5
Unfolding Event – short timespan (7)	1.1
Unfolding Event – long timespan (8)	1.3
Comparison Events (2)	2.7
All Events (30)	1.3

Table 9-3: Exact Match Sentences Between Human 10 Sentence Summaries

Summary Type	H1 to H2	H1 to H3	H2 to H3
Generic	41 (14%)	41 (14%)	46 (15%)
Query-Relevant	59 (20%)	54 (18%)	65 (22%)

A coarse examination of the 10 sentence human multi-document summaries indicates that they did not have a huge overlap with each other (Table 9-3). The best human overlap in the query relevant case (which constrains the relevant sentences that can be picked) is 22% exact matches.

We expect a lot lower exact sentence overlap than that of single document summaries due to the fact that some data sets were designed to have a high degree of sentence redundancy between articles, providing a variety of closely related choices for sentence selection. Radev provides a detailed analysis of these types of sentences of the types of relationships that can occur in these sentence clusters [Radev 2000].

We expect a lot lower exact sentence overlap than that of single document summaries due to the fact that some data sets have a high degree of sentence redundancy between articles, providing a variety of closely related choices for sentence selection. Radev provides a detailed analysis of these types of sentences of the types of relationships that can occur in these sentence clusters [Radev 2000].

This information overlap is apparent upon manually reviewing the summaries - reflecting the fact that that there are often several suitable sentences for a single point. We can crudely approximate this information overlap if we examine human summarizer agreement for the subtopics within a summary.

First, let us examine the human agreement on the important subtopics for each set. There were an average of 16 subtopics for each set. For the most important subtopic of the provided subtopics, assessors had 56% agreement and for the three most important subtopics of the topic set, assessors had 62% agreement (Table 9-4).

Table 9-4: Subtopic Agreement in Human 10 Sentence Summaries

Clusters	Human Subtopic Agreement on the Most Important Subtopic	Human Subtopic Agreement on the Three Most Important Subtopics
All Events (30)	56%	62%

For scoring the individual sentences, assessors could choose as many subtopics as they wanted to assign to the sentence. Two people agreed on at least one subtopic in 47% of the sentences and all three agreed on one subtopic in 57% of the sentences. Ten percent of the sentences had no majority agreement as to subtopic. The overall score for agreement on subtopic selections is 60%.

For the three most representative articles of the document set, assessors had 42% agreement on the most representative article and 67% agreement on the articles selected as the three most representative articles. All articles were presented in their ordered time sequence of article appearance (although assessors were allowed to work with them in any order) and the majority of articles selected as the most representative articles were in the latter half of the data sets (Table 9-5), supporting our summarizer algorithm's use of an additional weighting for documents with a more recent time stamp.

However, there appears to be only a slight bias, if any, towards summary sentence selection from latter articles. From Table 9-6, we can calculate that the average article number selected is 5.9 for both generic and query relevant (5.5 would be the average article number if one summary sentence was selected from each of the ten articles which are labeled 1-10). Approximately 55% of the summary sentences from both generic and query relevant summaries come from the last five articles (Table 9-6) and approximately 68% of the chosen most representative articles are in the last five articles (71% for the most representative article).

Table 9-5: Distribution of Articles Selected as Most Representative
Number of times the article was selected by all human summarizer as most representative compared to the number of times the articles was in the top three most representative articles selected.

Articles		
Article Number	First Article	Top Three Articles
1	4	10
2	5	17
3	7	16
4	3	20
5	7	33
6	8	27
7	10	33
8	9	35
9	17	40
10	20	69

Table 9-6: Sentence Distribution from Articles for Human Summarizers

	Percent of Summary Sentences	
Article Number	Generic	Query Relevant
1	7.2	9.4
2	8.1	9.9
3	9.8	7.9
4	10.0	7.2
5	10.7	10.0
6	9.6	7.8
7	6.9	8.9
8	9.6	8.8
9	13.2	15.2
10	14.4	14.6

9.3 Evaluation of our MDS system

In 3.5, Equation 3-3, we presented our algorithm for multi-document summarization. In this section, we present the experiments we performed to evaluate the MDS system.

As previously mentioned, Spark-Jones & Galliers define two types of summary evaluations: (i) intrinsic, measuring a system's quality, and (ii) extrinsic, measuring a system's performance in a

given task [Sparck-Jones and Galliers 1996]. Automatically produced summaries by text extraction can often result in a reasonable summary. However, this summary may fall short of an *optimal* summary, i.e., a readable, useful, intelligible, and relevant appropriate length summary which meets the user's information seeking goals. Thus, extrinsic evaluations ultimately determine the utility of summaries such as that performed by SUMMAC [Hand 97, Mani et. al. 1998].

Our current evaluations are intrinsic - we evaluated how similar our summaries are to the "gold standard" described in the previous two sections. We also performed a brief one person evaluation of summary quality. The assessment was performed by a English major focusing on journalism.

In particular, we want to evaluate the effectiveness of the anti-redundancy MMR-MD technique. Our summarizer uses various weights[13]. We can compare our machine generated summaries to our human generated summaries for various combinations. As a start, let us compare the anti-redundancy measures and $\lambda = 0.3$ (AD-3) and $\lambda = 0.6$ (AD-6) to no redundancy elimination $\lambda = 1.0$ (AD-SDS) with weights for coverage, content and time sequencing set to zero - w_2, w_3, and w_4 in Equation 3-3. We use AD-SDS as one of our baselines. This baseline is essentially equivalent to a single document summary of the concatenation of all documents.

In our evaluations, we use some combination of these four baseline summaries:

1. centroid document single document summary (CD-SDS): This is the human selected centroid document - selected to be most representative of the set of articles. Our summarizer creates a single document summary.
2. centroid document lead sentence summary (CD-LSS): This is the lead sentences from the human selected centroid document - a fast method and lead sentences often perform well for the newswire single document summarization case. This summary – composed of the lead sentences, should be fluent and readable.
3. the first sentences of relevant documents (AD-FS): This is also a fast method and would effectively create a summary of some of the main points of the document set, since for newswire articles, 70% have a highly relevant first sentence [Goldstein et. al. 1999]. In our case, each set has ten documents and all are relevant to the topic. This summary can suffer from redundancy as motivated in Section 1.1.
4. document concatenation single document summary (AD-SDS): concatenate all the documents and perform single document summarization with no anti-redundancy measures.

We evaluated our summaries in three ways:

1. cosine similarity metric
2. comparison with human selected subtopics for the summary
3. human judgment

[13] These weights were determined manually through trial and error. These weights could be learned empirically, using a hill climbing mechanism, if we had a sufficient large set of hand-generated summaries to allow for the computation of different summary types.

9.3.1 Cosine Similarity Evaluation

Ideally, there would be an effective methodology of evaluating the content of a summary as discussed in Chapter 4. It seems unlikely that this can be effectively done without an understanding of natural language, in particular, how to group and score semantically equivalent pieces of information. However, as a first pass we use the cosine similarity metric (Section 4.3.1). We compute the cosine similarity between two sentences and, instead of using this as a redundancy penalty as it is used in Maximal Marginal Relevance (see Section 3.2) and in Radev's Cross Sentence Information Subsumption (CSIS) [Radev et. al. 2000], we use this to score the machine generated sentences with respect to the human generated ones.

Our scoring algorithm functions as follows:

1. Calculate a score for each summarizer generated sentence with respect to each human generated sentence using cosine similarity with term frequency.
2. Perform N passes (where N is the number of sentences in the output summary) through the system, one for each sentence in the output summary, removing the highest scoring sentence pair.
3. Compute a score for the summarizer generated summary by averaging the scores for the extracted sentence pairs.
4. Compute a final score for the summarizer generated summary by averaging over the number of human generated summaries.

For the cosine similarity metric, we compared our results to our baselines. The results are shown in Table 9-7.

Table 9-7: Summarizer Type Results: Similarity Score

Type	Baseline HS: Human Comparison	Baseline CD-SDS: Centroid Doc Single Doc Summary	AD-SDS: Concatenate Docs - Single Doc Summary	AD-6: MMR-MD $\lambda = 0.6$	AD-3: MMR-MD $\lambda = 0.3$
Query-relevant	0.37	0.29	0.29	0.31	0.28
Generic	0.33	0.29	0.31	0.31	0.29

We performed the two sample t-test to determine significance of our results.

