ABSTRACT

Title of dissertation: Text Summarization Evaluation: Correlating Human Performance on an Extrinsic Task with Automatic Intrinsic Metrics

Stacy F. Hobson
Doctor of Philosophy, 2007

Dissertation directed by: Professor Bonnie J. Dorr
Department of Computer Science

Text summarization evaluation is the process of assessing the quality of an individual summary produced by human or automatic methods. Many techniques have been proposed for text summarization and researchers require an easy and uniform method for evaluation of their summarization systems. Human evaluations are often costly, labor-intensive and time-consuming, but are known to produce the most accurate results. Automatic evaluations are fast, easy to use and reusable, but the quality of their results have not been independently shown to be similar to that of human evaluations.

This thesis introduces a new human task-based summarization evaluation measure called Relevance Prediction that is a more intuitive measure of an individual's performance on a real-world task than agreement based on external judgments. Relevance Prediction parallels what a user does in the real world task of browsing a set of documents using standard search tools, i.e., the user judges

relevance based on a short summary and then that same user—not an independent user—decides whether to open (and judge) the corresponding document. This measure is shown to be a more reliable measure of task performance than LDC Agreement, a current external gold-standard based measure used in the summarization evaluation community.

Six experimental studies are conducted to examine the existence of correlations between the human task-based evaluations of text summarization and the output of current intrinsic automatic evaluation metrics. The experimental results indicate that moderate, yet consistent correlations exist between the Relevance-Prediction method and the ROUGE metric for single-document summarization.

This work also formally establishes the usefulness of text summarization in reducing task time while maintaining a similar level of task judgment accuracy as seen with the full text documents.

TEXT SUMMARIZATION EVALUATION: CORRELATION HUMAN PERFORMANCE ON AN EXTRINSIC TASK WITH AUTOMATIC INTRINSIC METRICS

by

Stacy F. Hobson

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2007

Advisory Commmittee:

Professor Bonnie J. Dorr, Chair/Advisor
Professor Avis Cohen
Professor David Huber
Professor James Reggia
Professor Thomas Wallsten

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

With the increased usage of the internet, tasks such as browsing and retrieval of information have become commonplace. Users often skim the first few lines of a document or prefer to have information presented in a reduced or summarized form. Examples of this include document abstracts, news headlines, movie previews and document summaries. Human generated summaries are often costly and time consuming to produce. Therefore, many automatic summarization algorithms/techniques have been proposed to solve the task of text summarization.

To measure the impact of summarization techniques, it is important to have a consistent and easy-to-use method for determining the quality of a given summary (how reflective the summary is of the original document's meaning) and for comparing a summary against other automatic and human summaries. Currently, numerous automatic and semi-automatic evaluation metrics have been developed and are becoming more widely used in the text summarization evaluation community. Many of these methods claim to correlate *highly* (Papineni et al., 2002) or

surprisingly well (Lin and Hovy, 2003) with human measures of task performance, and a goal of this work is to investigate these claims. Therefore, five relevance-assessment experiments were conducted to compare automatic evaluation metrics with judgments of human performance.

In the first two experiments, users were asked to determine the relevance of a particular document to a specified topic or event, based on the presented document summary or entire document text. Judgments made by individual users were compared to "gold standard" judgments as provided by the University of Pennsylvania's Linguistic Data Consortium (LDC, 2006); this evaluation approach is further referred to as *LDC Agreement*. These gold standards were considered to be the "correct" judgments, yet they yielded very low interannotator agreement rates and inconsistencies in the user's judgments. Thus, it was difficult to make strong statistical statements using the results of these earlier experiments.

This thesis introduces a new measurement technique, called *Relevance Prediction*, that yields better agreement levels than *LDC Agreement*. Relevance Prediction is a more intuitive measure of an individual's performance on a real-world task than interannotator agreement. Specifically, Relevance Prediction parallels what a user does in the real world task of browsing a set of documents using standard search tools, i.e., the user judges relevance based on a short summary and then that *same* user—not an independent user—decides whether to open (and judge) the corresponding document. This method eliminates the need for an externally induced "gold standard" by making use of the same user's relevance judgment on

both the summary and the corresponding full text.