For the query-relevant results, there was a significant difference ($p < 0.01$) between the human summaries and all other results. Excluding the human summaries, there were no other significant differences. For the generic results, there was no significant difference between the human summaries and AD-6 (MMR anti-redundancy with $\lambda = 0.6$), but there was ($p < 0.05$) between the human summaries and AD-3 ($\lambda = 0.3$) and CD-SDS.

This indicates that for the generic summaries, in which the task is less specified than in the query relevant case, that our machine generated anti-redundancy AD-6 summaries perform similarly to the human summaries. It also may indicate that too diverse of a summary $\lambda = 0.3$ is less ideal,

which is supported by the lower scores for λ = 0.3 in the exact match sentences evaluation (Table 9-8).

Table 9-8: Summarizer Type Results: Sentences Exact Match

Type	Baseline HS: Human Comparison	Baseline CD-SDS: Centroid Doc Single Doc Summary	AD-SDS: Concatenate Docs - Single Doc Summary	AD-6: MMR-MD λ= 0.6	AD-3: MMR-MD λ= 0.3
Query-relevant	2.0	0.9	1.2	1.3	1.0
Generic	1.4	0.9	1.4	1.4	1.0

Although, there is not much difference in the scores between methods - there is clear evidence of redundant information in the individual summaries as shown in Figure 1-14, Figure 1-15, Figure 1-16 and Figure 1-17. The cosine similarity scoring method does not penalize for redundancy. In fact, upon examination, some of the reason that the baseline and λ =0.6 scored so well seems due to the fact that the primary contributions to the scores comes from summary sentences which are exact matches, partial matches or c contain terms in the sentences that are very similar to the human summary sentences (Table 9-8).

These summaries clearly show the need for extraction and evaluation of information at a phrasal level, which would require an approach similar to that of Radev's subsumption evaluation method [Radev et. al. 2000], the extensions to the automatic metric ROUGE, ROUGE-BE (Section 4.4.3) [Hovy et. al. 2005] or the Pyramid Method [Passonneau and Nenkova 2005]. It also shows the need for information content analysis, such as cross document co-reference resolution to match variations in spelling, e.g., Mohammed and Muhammed, or written language spelling variants such as “tire” and “tyre”.

From an examination of the scores on individual data sets, sometimes the λ=0.3 summaries perform the best and at other times the λ=0.6 summaries are more effective. We hypothesize that this is partly due to the fact that certain techniques may be particularly effective for retrieving certain types of clusters. Our clusters consist of various types on information over various spans of time and geographic locations – including clusters with articles covering an event on a single day, articles with updates on a particular event as well as articles about similar events in multiple locations. We need to explore value of λ to determine which values (including others besides 0.3 and 0.6) are most effective in what situations. In addition, we need to examine the effect of the queries on the summaries and what type of queries are effective for what types of summaries. Furthermore, we need to identify the differences and similarities between the human summaries to gain a clearer idea on how to set and fine tune the weighting parameters in our system.

9.3.2 Subtopics Evaluation

For the subtopic evaluation, we use the combined subtopics of a summary for the calculation. From our data set collection, each sentence has one or more associated subtopics (using majority voting). For each human summary we determine the associated subtopic set. Multiple subtopics for a summary are ignored, so if a summary has two sentences that are labeled subtopic 4, then

subtopic 4 is only listed once in the subtopics for the summary. The final set of subtopics for the comparison is the union of all three human summaries subtopic sets. From this determination, we can compare the coverage of the subtopics of our machine generated summaries to that of the general summaries.

The results of the summary comparison by subtopic is shown in Table 9-9. Although it appears that the λ=0.6 summaries and the λ=0.3 summaries are introducing sentences from relevant subtopics in the summaries, this is only significant in the case of the query relevant subtopics λ=0.3 compared to the baseline ($p < 0.25$). However, we can see in both query relevant and the generic cases, the anti-redundancy measures are overall increasing the number of subtopic matches between the summaries (Table 9-10) with more subtopics covered at λ=0.3 as compared to λ=1.0 – 18 for query relevant and 16 for generic summaries. This indicates that indeed our anti-redundancy metrics are resulting in better subtopic coverage and that we need to be able to evaluate summaries more effectively than the cosine similarity metric by developing multi-document summarization scoring methods to distinguish finer grades of summary quality. ROUGE-BE might be a suitable metric for such a test – this is left for future work.

Table 9-9: Summarizer Type Results: Subtopic Coverage Score – system coverage compared to the coverage of the combination of human summaries

Type	AD-SDS: Concatenate Docs - Single Doc Summary	AD-6: MMR-MD λ= 0.6	AD-3: MMR-MD λ= 0.3
Query-relevant	0.48	0.55	0.59
Generic	0.52	0.56	0.58

Table 9-10: Subtopic Score Change with λ value.
As compared to the CD-SDS baseline summary (concatenation of documents, λ= 1.0)

Score	AD-6: MMR-MD λ= 0.6	AD-3: MMR-MD λ= 0.3
Query-relevant Summaries		
Increase	13	18
Decrease	4	4
No Change	13	8
Generic Summaries		
Increase	11	16
Decrease	6	8
No-change	13	6

9.3.3 Human Judgment Summary Evaluation

To further determine the overall quality of our summaries, for the generic summaries, we asked one human evaluator to rank eight summaries (rank 1 is best) from our possible summaries:

- H1, H2, H3: The three human summaries
- CD-LSS: (baseline) The lead sentences from the human selected centroid document. (baseline) This summary was chosen as it guaranteed consecutive sentences.
- AD-FS: (baseline) The first sentence from each document
- AD-SDS: (baseline) Concatenate all documents and compute a single document summary
- AD-6: MMR-MD with $\lambda = 0.6$
- AD-3: MMR-MD with $\lambda = 0.3$

All summaries were arranged in date order and document order (with the exception of the human summaries). All sentences from the earliest document were output first in the order which they appeared in the document.

For each summary, we also asked the users a series of questions (partly constructed based on the DUC-2001 evaluation) [DUC]. We used the ranking:

1. All (score 1)
2. Most (score 2)
3. Some (score 3)
4. Hardly Any (score 4)
5. None (score 5)

The evaluators were to select word choices (rankings) for the following six questions:

1. _____ of the sentences belong with the surrounding sentences
2. _____ of the summary is well organized, i.e. the content is expressed and arranged in an effective way
3. _____ of the sentences are related to the main topic(s)
4. _____ of the sentences should be included in a summary of the main topic(s)
5. _____ of the sentences include redundant or repetitive information with the other sentences
6. _____ of the sentences are highly informative

The evaluator was given the sentence titles as the topics for the set. The summaries were presented to the evaluator numbered in a random order. The evaluator ranked all 30 sets.

Table 9-11 shows the results of our human evaluation. In particular, for question 5, which measure anti-redundancy, AD-6 is significantly different from AD-SDS ($p < 0.1$) and AD-3 is significantly different from AD-SDS ($p < .0005$). This demonstrates that the anti-redundancy measures are assisting. Question 5 also shows that the human summaries and centroid document lead sentence summary (CD-LSS) performs better for anti-redundancy measures and that the machine generated summaries need to improve to approach these results. In fact, CD-LSS scores very close to the human summaries. In question 1, there is only a significant difference between H1 and CD-LSS ($p < 0.01$), in question 2 there is no significant difference, in question 3 CD-LSS is significantly different from H1 ($p < 0.025$), and from H2 and H3 ($p < 0.001$). For

question 4, CD-LSS is significantly different from H1 ($p < 0.1$), H2 ($p < .05$) and H3 ($p < 0.01$). In question 5 there is no significant difference and in question 6, CD-LSS is significantly different from all human summaries ($p < 0.025$) or better).

Table 9-11: Results of Human Evaluation of Generic Multi-Document Summaries

	H1	H2	H3	AD-FS	CD-LSS	AD-SDS	AD-6	AD-3
Ranking (1 best)	3.6	3.1	3.4	6.3	**3.2**	6.0	4.9	5.4
Q1 Cohesion (1 best)	2.1	1.8	2.0	3.4	**1.5**	3.1	2.8	2.9
Q2 Organization (1 best)	2.1	1.9	2.0	3.7	**1.9**	3.4	3.2	3.0
Q3 Topic Related (1 best)	2.1	1.9	2.1	2.0	2.5	**2.0**	2.3	2.2
Q4 Inclusion (1 best)	2.5	2.5	2.3	3.0	2.9	2.9	**2.7**	2.9
Q5 Redundancy (5 best)	4.1	4.2	4.2	2.6	**4.3**	2.6	3.0	3.4
Q6: Informative (1 best)	2.4	2.3	2.4	2.6	2.9	2.9	**2.6**	2.7

Further research needs to be conducted to determine whether the system can locate these human selected centroid documents and if the system fails, what is the resultant degradation in summary quality. It is worth noting (from the rankings), that the AD-6 (anti-redundancy measure summary) is ranked the highest after the human produced summaries, which includes CD-LSS since it consists of text extracts in a human selected order.