Relevance Prediction provides a stable framework within which developers of new automatic measures may verify more reliably—through correlation studies—the effectiveness of their measures in predicting summary usefulness. It is demonstrated—as a proof-of-concept methodology for automatic metric developers—that current automatic evaluation measures have better correlations with Relevance Prediction than with LDC Agreement and that the significance level for detected differences is higher for the former than for the latter. As such, automatic metric developers may use Relevance Prediction to make stronger statistical statements about the effectiveness of their measures in predicting summary usefulness.

1.1 Motivation

Text summarization evaluation is an area wrought with many challenges. Human evaluations of summary quality are very expensive, labor intensive and time consuming. Participants are usually compensated financially or assigned assessment tasks as part of their normal daily job requirements. Tasks can last from one to a few hours per participant depending upon the number of documents and summaries to be judged.

Participants' judgments vary greatly and generally do not match gold standard judgments. Very low agreement rates have been reported by Mani (2001) and

Tombros and Sanderson (1998) in studies that use such standards. At least four total participants are usually needed to produce representative results, although more participants are needed for the most reliable results.

These and other challenges have led researchers to investigate the use of automatic summarization evaluation methods. Such methods are fast, inexpensive, easy to use, and reusable; moreover, they allow developers to continuously check for improvements based on small changes to their summarization system. Two examples of fully automatic intrinsic measures are BLEU (Papineni et al., 2002), a modified n-gram precision-based metric, and ROUGE (Lin and Hovy, 2003; Lin, 2004), a modified n-gram recall-based metric. Recently, two content-based measures, Basic Elements (BE) (Hovy et al., 2005, 2006) and The Pyramid Method (Nenkova and Passonneau, 2004) have been proposed for text summarization evaluation.

One issue with these methods is that they adopted an evaluation design that was *intrinsic* in nature, i.e., assessments of summary quality are made without reference to a particular task. Of these, *human* intrinsic evaluations have been used to assess the summarization system itself, based on factors such as clarity, coherence, fluency and informativeness (Jing et al., 1998). Alternatively, *automatic* intrinsic evaluation measures have been used to compare a candidate summary (output of a summarizer) against an 'ideal' or model human summary (Mani et al., 2002).

While important, intrinsic measures do not address an *extrinsic* question that

is central to this work: *how is text summarization useful*? Although summaries are thought to help reduce cognitive load (Tombros and Sanderson, 1998), two other possible benefits of using a summary over the full text are investigated: (1) Summaries should reduce the reading and judgment time for relevance assessments or other tasks; and (2) Summaries should provide enough information for a reader to get the general meaning of a document so that he/she can make judgments that are as accurate as the judgments on full texts in a relevance assessment task.

Previous work (Mani et al., 2002) demonstrated that users can read summaries faster than the full text, with some loss of accuracy; however, researchers have found it difficult to draw strong conclusions about the usefulness of summarization due to the low level of interannotator consistency in the gold standards that they have used. This thesis defines a new extrinsic method, *Relevance Prediction*, that eliminates gold-standard judgments and is thought to be a better indicator of "summary usefulness" and a more reliable predictor of task performance than previous gold-standard methods. This method is used to demonstrate that human task performance measures correlate with intrinsic automatic summarization measures. This work yields a usable framework for drawing definitive conclusions about summary usefulness, thus justifying continued research and development of new summarization methods.

1.2 Experimental Studies

A major goal of this research is to objectively study and compare various evaluation methods and to provide the summarization evaluation community with empirically grounded findings and suggestions on improving current methods and techniques. The six experimental studies conducted as part of this research are briefly outlined below.

- Experiment 1: *LDC General*. This study aims to determine whether two automatic evaluation metrics, BLEU and ROUGE, correlate with human performance on a relevance assessment task. Six summarizers (four automatic, two human) are tested using NIST topic and document sets. The evaluation uses LDC Agreement, i.e., comparison to an externally produced gold-standard, as the basis for the analysis.

- Experiment 2: *LDC Event Tracking*. This study continues to investigate correlations with BLEU and ROUGE, but uses the TDT-3 document collection for an event tracking relevance assessment task. Event tracking is similar to the real-world task of web browsing and information retrieval and is thought to be more reliable than assessment task in previous experiment. Nine summarizers (six automatic, two human and first-75 character baseline) are tested and the results are intepreted using the LDC-Agreement method.

- Experiment 3: *Memory and Priming Study*. This study explores the effects of ordering of documents and summaries on user performance. Re-

sults of ten different orderings are compared in a two part experiment, with part 2 at least one week after part 1, to minimize memory effects. The performance scores are produced by comparing the judgment made on the summary with the judgment made on the corresponding full text document (within the same experimental trial), or by comparing the judgment made on a summary/document on week 1 with the judgments made on the same summary/document on week 2.

- Experiment 4: *RP with Human Summaries*. This study introduces the Relevance-Prediction measurement technique and compares human performance scores produced by the new Relevance-Prediction (RP) method with scores produced by LDC Agreement. The correlation of the ROUGE metric with human performance in an event tracking task is investigated. Two human summary types are tested: the original document headline, and human-generated summaries.

- Experiment 5: *RP Dual Summary*. This study continues to compare the Relevance-Prediction and LDC-Agreement methods and examines correlations with the ROUGE metric. A new content-based method, Basic Elements (BE) is introduced and used for intrinsic evaluations and correlations with the two extrinsic methods. Two human summary types, one automatic summarizer, and the first-75 character baseline are tested.

- Experiment 6: *RP with Multi-Document Summaries*. This study explores

the extension of Relevance Prediction to the problem of multi-document summarization and examines correlation differences in this context. The extension to the multi-document framework involves three steps: (1) varying the distribution of relevant and non-relevant documents among five distinct combinations; (2) introducing a five-point likert scale for judgments; and (3) devising two additional scoring methods basec on the new judgment scale. Correlations of the human performance results with ROUGE are also investigated.

1.3 Contributions

The experiments detailed in this work show that the Relevance-Prediction method is a better performance metric than LDC-Agreement method and that the elimination of external gold standards produces more stable results. The findings also show small, positive correlations with some automatic intrinsic evaluations metrics and human task-based Relevance-Prediction measurements.

The specific contributions of this thesis are:

- Provision of a means for determining quality of current summarization evaluation methods based on the level of correlation with human judgment measurements.

- Development of a methodology for conducting human evaluations to determine the usefulness of text summarization.

- Introduction of a new method, Relevance Prediction, that is more reliable than current "gold standard" methods for measuring human performance.
- Exploration of the factors that affect performance scoring including Single versus Multi-document summarization, summary length, summary type (abstractive versus extractive and indicative versus informative).
- Creation and implementation of a new evaluation approach incorporating a 5-point likert scale for evaluation of multi-document summaries.
- Use of the results of the human evaluations to compare summarization techniques.

1.4 Outline

The next chapter will detail some of the background of the field and discuss related work. Both intrinsic and extrinsic evaluation methods are described. Chapter 3 discusses a previous external gold-standard relevance assessment method and describes a new and more intuitive method, Relevance Prediction. Chapter 4 uses the previous relevance assessment method to evaluate human experimental results and investigates correlation with automatic evaluation methods. This chapter also includes details of a memory and priming study—used to determine whether the experimental design influences the results. Chapter 5 introduces the Relevance-Prediction method in an experimental study of human summaries and contrasts scoring and correlation results with those of the LDC-Agreement method. Chap-

ter 6 continues the investigation of the Relevance-Prediction method with a study that also includes automatic summaries. Chapter 7 introduces multi-document summaries as part of the experimental data set and notes the differences produced in the human and automatic evaluation results. Finally, the overall findings, the specific contributions of this thesis and directions for future work are presented in Chapter 8.