Chapter 10 Conclusions and Future Work

> **"And so we come to the end of our journey. We have spoken of many and diverse things. But through all their differences ran one unifying thread, one invariant under transformation, the structure of the language used to describe and define them. Now our portion of the tapestry is woven. I hope the design is clear, the pattern appealing, and I hope this weaving is not the end."**
>
> Professor Harry L. Weinberg, Levels of Knowing and Existence: Studies in General Semantics 1959.

Summarization is a very important area of research due to the vast amounts of electronic information available on the internet and through digital libraries. As more countries produce information that is available for perusal, multilingual summarization becomes increasingly important as the "answers" to user's information seeking needs might be part of a document in another language. As an example of the information explosion, Wikipedia, which was created in 2001, currently (2008) has over 684 million visitors yearly, 75,000 active contributes, working on more than 10,000,000 articles in more than 250 languages. In 2001, Wikipedia had entries in 16 languages, in 2003, it had grown to 54 languages [Wikipedia].

In this thesis, we have only focused on a small portion of the summarization information space. We have considered only summaries in one language, English, and only summaries for a few of the many available genres and only multi-document summarization techniques within one genre. A truly comprehensive summarization system requires summarizing across multiple genres when necessary and across multiple languages as well.

As digital communications contain to grow, we see variants of English forming, including abbreviations that are prevalent in text messages and chat, as well as the use of emoticons to convey the feelings of the writer. At some point, summarization systems need to address these variants and nuances as well.

In the following sections we summarize the contribution of this thesis and describe some ideas for future work.

10.1 Summary of Contributions

There are many contributions that have been made in the course of our exploration of summarization. Some are summarized in the list below.

- **User's Information Seeking Goals - Goal Focused Summarization**
  We motivated the topic of Goal Focused Summarization based on Xie's framework for information seeking goals [Xie 2000]. We demonstrated the different types of summaries that can be produced using these principles.

- **Summary Length**

- We suggested that there might be an appropriate summary length for different genre oriented summaries based on empirical evidence.
- We proposed various lengths for the composition of goal focused summaries in various genres, including movie reviews, product reviews, product press releases, bibliographies, interviews, new articles and editorials.
- We performed a user study for email summary length using the Enron email corpus, which indicated that the subject line and a one sentence summary from the text body may be an effective summary for a recipient of the email to determine some course of action and priority for the email.

- **Evaluation**
  - We suggested other criteria for summary evaluations based on Endres-Niggemeyer's principles for creating summaries [Endres-Niggemeyer 1998].
  - We demonstrated in the Multilingual Summarization Evaluation MSE 2006 [Goldstein et. al. 2006b] that evaluation of summaries might possibly lead to inflated summary scores due to the fact that assessors are evaluating a summary without knowledge of the actual documents.
  - We showed that humans produce poor summaries as judged by other humans, even for formal evaluations. This motivates the requirement for well trained summarizers or professional summarizers for formal evaluations.

- **Data Sets**: We created four data sets for use by the community:
  - Single document summarization: Newswire documents marked by three assessors with relevant, partially relevant and not-relevant. A three sentence summary was created for each document.
  - Multi-document newswire summarization: Three human assessors created 10 sentence extract multi-document summaries for 30 clusters consisting of 10 topically related articles. The selected sentences were marked in *rank order* – the sentence judged most important first, followed by the second most important, etc. In addition, each 10 sentence extract was also ordered by the creator in a most *readable order* based on combining the sentence extracts in most the most fluent, intelligible and coherent manner. For each topic cluster, subtopics relating to the articles were created. The assessors assigned subtopics (one or more) to each sentence in each article. Single document 3 sentence extract summaries were also created for each article. The topic clusters were designed towards exploration of various types of summarization, including summarization over short spans of time (high redundancy), longer periods of time, as well as related clusters for the exploration of update summarization.
  - Genre Oriented Goal-focused Summarization: Goal-focused sentence extract summaries were designed and created by three humans for 7 genres. Genre-oriented goal-focused summaries varied from 5-9 sentences in length depending on the genre and genre-topic. Sentence extract summaries were created for each genre oriented goal-focused summary. These summaries were provided both in rank order and the most "readable" order.
  - Genre Identification: We created the largest known tagged web collection of genre related articles. This consisted of approximately 1000 tagged items for 7 genres. Within the 7 genres, there were subtopics, 8 for biography web pages and 20 for store products.

We combined this with another CMU collection of 9 genres, which had 1000 or more documents per genre, allowing studies on 42 distinct genre and genre-topic classes.

- **Genre Identification**
  - We showed that in most cases an average genre identification precision and recall of over 85% can be achieved, especially if the number of documents used for training and tested is over 100 [Goldstein et. al. 2007].
  - The "out of the box" classifier Random Forests performed better than SVM-light in terms of: (1) overall performance, (2) the ability to correctly classify topically confusable sets, and (3) the amount of time required for training.
  - Random Forests performs very well with our baseline set of features 66, which has very limited content and topically related items.
  - The addition of "random files" collected from the internet does not hugely impact genre identification performance.

- **Genre Oriented Summarization**
  - We designed algorithms to create goal-focused summaries for genre oriented summarization. We show that in most cases, these algorithms outperformed the baselines [Goldstein et. al. 2007].
  - We demonstrated that with the knowledge of the genre of the document, genre specific information can be extracted to create summaries that ought to address a user's information seeking goals.

- **Email Summarization**
  - Speech Acts: We extended speech acts into the email domain to form email acts and created a framework of 12 major categories, 30 sub categories based on the dialog research for speech acts. We demonstrated that for a small subset of tagged data that we could achieve both high inter-annotator agreement and reasonable performance for the 5 of the major categories [Goldstein and Sabin 2006]. We showed that verb classes alone do not perform as well in distinguishing the email acts as a set of email specific features. The email specific features combined with the verbs perform slightly better than the email feature alone.
  - Subject Line and Summary Sentence. We analyzed human tagging of a subset of the Enron corpus for subject line and selection of one summary sentence. We showed that the subject line combined with the summary sentence forms a good indicative summary for the task of taking some action in regards to the email.

- **Multi-Document Summarization**
  - We extended the anti-redundancy maximal marginal relevance framework MMR [Carbonell and Goldstein 1998] for the case of multi-document summarization MMR-MD [Goldstein et. al. 2000b].
  - A study of one human evaluator examining multi-document summaries from 30 clusters, showed rated the lead sentences from the centroid document of the cluster as the best automated summary. Of the two summaries formed using MMR-MD, summaries using the anti-redundancy measure $\lambda = 0.6$ scored better than $\lambda = 0.3$.

- Using the MMR-MD anti-redundancy measures $\lambda = 0.6$ and $\lambda = 0.6$, the resultant sentence extract summaries were able to include more sub-topic coverage of the human generated summaries as compared to the baseline of no anti-redundancy measures.
- Using the cosine similarity metric for evaluation MMR-MD $\lambda = 0.6$ and the two sample t-test, there was no significant difference between these summaries and the human generic summaries. There was, however, a significant difference for $\lambda = 0.6$ ($p < 0.5$). For the human generated query relevant summaries, there was a significant difference ($p < 0.01$) compared to summaries created with both anti-redundancy measures.

10.2 Future Work

- For *Summary Creation,* future research might
  - Develop algorithms to address Endres-Niggemeyer's analysis of methods utilized by professional summarizations (Section 2.2.2).
  - Explore what are optimal length summaries for particular genres and why?
- For *Summary Characteristics,* future research might
  - Explore sentence ordering. Our multi-document and genre oriented data set have both rank sentence orderings and "most readable" sentence orderings that could be used for this purpose.
  - Analyze topical coverage and overlap in a human and machine generated summaries. Each sentence in our multi-document evaluation set has assignments to subtopics for the corresponding news cluster.
- For *Evaluation*, future research might
  - Develop metrics to address Endres-Niggemeyer's summary assessment for interestingness and innovation as well as author intent in Section 2.2.2.
  - Investigate techniques to determine whether the summary has successfully employed the meaning reduction strategies.
  - Perform extrinsic studies to determine whether users have a preferences for genre oriented summarization, and to evaluate our summaries.
- For *Genre Identification*, future research might
  - Test the new multi-class support vector machines and compare it to Random Forests.
  - Determine what features are contributing to the scores and eliminate unnecessary features.
  - Evaluate errors and determine methods to improve results including the development of additional features to distinguish genres.
- For *Genre Oriented Summaries*, future research might
  - Investigate the effects of misclassification on summary output using the ground truth data.
  - Determine better algorithms for the interview genres.
  - Conduct user studies to determine people's responses to genre-oriented goal-focused summaries.
- For *Email Summarization*, future research might
  - Investigate the ability to categorize the 30 categories of communicative intents.
  - Investigate the ability to determine whether an email thread is about a meeting.