Chapter 2

Background

This chapter provides the background and motivation for summarization evaluation. The factors involved in text summarization often influence summarization evaluation methods and can explain some differences in human judgment performance from one text summarization system to another. Section 2.1 defines text summarization and describes some of the main summarization types including human, automatic, single, and multi-document. Section 2.2 introduces the two summarization evaluation methods and the specific task-based evaluation method that is used in the experimental studies.

2.1 Text Summarization

Text summarization is the process of distilling the most important information from a set of sources to produce an abridged version for particular users and tasks (Maybury, 1995). Producing a summary that accurately reflects the meaning of the source text is a difficult task. One would not expect the summary to

contain all of the information present in the original text, but enough information that conveys the most important concepts from the source. The sections below discuss the types of text summarization and how summarization may be evaluated for usefulness.

2.1.1 Types of Text Summarization

There are many factors involved in text summarization and numerous summarization methods. Texts may be summarized by a human as in news story headlines or movie previews, or automatically as done by search engines such as Google and AltaVista.

Human summarization is currently the most preferred and reliable form of text summarization. News story headlines, movie previews and movie reviews are all examples of human summaries. They are usually considered to be of high quality, coherent and reflective of the source document.[1] However, human summaries are often time consuming and labor intensive to produce.

Automatic summarization is machine-generated output that presents the most important content from a source text to a user in a condensed form and in a manner sensitive to the user's or application's needs (Mani et al., 2002). Automatic summaries of text documents are faster and less expensive to generate

[1]News story headlines are usually intended to be 'eye-catchers' to capture a reader's interest and encourage them to read the entire article thus, they may not be directly reflective of the text source.

in comparison to human summaries. However, automatic summaries have not achieved the level of acceptance achieved by human summaries, and it has previously been shown that human summaries provide at least 30% better information than automatic summaries.[2] Various methods for automatic summarization have been proposed, and large scale evaluations such as the Document Understanding Conference (DUC), (Harman and Over, 2004) and SUMMAC (Mani et al., 2002) have been conducted to judge systems and understand issues with summarization.

Single document summarization is the summarization of only one text document and can be thought of mostly as an aid to information retrieval. When users search for information online, they may require a single document to answer their question or to provide the information they need. For example, a middle schooler writing a report on the life of Abraham Lincoln may search for 'Abraham Lincoln biography' and may find it sufficient to examine a single document detailing Lincoln's life, his ascension to the presidency and his death.

Multi-document summarization is the summarizing of information from more than once source document. It is thought to be harder than single document in that more information has to be condensed into a single summary and the summary has to be reflective of more than one text source. Some summarizers rank the documents and the sentences within them using current information retrieval technologies. They can then choose the top ranked sentence (or sentences) from each document for inclusion as part of the summary. If this procedure creates a

[2]K. McKeown, personal communication, July 2005

summary that is too large, techniques to remove redundant sentences or terms can be used or the summary can be truncated.

Extractive summaries use information directly from the source document(s). Automatic summarizers are more likely to produce extractive summaries than their abstractive counterparts (described next). Many of these rank the sentences contained within a single document or set of multiple documents and use the higher-ranking sentences as the summary. It has also been shown that using the lead sentence or leading characters (the first sentence or first few characters of a document) can provide a relatively good summary (Brandow et al., 1995; Erkan and Radev, 2004). Highly extractive summaries contain only words found in the source document. Less extractive summaries pull information from the source text but add in conjunctive or limited modifying information.

Abstractive summaries may contain material not present in the source text. These are more likely to be produced by humans where synonyms, or even entire rephrasing of words appearing in the document(s) may be used to condense the meanings of multiple words into one *("The assailant fired six shots$_1$ and fatally wounded$_2$ a man who was not involved$_3$ with the..."* becomes *"gunman$_1$ killed$_2$ bystander$_3$")*. News story headlines, which are usually intended to catch a reader's interest and may not accurately reflect the contents of the document, are good examples of abstractive summaries.

Indicative summaries identify what topics are covered in source text, and alert the user to source content. These summaries generally provide a few sentences

or even a few keywords related to just one information area, sometimes in relation to a topic-based query. Indicative summaries are used in information retrieval tasks (e.g. Google searches), where a user determines whether a document (based on the summary) contains the information/topic they are looking for. If this information is indicated through the summary, the user will then retrieve or open the full text document for further reading.

Informative summaries identify the central information about an event (who, what, where, etc.). They may be used as document "surrogates," i.e., they are used to stand in place of the source document(s) when the user has to find information quickly (usually for a question answering task) and does not have time to open the full text. Many tend to include the first sentence of the source document as part of the summary. In newswire text, the first sentence is sometimes introductory, giving a general overview of the contents of the document.

Compression is also an important part of text summarization. Compression determines the size of the summary as a function of the document size. The summarization compression ratio is the ratio of the size of the compressed data to the size of the source data. This is usually set at a specific length for comparison and evaluation of summarization systems. The compression method may apply at the level of sentences, words or characters.

In evaluations where compression is required at different levels, it has been shown that informative summaries perform better at a higher compression ratio, about 35-40% (Mani and Bloedorn, 1999) because at longer lengths they are able

to include more sentences reflecting different parts or information areas from the entire document rather than providing a few sentences related to just one information area or topic (which is the goal of indicative summaries). In another study (Jing et al., 1998), it was shown that the results of evaluating a given system can change dramatically when evaluated at different summary lengths.

2.1.2 Usefulness of Summarization

A key question motivating this research is: *how is text summarization useful*? Summaries are thought to help reduce cognitive load (Tombros and Sanderson, 1998), but there are also other benefits to the use of a summary over the full text. This work focuses on two other possible benefits of using a summary over the full text: (1) Summaries should reduce the reading and judgment time for relevance assessments or other tasks; and (2) Summaries should provide enough information for a reader to get the general meaning of a document so that he/she can make judgments that are as accurate as the judgments on full texts in a relevance assessment task.

Although researchers have demonstrated that users can read summaries faster than the full text (Mani et al., 2002), with some loss of accuracy, researchers have found it difficult to draw strong conclusions about the usefulness of summarization due to the low level of interannotator consistency in the gold standards they have used. Definitive conclusions about the usefulness of summaries would provide justification for continued research and development of new summarization

methods.

The next chapter of this thesis presents a new extrinsic measure of task based usefulness called *Relevance Prediction* where a user's summary-based decision is compared to his or her own full-text decision rather than to a different user's decision. The experiments discussed in Chapters 4 through 7 show it is possible to save time using summaries for relevance assessments without greatly impacting the degree of accuracy that is achieved with full documents.

2.2 Summarization Evaluation

There are two types of summarization evaluations: intrinsic and extrinsic. The two types of intrinsic summarization evaluations are human and automatic. Human intrinsic evaluations assess the summarization system itself, based on factors such as clarity, coherence, fluency and informativeness (Jing et al., 1998). These will be discussed below in Section 2.2.1.

Automatic intrinsic evaluation measures usually compare a candidate summary (output of a summarizer) against an 'ideal' or model human summary (Mani et al., 2002). These will be discussed below in Section 2.2.2. The majority of this thesis will focus on automatic intrinsic evaluations and their correlations with human extrinsic evaluations (to be described in more detail in Chapters 3 and 4).

Extrinsic evaluations study the use of summarization for a specific task. Examples of such a task are: (1) execution of instructions, (2) information retrieval,

(3) question and answering, and (4) relevance assessments (Mani, 2001). Extrinsic evaluation measures will be discussed below in Sections 2.2.3 and 2.2.4.