    - Study how to form an effective email summary of a thread.
- For *Multi-document Summarization*, future research might
    - Explore genre oriented multi-document summarization, including multiple viewpoints on an event or multiple reviewers' opinions.
    - Explore multi-document summarization within a single genre other than newswire.
    - Explore multi-document summarization across multiple genres.
    - Investigate summarization by viewpoints, types of comments, a particular view or a facet of the information, such as the cinematography facet in movie reviews.
    - Investigate summarization by timeline – including anchoring events on the timeline.
    - Address summarizing an area of knowledge. For example in the topic of astronomy, one could summarize theories about the creation of the universe as well as instruments used for astronomy, e.g., the telescope. This type of summarization would require clustering the various types of information by these subcategories and then creating a single or multi-genre summary (if there were multiple genres, e.g., newswire and scientific articles).
    - Study update summarization.
    - Investigate how to create comparative summaries.
    - Analyze human selected most readable summaries.
    - Analyze human selected generic summaries for single document summarization. Compare to our previous results and analyze if there is any useful information for the resultant multi-document summaries.

10.3 Conclusion

As we conclude this dissertation, we present a quote for Inderjeet Mani from his well cited book on automatic summarization [Mani 2001]:

> **"Summarization offers the promise of helping humans harness the vast information resources of the future in a more efficient manner. Before this promise fully materializes, there is much research in terms of both theory and practice, that must be carried out. The success of the effort will depend, in part, on a precise formulation of the different summarization requirements for various practical tasks and a clear picture as to what counts as an effective solution."**

In the past 7 years since this statement was written, the amount of information available has taken one giant step forward in the quantities of information available – currently 26.8 billion webpages are estimated to be indexed in October 2008. From 2000 to 2007 the number of website hosts have been estimated to have grown from 7 million to 433 million. The Internet World Stats has counted over 1 billion Internet users in 2007. The average web page size has tripled in the past 5 years from 94K to 312K and increased 22 times since 1995 [Domenech et. al. 2007, Finn & Betcher 2008]. In the same five years, the number of objects in the average web page nearly doubled from 26 to 50.

During this period of growth, the summarization field has only taken baby steps even though workshops are held yearly and many papers are published in top tier conferences. In terms of formal evaluations open to the summarization community, the majority have been focused on newswire and are only recently, in the past two years, starting on multi-modal summaries as well as blogs [DUC, TAC, NTCIR]. Summarization for various practical tasks has been addressed more within industry than within the research communities and very little has been published on what counts as an effective solution for a summary.

Our research work, through the production of goal-focused summaries for particular user tasks, including various web genres, email, and multi-document newswire articles, has attempted to address some of the specific summaries that might be of interest to users.

However, even with all the progress in the past decade, current summarization research is only the tip of the iceberg. Much research needs to occur before we can define and create the types of summarization systems that effectively meet the user's information seeking goals. It is our hope that this thesis will inspire other researchers to continue in these efforts and further advance the field of summarization.

Bibliography

[ACL] ACL Summarization Workshops:

[Allan et. al. 1998] J. Allan, J. Carbonell, G. Doddington, J. Yamron, J., and Y. Yang. Topic detection and tracking pilot study: Final Report. *In Proceedings of the DARPA Broadcast News Transcription and Understanding Workhop*, 1998.

[Allen and Core 1997]. J. Allen and M. Core. *Draft of DAMSL: Dialog Act Markup in Several Layers*, 1997.

[Amitay and Paris 2000] Automatically Summarising Web Sites: Is There a Way Around It? In *Proceedings of the 9th International Conference on Information and Knowledge Management*, pages 173-179, New York, NY USA 2000.

[Aone et. al 1997] C. Aone, M.E. Okurowski, J. Gorlinksy, and B. Larsen, A scalable summarization system using robust NLP, In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization* pages 67-73.

[Austin 1962] J.L. Austin, *How to do things with words*. Harvard University Press, Boston, MA, 1962.

[Avrahami et. al. 2006] T. T. Avrahami, L. Yau, L. Si, and J. Callan. The FedLemur project: Federated search in the real world. In *Journal of the American Society for Information Science and Technology*, 57(3) (pp.347-358).

[Bach and Harnish 1979], K. Bach K. and R.M. Harnish, *Linguistic Communication and Speech Acts*, The MIT Press, Cambridge, MA, 1979.

[Baldwin and Morton] Breck Baldwin and T.S. Morton T. Dynamic coreference-based summarization. *In Proceedings of the Third Confrence on Empirical Methods in Natural Language Processing (EMNLP-3),* Granada, Spain June 1998.

Banko, M., Mittal, V., Kantrowitz, M, and Goldstein, J. Generating extraction based summaries from handwritten summaries by aligning text spans. *In Proceedings of PACLING-99*, Waterloo, Ontario, July 1999.

[Ballmer and Brennenstuhl 1981] T.T. Ballmer and W. Brennenstuhl, *Speech Act Classification. A Study in the Lexical Analysis of English Speech Acts,* Springer-Verlag, Berlin, 1981.

[Baron 2000] N.S. Baron, *Alphabet to Email: How Written English Evolved and Where It's Heading,* Routledge, London, 2000.

[Baron 2003] N.S. Baron, Why Email Looks Like Speech in *New Media Language*, Aitchison, J. and Lewis, D. (ed.), Routledge, London, 2003.

[Barzilay and Elhadad 1997] R. Barzilay and M. Elhadad, Using Lexical Chains for Text Summarization. *In Proceedings of the ACL'97/EACL-97 Workshop on Intelligent Scalable Text Summarization*. Madrid, Spain 1997.

[Barzilay et. al 2002] Regina Barzilay, Noemie Elhadad and Kathleen McKeown. Inferring Strategies for Sentence Ordering in Multidocument News Summarization. In Journal of Artifical Intellgience Research 17 35-55 2002.

[Barzilay and Lee 2004] Regina Barzilay and Lillian Lee. Catching the Drift. Probabilistic Content Models, with Applications to Generation and Summarization. In *Proceedings of HLT-NAACL*, 113-120. 2004.

[Barzilay and Lapata 2005] Regina Barzilay and Mirella Lapata. Modeling Local Coherence: An Entity-Based Approach. In ACL 2005.

[Biber 1988] D. Biber, *Variation across speech and writing*, Cambridge University Press, Cambridge, UK. 1988.

[Boguraev and Kennedy 1997] B. Boguraev and C. Kennedy. Salience based content characterization of text documents. *In Proceedings of the ACL'97/EACL-97 Workshop on Intelligent Scalable Text Summarization*. Madrid, Spain 1997.

[Bollegala et. al. 2005]. D. Bollegala, N. Okazaki and M. Ishizuka. A Machine Learning Approach to Sentence Ordering for Multi-document Summarization and its Evaluation. In *Proceedings of IJCNLP* 2005.

[Brants 2000] Thorsten Brants, TnT: A Statistical Part-Of-Speech Tagger, In *Proceedings of ANLP-2000*, Seattle, WA, 2000.

[Breiman 2001] L. Breiman, *Random Forests*, U.C. Berkeley Technical Report for Version 3, Berkley, CA, 2001.

[Breiman 2004] L. Breiman, Consistency for a Simple Model of Random Forests, U.C. Berkeley Technical Report 670, Berkeley, CA 2004.

[Buckley 1985] C. Buckley, Implementation of the SMART Information Retrieval System. Technical Report TR 85-686, Cornell University, 1985.

[Byrd 2005] Byrd, P., "Irregular verbs," http://www2.gsu.edu/~wwwesl/egw/verbs.htm 2005.

[Carbonell and Goldstein 1998] Jaime G. Carbonell and Jade Goldstein, The Use of MMR, Diversity-based Reranking for Reordering documents and Producing summaries. In Proce*edings of SIGIR-98*, Melbourne, Australia, August 1998.