2.2.1 Human Intrinsic Measures

For human intrinsic evaluations, experimental participants or laborers are asked to quantify factors of coherence, referential clarity, fluency and informativeness of a summary, or are asked to assign a score to a candidate summary in comparison to an "ideal" or reference summary. Coherence and fluency focus on the readability and grammaticality of a summary, whereas referential clarity and informativeness concentrate on the actual content of the summary.

For a measure of coherence, users rate the summary in terms of subjective grading of readability, lapses in grammaticality, presence of dangling anaphors (a common problem when extracting sentences out of context), or ravaging of structured environments like lists or tables (Mani et al., 2002). Referential clarity focuses on whether any nouns or pronouns are clearly referred to in the summary. For example, the pronoun *he* has to mean something in the context of the summary (Farzindar et al., 2005). Fluency judgments determine whether the summary presents the information in an order where there are smooth transitions from one statement, sentence, or idea to the next, or whether the information is presented in an order consistent with the source document. The informativenes measure can have the users compare the summary to the full text document (or "ideal" summaries) and determine whether the most salient information in the text is

preserved in the summary (Mani et al., 2002).

Human intrinsic measures are generally used in combination with another because a summarizer might perform well for one measure but not another. For example, one can have a coherent but bad summary (Mani et al., 2002), or an informative but poorly formed summary. Therefore, users are often asked to score summaries on multiple factors (e.g. coherence **and** informativeness). In the 2005 Document Understanding Conference (Dang, 2005), users rated the summaries on factors including clarity and coherence.

2.2.2 Automatic Intrinsic Measures

Automatic intrinsic summarization evaluation measures usually compare a candidate summary with an ideal human generated summary and use the overlap between the two for scoring. Numerous methods have been proposed; some for direct use as summarization evaluation metrics, others for use in other natural language processing application communities and sometimes extended for use with text summaries.

Eight evaluation metrics, BLEU, ROUGE, BE, the Pyramid Method, GTM, Meteor, Pourpre, and (H)TER, are described below. The BLEU metric was the first automatic evaluation method to be developed and was originally created for machine translation evaluation. After gaining popularity in the MT community, the developers later suggested that it could be used to evaluate text summaries. However, the ROUGE metric and Basic Elements (BE) were created specifically

for text summarization evaluation. These three metrics will be used as the intrinsic evaluation methods in the experiments of Chapters 4 through 7.

The Pyramid Method was also created for text summarization evaluation, but is only semi-automatic in that it still requires part of the core evaluation task to be completed by humans. The rest of the methods were designed for other areas—GTM, Meteor, and (H)TER for machine translation evaluation, and Pourpre for evaluation of question-and-answering tasks. These methods could be extended to the task of text summarization evaluation but are not widely used in the summarization evaluation community.

Each of the measures take a slightly different approach to evaluation of their respective NLP applications, some built upon the shortcomings of previous methods. The methods not specific to the summarization evaluation community are described in detail to provide additional background about types of NLP application evaluation methods and the differences between them. All the methods do have in common the claim to correlate highly with human extrinsic evaluations (described below in Section 2.2.3).

2.2.2.1 BLEU

Bilingual Language Evaluation Understudy (BLEU) (Papineni et al., 2002) is an n-gram precision based evaluation metric initially designed for the task of machine translation evaluation. It has become the standard metric in the machine translation community. The developers of BLEU also suggest that this metric

could be used for summarization evaluation.

BLEU's precision can be computed as the number of words in a candidate translation that matches words in the human generated reference translation divided by the total number of words in the candidate translation. The authors point out an issue with regular unigram precision: machine translation systems can 'overgenerate' words for the candidate that are sure to appear in the reference, allowing them to achieve a very high precision score. To combat this, a modified n-gram precision score is used in which a reference word is 'exhausted' once a matching candidate word has been counted. BLEU's modified n-gram precision is defined as:

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

where $Count_{clip}(n\text{-}gram)$ is the maximum number of $n\text{-}grams$ co-occuring in a candidate translation and a reference translation, and $Count(n\text{-}gram)$ is the number of $n\text{-}grams$ in the candidate translation. This generates what they term BLEU's modified precision score, p_n. The equation is known as precision based because the denominator is the total number of n-grams in the *candidate* translation.