[Carvalho and Cohen 2006] Vitor R. Carvalho. and William W. Cohen, Improving Email Speech Act Analysis via N-Gram Selection, *In Proceedings of Human Language Technology Conference North American Chapter of the Association for Computational Linguistics, ACTS Workshop*, New York City, 2006.

[Cohen 1960]. J. Cohen, A Coefficient of Agreement for Nominal Scales, *Education and Psychological Measurement*, 20, 1960, 37-46.

[Cohen et. al. 2004] William W. Cohen, Vitor R. Carvalho, and Tom M. Mitchell, ,"Learning to Classify Email into 'Speech Acts,'" *Proceedings of Empirical Methods in Natural Language Processing (EMNLP2004*), 2004.

[Collot and Belmore 1996] M. Collot and N. Belmore, Electronic Language: A New Variety of English, *Computer-Mediated Communication*, Herring, S. C. (ed.), John Benjamins, Amsterdam, 1996.

[Conroy et. al 2006a]. John M. Conroy, Judith D. Schlesinger, Dianne P. O'Leary, and Jade Goldstein, Back to Basics: CLASSY 2006. In *Proceedings of DUC 06*. 2006.

[Conroy et. al 2006b] CLASSY Arabic and English Multi-Document Summarization. Report for the Multilingual Summarization Evaluation 2006.

[Corston-Oliver et. al. 2004] S. Corston-Oliver, E. Ringger, E., M. Gamon and R. Campbell. Task-focused Summarization of Email. In *Proceedings of Text Summarization Branches Out Workshop, ACL 2004*, 2004.

[Crammer and Singer 2007]. K. Crammer and Y. Singer. On the Algorithm Implementation of Multi-class SVMs. In *Journal of Machine Learning Results v. 2*. pp. 265-292, 2001.

[Crowston and Kwasnik] K. Crowston, and B. Kwasnik, Can document-genre metadata improve large digital collections. Draft submitted to *Library Trends,* Sepmber 21, 2003.

[Crystal 2001] D. Crystal, *Language and the Internet.* Cambridge University Press, Cambridge, UK, 2001.

[Dang 2006] Hoa Dang. Document Understanding Conference DUC 2006 slides. Presented at the Document Understanding Workshop at *HLT-NAACL 2006*. Brooklyn, New York June 2006. http://duc.nist.gov/pubs.html#2006.

[Dang and Lin 2007] Hoa Trang Dang and Jimmy Lin. Different Structures for Evaluating Answers to Complex Questions: Pyramids Won't Topple, and Neither will Human Assessors. *In Proceedings of Association of Computational Lingustics*, pages 768-775, Prague, Czech Republic, June 2007

[Dewdney et. al. 2001] N. Dewdney, C. VanEss-Dykema, and R. McMillan, The form is the substance: Classification of genres in text. In *ACL Workshop on Human Language Technology and Knowledge Management*, 2001.

[Diesner and Carley 2005] J. Diesner and K. M. Carley, Exploration of Communication Networks from the Enron Email Corpus, In *Proceedings of Workshop on Link Analysis, Counterterrorism, and Seciurity*, Newport Beach, CA, April, 2005.

[Donway et. al. 2000] R. Donway, K. Drummey and L. Mather. A Comparison of Rankings produced by Summarization Evaluation Measures", *In NAACL-ANLP Workshop on Automatic Summarization* 2000.

[Dorr and Jones 1996] Bonnie Dorr and D. Jones. Acquisition of semantic lexicons: Using Word Sense Disambiguation to Improve Precision," *Proceedings of the SIGLEX Workshop on Breadth and Depth of Semantic Lexicons*, Santa Cruz, CA. 1996.

[DUC] http://duc.nist.gov/

[DUC-QUALITY 2003] http://duc.nist.gov/duc2003/quality.html

[DUC-QUALITY 2005] http://duc.nist.gov/duc2005/quality-questions.txt

[DUC-ROADMAP] B. Baldwin, R. Donaway, E. Hovy, E. Liddy, I. Mani, D. Marcu, K. McKeown, V. Mittal, M. Moens, D. Radev, K. Sparck-Jones, B. Sundheim, S. Teufel, R. Weischedel and M. White. An Evaluation Road Map for Summarization Research. http://www-nlpir.nist.gov/projects/duc/roadmap.html

[Endres-Niggemeyer 1998] Brigitte Endres-Niggemeyer. *Summarizing Information.* Springer-Verlag 1998.

[Erikson 2000] T. Erikson, "Making Sense of Computer-Mediated Communication (CMC): Conversations as Genres," *Proceedings of Hawaiian International Conference on System Services (HICSS2000)*., 2000.

[Feldbaum 1998] C. Felbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA 1998.

[FERC] FERC online elibrary, docket e103-180. *http://elibrary.ferc.gov/idmws/search/fercgensearch.asp*

[Finn and Kushmerick 2003] A. Finn and N. Kushmerick. Learning to classify documents according to genre. In IJCAI-03 Workshop on Computation Approaches to Style Analysis and Synthesis. 2003.

[Gospodnetic and Hatcher] Otis Gospodnetic and Erik Hatcher. *Lucene in Action*. Manning Publications 2004.

[GALE] http://www.darpa.mil/ipto/programs/gale/gale.asp

[Goldstein et. al. 1999] J. Goldstein, M. Kantrowitz, V.O. Mittal, and J. Carbonell, Summarizing Text Documents: Sentence Selection and Evaluation Metrics. In *Proceedings of SIGIR-99*, Berkeley, CA, August 1999.

[Goldstein et. al 2000a] Jade Goldstein, Vibhu Mittal, Mark Kantrowitz and Jaime Carbonell, Multi-Document Summarization by Sentence Extraction In *ANLP/NAACL Workshop*. Seattle, WA April 2000.

[Goldstein et. al. 2000b] Jade Goldstein, Vibhu O. Mittal, Jaime G. Carbonell, James P. Callan, Creating and Evaluating Multi-Document Sentence Extract Summaries. In *Proceedings of the Conference on Information and Knowledge Management CIKM 2000,* pages 165-172, McLean, Virginia November 2000.

[Goldstein and Sabin 2006] Jade Goldstein and Roberta E. Sabin, Using Speech Acts to Categorize Email and Identify Email Genres. In *Hawaii International Conference on System Sciences* (HCSS 2006), Hawaii January 2006.

[Goldstein et. al. 2006a] Jade Goldstein, Andrew Kwasinski, Paul Kingsbury, Roberta Evans Sabin and Albert McDowell. Annotating Subsets of the Enron Email Corpus. Presented at the Conference on Email and Anti-Spam (CEAS) July 2006, Mountain View, CA.

[Goldstein et. al. 2006b] Jade Goldstein, Paul Chase and Lucy Vanderwende Multilingual Summarization Evaluation 2006. Presented at the Workshop on Task-Focused Summarization and Question Answering at ACL 2006, Sydney, Australia.

[Goldstein et. al. 2007] Jade Goldstein, Gary M. Ciany, Jaime G. Carbonell. Genre Identification and Goal-focused Summarization. In *Proceedings of the Conference on Information and Knowledge Management CIKM 2007*, pages 889-892, Lisbon, Portugal November 2007.

[Hand 1997] T.F. Hand. A proposal for task-based evaluation of text summarization systems. *In ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization,* pages 31-36, Madrid, Spain, July 1997.

[Harabagiu 2001] S. Harabagiu, R. Bunescu, and S. Maiorano, Text and Knowledge Mining for Coreference Resolution. *Proc. 2nd Meeting of the North America Chapter of the Association for Computational Linguistics (NAACL-2001)*, 2001, 55–62.

[Harman and Over 2004] Donna Harman and Paul Over. The Development and Evolution of TREC and DUC. In *Proceedings of the ACL-04 Workshop: Text Summarization Branches Out*, pages 10-17 Barcelona, Spain 2004.

[Hearst 1997] M.A. Texttiling: Segmenting text into multi-paragraph subtopic passages. In *Computational Linguistics*, 23(1):33-65, March 1997.

[Hoskinson 2005] Andy Hoskinson. Creating the Ultimate Research Assistance. In *IT Systems Perspectives* November 2005.

[Hori et. al 2002] Chiori Hori, Sadaoki, Furui, Rob Malkin, Hua Yu and Alex Waibel. Automatic Speech Summarization Applied to English Broadcast News Speech. In Proceedings of HLT, San Diego, CA 2002.

[Hori et. al. 2003] Chiori Hori, T. Hori and S. Furui, "Evaluation methods for Automatic Speech Summarization," In EUROSPEECH 2003, Geneva, Switzerland 2003.

[Hovy and Lin 1997] E. Hovy, E. and C.-Y. Lin, Automated text summarization in SUMMARIST. *In ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization,* pages 18-24, Madrid, Spain, July 1997.

[Hovy et. al. 2005] Evaluation DUC 2005 Using Basic Elements. In Proceedings of the Fifth Document Understanding Conference (DUC), Vancouver, Canada 2005.

[Hovy et. al. 2006] Eduard Hovy, Chin-Yew Lin, Liang Zhou and Junichi Fukumoto. Automated Summarization Evaluation with Basic Elements. Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC). Genoa Italy 2006.

[Hu and Bing 2004] M. Hu and L. Bing, Mining and Summarizing Customer Reviews. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004), Seattle, WA Aug. 22-25, 2004.

[IRESEARCH] http://www.iresearch-reporter.com/

[Jindal and Liu 2006] N. Jindal and B. Liu. Identifying Comparative Sentences in Text Documents. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval (SIGIR-06*) Seattle August 2006.

Jing, H., Barzilay, R., McKeown, K., and Elhadad, M. Summarization Evaluation Methods Experiments and Analysis. In AAAI Intelligent Text Summarization Workshop, pages 60-68, Stanford, CA March 1998.

[Joachims 1998] T. Joachims, Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *Proceedings of the European Conference on Machine Learning (ECML)*, 1998.

[Joshi and Rose 2007] Mahesh Joshi and Carolyn Penstein Rose. Using Transactivity in Conversation for Summarization of Educational Dialogue. In *SlaTe Workshop on Speech and Language Technology in Education*. Farmington, Pennsylvania October 2007.

[Jurafsky et. al. 1997] D. Jurafsky, E. Shriberg, and D. Biasca, Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual, Draft 13. *Institute of Cog. Sc. Tech Report 97-02*. University of Colorado, Boulder CO, 1997.

[Kalgren and Cutting 1994] J. Kalgren and D. Cutting,. Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. In *Proceedings of the 15th International Conference on Computational Linguistics* (COLING 1994) (Kyoto, Japan, 1994), 1071-1075.

[Kazantseva and Szpakowicz 2006] Anna Kazantseva and Stan Szpakowicz. Challenges in Evaluating Summaries of Short Stories. In *Proceedings of the Workshop on Task-Focused Summarization and Question Answering,* at ACL 2006, pages 8-15, Sydney, Australia July 2006.

[Kessler et. al. 1997]. B. Kessler, G. Nunberg, and H. Schutze,, Automatic Detection of Text Genre. In Proceedings of the 35th Annual Meeting of the ACL and 8th Conference of the EACL (Somerset, NJ 1997) 32-38.

[Klavans and Shaw 1995] J.L. Klavans and J. Shaw, Lexical Semantics in Summarization. *In Proceedings of the First Annual Workshop of the IFIP Working Group for NLP and KR*, Nantes, France, April 1995.

[Klimit and Yang 2004] B. Klimit and Y. Yang, The Enron Corpus: A New Dataset for Email Classification Research, In *Proceedings of the European Conference on Machine Learning* (*ECML),* 2004.

[Kupiec et. al 1995] J.M. Kupiec, J. Pedersen, and F. Chen. A Trainable Document Summarizer. *In Proceedings of the 18th Annual International ACM/SIGIR Conference on Research and Development in IR*, pages 68073, Seattle, WA, July 1995.

[Kwasnik and Crowston 2005] B. Kwasnik and K. Crowston, Genres of Digital Documents Introduction to the *Special Issue of Information, Technology & People,* 18 (2), 76-88 2005.

[Langridge 1989] D.W. Langridge. *Subject Analysis: Principles and Practices*. London: Bowker-Saur 1989.

[Lapata 2003] Mirella Lapata. Probabilistic text structuring. Experiments with Sentence Ordering. In *Proceedings of the ACL 2003*, 545-552.

[Lavie et. al. 2004] Alon Lavie, Kenji Sagae and Shyamsundar Jayaraman. The Significance of Recall in Automatic Metrics for MT Evaluation. In *Proceedings of the 6th Conference of the Assocation for Machine Translation in the Americas (AMTA-2004),* Washington, D.C., September 2004.

[Lavie and Agarwal 2007] Alon Lavie and Abhaya Agarwal. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation at the 45th Meeting of the Association for Computational Linguistics (ACL-2007),* Pages 228-231, Prague, Czech Republic, June 2007.

[Lemur] Lemur: http://www.lemurproject.org/

[Levin 1993] B. Levin, *English Verb Classes and Alternations: A Preliminary Investigation*., University of Chicago Press Chicago, IL, 1993.

[Lin and Hovy 2002[. Chi-Yew Lin and Eduard Hovy. Manual and Automatic Evaluation of Summaries. In *Proceedings of the ACL 2002 Workshop on Text Summarization*, pages 45-51, Philadelphia, PA July 2002.

[Lin and Hovy 2003] Chin-Yew Lin and Eduard H. Hovy. Automatic Evaluation of Summaries using N-gram Co-occurrence Statistics. In *Proceedings of 2003 Language Technology Conference (HLT-NAACL 2003)*, Edmonton, Canada, May 27 - June 1, 2003.

[Lin 2004] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of the Workshop of Text Summarization Branches Out (WAS 2004),* Barcelona, Spain, July 25-26, 2004.

[Lin and Demmer-Fushman 2005] Jimmy Lin and Dina Demmer-Fushman. Automatically Evaluating Answers to Definition Questions. *Proceedings of the 2005 Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005},* pages 931-938, Vancouver, British Columbia, Canada, October 2005.

[Lin 2006] http://www.umiacs.umd.edu/~jimmylin/downloads/index.html.

[Lin and Demmer-Fushman 2006] Jimmy Lin and Dina Demmer-Fuschman. Will Pyramids Built of Nuggets Topple Over? *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, page 383-390. New York, June 2006.

[Liu et. al. 2007] Jingjing Liu, Yunbo Cao, Chin-Yew Lin, Yalou Huang, and Ming Zhou. Low-Quality Product Review Detection in Opinion Summarization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 334-342, Prague, June 2007.

[LCS] LCS Database: www.umiacs.umd.edu/~bonnie/LCS_Database_Documentation.html

[LUCENE] Lucene: http://lucene.apache.org/

[Luhn 1958] P.H. Luhn. Automatic Creation of Literature Abstracts. IBM Journal, pages 159-165, 1958.

[Madnani et. al. 2007] Nitin Madnani, Rebecca Passonneau, Necip Fazil Ayan, John M. Conroy, Bonnie J. Dorr, Judith L. Klavans, Dianne P. O'Leary, and Judith D. Schlesinger, `Measuring Variability in Sentence Ordering for News Summarization, In *11th European Workshop on Natural Language Generation (ENLG07)* Schloss Dagstuhl, Germany, June 17-20, 2007

Mani, I. And Bloedern, E. Multi-document Summarization by Graph Search and Merging. In Proceedings of AAAI-97, pages 622-628. AAAI, 1997.

[Mani et. al 1998] Mani, I., House, D., Klain, G., Hirschman, L, Obrst, L., Firmin, T., Chrzanowki, M., and Sundhim, B. *The TIPSTER SUMMAC Text Summarization Evaluation.* Technical Report MTR 98W0000138, MITRE, October 1998.

[Mani 2001] Inderjeet Mani, *Automatic Summarization*, John Benjamins Publishing 2001.

[Marcu 1997] D. Marcu. From Discourse Structures to Text Summaries, In *Proceedings of the Workshop on Intelligent Scalable Text Summarization*, Madrid, Spain, July 1997. p. 82-88.

[Marton and Radul 2006] Gregory Marton and Alexey Radul. Nuggeteer: Automatic Nugget-Based Evaluation using Descriptions and Judgements. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pages 375-382. Brooklyn, New York, June 2006.

[McKeown et. al. 1995] K. McKeown, J. Robin, and K. Kukich Designing and evaluation a new revision-based model for summary generation. *Information Processing and Management*, 31(5), 1995.

[McKeown et. al. 1999a] K. R. McKeown, J.L. Klavans, V. Hatzivassiloglou, R. Barzilay and E. Eskin, E. Selecting Text Spans for Document Summaries: Heuristics and Metrics. *In Proceedings of AAAI-99*, Orlando, FL, July 1999.

[McKeown et. al. 1999b] Kathleen McKeown, Judith Klavans, Vasileois Hatzivassiloglou, Regina Barzilay and Eleazar Eskin. Towards Multidocument Summarization by Reformulation: Progress and Prospects. *In Proceedings of AAAI/IAAI 1999*.

[Mihalcea and Ceylan 2007] Rada Mihalcea and Hakan Ceylan. Explorations in Automatic Book Summarization. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 380-390. Prague, June 2007.

[Mitra e. al. 1997] M. Mitra, A. Singhal, and C. Buckley, Automatic Text Summarization by Paragraph Extraction. In *Proceedings of the Workshop on Intelligent Scalable Text Summarization*, Madrid, Spain, July 1997. p. 31-36.

[Mittal et. al. 1999] Vibhu O. Mittal, Mark Kantrowitz, Jade Goldstein, and Jaime Carbonell. Selecting Text Spans for Document Summaries: Heuristics and Metrics. In *Proceedings of AAAI-99*, Orlando, Florida, July 1999.

[MSE] http://research.microsoft.com/~lucyv/MSE2006.htm

[Murakoshi et. al 1999]. Hiroyuki Murakoshi, Akira Shimazu and Koichiro Ochimizu. Construction of Deliberation Structure in Email Communication. In *Proceedings of the Pacific Association for Computational Linguistics* (PACLING '99). August 1999.

[Myka 1997] Myka. "Postings on a Genre of Email," *Genre and Writing*, Bishop, W. and Ostrom, H. (eds.), Boynton/Cook-Heinemann, Portsmouth, NH, 1997.

[NaiveBayes] Naïve Bayes – WEKA implementation http://www.cs.waikato.ac.nz/ml/weka

Nenkova, A. and Baggga, A., "Email Classification for Contact Centers.," *Proceedings of 2003 ACM Symposium on Applied Computing*, 2003, 789-792.

[Nenkova and Bagga 2003]. A. Nenkova and A. Bagga. Faciliating Email Thread Access by Extractive Summary Generation. In Proceedings of RANLP, Bulgaria 2003.

[Nenkova and Passonneau 2004] Ani Nenkkova and Rebecca Passonneau. Evaluation Content Selection in Summarization: The Pyramid Method. In *Proceedings of the HLT-NAACL Conference*. Boston, MA 2004.

[Nenkova et. al 2007] Ani Nenkova, Rebecca Passonneau and Kathleen McKeown. The Pyramid Method: Incorporating Human Content Selection Variation in Summarization Evaluation. ACM Transactions, on Speech and Language Processing, volume 4, Issue 2, April 2007.

[NEWSBLASTER] http://newsblaster.cs.columbia.edu/

[NEWSINESSENCE] http://lada.si.umich.edu:8080/clair/nie1/nie.cgi

[NTCIR] http://research.nii.ac.jp/ntcir/

[Nguyen et. al 2007] Patrick Nguyen, Milind Mahajan and Geoffrey Zeig. Summarization of Multiple User Reviews in the Restaurant Domain. Microsoft Technical Report. MSR-TR-2007-126 September 2007.

[Ogilvie and Callan 2002] Paul Ogilvie and Jamie Callan. Experiments using the Lemur toolkit. In *Proceedings of the 2001 Text REtrieval Conference (TREC 2001)* (pp. 103-108). National Institute of Standards and Technology, special publication 500-250.

[Okamura] Manabu Okamura, Takahiro Fukusima and Namba Hidestsugu, Text Summarization Challenge 2 Text summarization evaluation at NTCIR Workshop 3.

[Olesker 2005] M. Olesker. E-mails show Steffen not 'irrelevant,' 'mid-level'. *Baltimore Sunpaper*, Baltimore, MD, Mar 14, 2005.

[Paice 1990] C.D. Paice, Constructing Literature Abstracts by Computer: Techniques and Prospects. Information Processing and Management, 26:171-186, 1990.

[Papineni et. al. 2002] Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th

Annual Meeting of the Association of Computational Linguistics (ACL), pages 311-318, Philadelphia, PA July 2002.

[PASCAL-RTE] http://pascallin.ecs.soton.ac.uk/Challenges/RTE/

[Passonneau et. al 2005] Rebecca J. Passonneau, Ani Nenkova, Kathleen McKeown and Sergey Sigelman. Applying the Pyramid Method in DUC 2005. In Pr*oceedings of the Fifth Document Understanding Conference (DUC),* Vancouver, Canada 2005.

[Pew 2005] Pew Internet and American Life Project . 2005. http://www.pewinternet.org

Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization. Madrid, Spain 1997.

[RandomForests]: www.stat.berkeley.edu/users/breiman/RandomForests

[Radev and McKeown 1998] Dragomir Radev and Kathleen McKeown. Generating Natural Language Summaries from Multiple Online Sources. *Computational Linguistics*, 24(3);569-501, September 1998.

[Radev et. al. 2000] Dragomir Radev, Hongyan Jing and Malgorzata Budzikowska. Centroid-based Summarization of Multiple Documents: Sentence Extraction, Utility-Based Evaluation and User Studies. In *ANLP/NAACL 2000 Workshop,* April 2000.

[Radev 2000] Dragomir Radev. A Common Theory of Information Fusion from Mutliple Text Sources Step One: Cross-Document Structure. In Proceedings of the First SIGdial Workshop on Discourse and Dialogue. Hong Kong October 2000.

[Radev et. al. 2002] Dragomir Radev, Simone Teufel, Horacio Saggion, Wai Lam, John Blitzer, Arda Celebi, Hong Qi, Elliott Drabek and Danyu Liu. Evaluation of text summarization in a cross-lingual information retrieval framework. Technical Report, Center for Language and Speech Processing, John Hopkins University, Baltimore, MD June 2002. John Hopkins University 2001 Summer Workshop Final Report.

[Radev and Tam 2003] Dragomir Radev and D. Tam. Single document and Multi-Document Summarization Evaluation via Relative Utility, CIKM 2003.

[Rijsbergen 1979] C.J. Van Rijsbergen, *Information Retrieval*, Butterworths 1979.

[Riloff et. al. 2003]. Ellen Riloff, Janyce Wiebe and Theresa Wilson. Learning Subjective Nouns using Extraction Pattern Bootstrapping. In Proceedings of the 7th CoNLL conference held at HLT-NAACL 2003, p. 25-32. Edmonton, Canada, May-June 2003.

[Roussinov et. al 2001] D. Roussinov, K. Crowston, M. Nilan, .B. Kwasnik, X. Liu, & J. Cai., Genre-based navigation on the Web. In *34th Hawaii International Conference on System Science (HICSS-34)* (Maui, HI), January 2001.

Salton, G., Automatic Processing of Foreign Language Documents. *Journal of American Society for Information Sciences*, 21:187-194, 1970.

[Salton 1989] G. Salton., *Automatic Text Processing*: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley, 1989.

[Salton and Buckley 1990] G. Salton and C. Buckley. Improving Retrieval Performance by Relevance Feedback. *Journal of American Society for Information Sciences*, 41:288-297, 1990.

[Santini 2006] M. Santini, Common Criteria for Genre Classification: Annotation and Granularity, In *Workshop on Text Based Information Retrieval (TIR-06), In Conjunction with ECAI 2006*, Riva del Garda, Italy – Aug 29th, 2006.

[Schneider 2003]. K.-M. Schneider, A Comparison of Event Models for Naïve Bayes Anti-Spam E-Mail Filtering, In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003),* Budapest, Hungary, 2003.

[Schlesinger et. al 2008] Judith D. Schlesinger, Dianne P. O'Leary and John M. Conroy, Arabic/English Multi-document Summarization with CLASSY – The Past and the Future. In A. Gelbukh (Ed.): CICLing 2008, LNCS 4919, pp. 568-581, 2008.

[Seki et. al. 2006] Y. Seki, K. Eguchi, N., Kando, and M. Aono, Opinion-focused Summarization and its Analysis at DUC 2006. In *Document Understanding Workshop in conjunction with HLT-NAACL 2006*. Brooklyn, NY June 8-9, 2006.

[Shaw 1995] J. Shaw, Conciseness through Aggregation in Text Generation. *In Proceedings of 33rd Association for Computational Linguistics*, pages 329-331, 1995.

[Shepherd and Watters] M. Shepherd and C. Watters., The Functionality Attribute of Cybergenres, In *Proceedings of the 32nd Hawaii International Conference on System Sciences (HICSS1999),* 1999.

[SOLR]: http://lucene.apache.org/solr/

[Sparck-Jones and Galliers 1996] Karen Sparck-Jones, and J.R. Galliers. *Evaluating Natural Language Processing Systems*: an Analysis and Review. Springer, New York, 1996.

[Sparck-Jones 1995] Karen Sparck Jones, Discourse Modelling for Automatic Summarizing in Prague Linguistic Circle Papers, pages 2001-227 John Benjamins Publishing Company 1995.

[Stamatos et. al. 2000] E. Stamatatos, N. Fakotakis, and G. Kokkinakis., Text genre detection using common word frequencies. In *18th International Conference on Computational Linguistics*, 2000.

[Strzalkowski et. al. 1998] T. Strzalkowski, J. Wang, J., and B. Wise. A Robust Practical Text Summarization System. In *AAAI Intelligent Text Summarization Workshop,* Stanford, CA March 1998, 26-30.

[Stolcke et. al. 2000] Andreas. Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Caron Van Ess-Dykema and Marie Meteer. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech., In *Computational Linguistics, 26(3), 2000*, 339 – 373.

[Sun et. al. 2005] Jian-Tao Sun; Dou Shen, Haujen, Zeng, Qiang Yang, Yuchang Lu, Zheng Chen. Web-Page Summarization Using Clickthrough Data. . In Proceedings of the 28th Annual International ACM SIGIR Conference (SIGIR'2005) Salvador, Brazil, August 2005.

[SVM_LIGHT] SVM-light webpages http://www-ai.cs.uni-dortmund.de/SOFTWARE/SVM_LIGHT/svm_light.html

[SVM-Multiclass] http://svmlight.joachims.org/svm_multiclass.html

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

[Tait 1983] J.I. Tait, *Automatic Summarization of English Texts*. PhD thesis, University of Cambridge, Cambridge, UK, 1983.

[Taylor 1992] P. Taylor, Social Epistemic Rhetoric and Chaotic Discourse, In *Re-Imagining Computers and Composition*, Hawisher, G and LeBlanc, P. (ed.), Boynton/Cook-Heinemann, Portsmouth, NH, 1992.

[Teufel and Moens 1997] Simone Teufel and Marc Moens, Sentence Extraction as a Classification Task. In *ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization*, Madrid, Span, July 1997, 58-65.

[Teufel and Moens 1999] Simone Teufel and Marc Moens, Argumentative Classification of Extracted Sentences as a First Step Toward Flexible Abstracting. In I. Mani and M. Maybury (eds), *Advances in Automatic Text Summarization*, MIT Press 1999.

[Teufel and Moens 2002] Simone Teufel and Marc Moens. Summarizing Scientific Articles – Experiments with Relevance and Rhetorical Status. In *Computational Linguistics, 28 (4)* December 2002.

[Teufel and van Halteren 2004] Simone Teufel and Hans van Halteren. Evaluating Information Content by Factoid Analysis: Human Annotation and Stability. In the *2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*, Barcelona, Spain July 2004.

[TIDES] http://en.wikipedia.org/wiki/DARPA_TIDES_program

TIPSTER Text Phase III 18 month Workshop Notes, May 1998. Fairfax, VA.

TIPSTER Text Phase III Workwhop, 1998. Baltimore, MD.

[Titov and McDonald 2008]]. Ivan Titov and Ryan McDonald. A Joint Model of Text and Aspect Ratings for Sentiment Summarization. In Proceedings of ACL 2008, Ohio, June 2008.

[Tombros 1998] A. Tombros. and M. Sanderson. Advantages of Query Biased Summaries in Information Retrieval. In Proceedings of SIGIR-98, Melbourne, Australia, August 1998.

[TnT] TnT webpages:
www.coli.uni-saarland.de/~thorsten/TnT

[TREC] http://trec.nist.gov/

[TREC-QA} http://trec.nist.gov/data/qa.html

[ULTIMATERESEARCH] http://ultimate-research-assistant.com/

[Turenne 2003] Turenne, N., "Learning Semantic Classes for Improving Email Classification," *Proceedings of Text Mining and Link Analysis Workshop*, 2003.

[van Halteren and Teufel 2003] Hans van Halteren and Simone Teufel. Examine the Consensus between Human Summaries: Initial Experiences with Factoid Analysis. In *Proceedings of the HLT-NAACL Workshop on Automatic Summarization*, Edmonton, Canada 2003.

[van Rijsbergen 1979] C. Van Rijsbergen. Information Retrieval. Butterworths, London 1979.

[Vendler 1972] Z. Vendler, *Res cogitans*, Cornell University Press, Ithaca, NY, 1972.

[VerbNet] VerbNet web pages:
www.cis.upenn.edu/~mpalmer/project_pages/VerbNet.htm

[Vorhees 2005] Ellen Vorhees. Using Question Series To Evaluation Question Answering System Effectiveness, *In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP),* pages 299-306, Vancouver, British Columbia, Canada, October 2005.

[Wan and McKeown 2004] Stephen Wan and Kathy McKeown. Generating Overview Summaries of Ongoing Email Thread Discussions. In the Proceedings of the 20th International Conference on Computational Linguistics, Geneva, Switzerland 2004.

[Weibe et. al. 2005] Janyce Wiebe, Theresa Wilson and Claire Cardie. Annotating Expressions of Opinions and Emotions in Language. In *Language Resources and Evaluation*, volume 39, issues 2-3, pp. 165-210.

[White et. al. 2007] J.V. White, D. Hunger and J.D. Goldstein. Statistical Evaluation of Information Distillation Systems. In Proceedgins of the Sixth International Language Resources and Evaluation LREC '08, Marrakech, Morocco 2008.

[Wikipedia] http://www.wikipedia.org

[Wilson et. al. 2005]. Theresa Wilson, Janyce Wiebe and Paul Hoffman. Recognizing Content Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005),* Vancouver, Canada.

[Wu et. al. 2008] Yu-Chieh Wu, Yue-Shi Lee and Jie-Chi Yang. Robust and efficient multiclass SVM for phrase pattern recognition. In *Pattern Recognition* Volume 41, Issue 9, September 2008.

[Xie 2000] Hong Xie. Shifts of Interactive Intentions and Information-Seeking Strategies in Interactive Information Retrieval. In *Journal of the American Society for Information Science*: 51(9):841-857, 2000.

[Xu and Croft 1996] J. Xu, and B. Croft, Query Expansion Using Local and Global Document Analysis. In Proceedings of the 19th ACM/SIGIR (SIGIR-96) 1996. 4-11.

[Yang and Liu 1999] Y. Yang, and X. Liu, A Re-examination of Text Categorization Methods. In *Proceedings of the 22nd ACM International Conference on Research and Development in Information Retrieval (SIGI*R), Berkeley, CA, August 1999, 42-49.

[Yang and Pederson 1997] Y. Yang and J. Pedersen. A Comparative Study on Feature Selection in Text Categorization, in *Proceedings of the 1997 International Conference on Machine Learning (ICML)* 412-420.

[Yang et. al. 1998] Y. Yang, T. Pierce T., and J.G. Carbonell. A study on retrospective and on-line event detection. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98),* pages 28-36, 1998.

[Ye et. al 2005] Shirin Ye, Long Qiu, Tat-Seng Chua, Min-Yen Kan. NUS at DUC 2005: Understanding Documents via Concept Links. In Pr*oceedings of the Fifth Document Understanding Conference (DUC),* Vancouver, Canada October 2005.

[Zajic et. al 2008] David Zajic, Bonnie Dorr and Jimmy Lin. Single-document and Multi-document Summarization Techniques for Email Threads Using Sentence Compression. In Information Processing and Management: an International Journal, Volume 44, Issue 4 (July 2008)

[Zhou et. al 2005] Liang Zhou, Erin Shaw, Chin-Yew Lin and Eduard Hovy. Classummary: Introduction Discussion Summarization to Online Classrooms. In *Proceedings of HLT 2005 (demo)*. Vancouver, Canada October 2005.

[Zhu and Penn 2005] Xiaodan Zhu and Gerald Penn. Evaluation of Sentence Selection for Speech Summarization". *In Proceedings of RANLP Workshop on Crossing Barriers in Text Summarization Research*, 2005.

[Zhuang et. al. 2006] Li Zhuang, Feng Jing, and Xiao-yan Zhu. Movie Review Mining and Summarization. In Proceedings of the 15th ACM International Conference on Information and Knowledge Management, Arlington, Virginia. November 2006.