BLEU also imposes a brevity penalty to ensure that extremely short candidate translations are not unfairly scored very highly. If a candidate's length matches the reference translation, the penalty is set to 1.0 (meaning no penalty). If the candidate is shorter than all the reference translations, a brevity penalty is included in the translation scoring.

For evaluation, human participants scored the readability and fluency of Chinese to English translations produced by five systems. The BLEU metric was also used to generate system scores based on these translations. Using linear regression, the authors report a correlation coefficient of 0.99 with monolingual English participants, and 0.96 with the bilingual Chinese/English participants.

It is important to note that the evaluation criterion for machine translation can be defined precisely, yet it is difficult to elicit stable judgments for summarization (Rath et al., 1961; Lin and Hovy, 2002), which may explain the reason BLEU and two additional machine translation evaluation metrics described below (GTM and Meteor) have not achieved similar acceptance in the summarization evaluation community.

2.2.2.2 ROUGE

Since BLEU uses precision-based scoring and the human evaluations at the Document Understanding Conferences (DUC) that were used for the correlations at that time were recall based, researchers at the University of Southern California's Information Sciences Institute (ISI) proposed a new recall-based evaluation metric, Recall Oriented Understudy of Gisting Evaluation (ROUGE). ROUGE is an n-gram recall between a candidate summary and a set of reference summaries (Lin and Hovy, 2003; Lin, 2004), and has surpassed BLEU in usage in the summarization community. ROUGE has also recently been adopted as the National Institute of Standards and Technology's (NIST) method for automatic intrinsic evaluation of

summarization systems.

ROUGE scoring is computed as:

$$c_n = \frac{\sum_{C \in \{ModelUnits\}} \sum_{n\text{-}gram \in C} Count_{match}(n\text{-}gram)}{\sum_{C' \in \{ModelUnits\}} \sum_{n\text{-}gram' \in C'} Count(n\text{-}gram')}$$

where $Count_{match}(n\text{-}gram)$ is the maximum number of $n\text{-}grams$ co-occuring in a candidate (peer) summary and a reference (model) summary, and $Count(n\text{-}gram)$ is the number of $n\text{-}grams$ in the reference (model) summary. This equation is recall-based because the denominator is the total number of n-grams in the *reference* summaries. (Note BLEU's precision-based equation uses the *candidate* translation for the denominator.)

The previous equation only applies when there is a single reference summary. It has been shown that the correlation between an automatic intrinsic measure (i.e. BLEU, ROUGE) increases when more than one reference summary is used. Therefore, for multiple reference summaries, a pairwise summary-level score is computed between a candidate summary, s, and every reference, r_i, in the reference set. The maximum pairwise score is used as the final ROUGE score. This is computed as:

$$ROUGE_{multi} = argmax_i ROUGE(r_i, s)$$

ROUGE does not impose a brevity penalty as BLEU does, but instead offers a brevity bonus, since a shorter, correct summary is preferred over a larger summary containing many of extraneous terms.

Currently, there are five different versions of ROUGE available[3]:

- ROUGE-N: the base recall n-gram measure as described above.
- ROUGE-L: uses a combination of recall, precision, and the longest common subsequence between a candidate and reference summary to compute the resulting f-measure score.
- ROUGE-W: similar to ROUGE-L, but also includes a weighting factor for the maximum number of matching words that appear consecutively.
- ROUGE-S: a measure of the overlap of skip-bigrams[4] between a candidate and a set of reference translations.

Currently, in the summarization community, ROUGE 1-gram is preferred for evaluating single document summaries and ROUGE 2-gram is preferred for multi-document summaries.

For correlations, the data from the 2001 Document Understanding Conference (DUC) (Harman and Marcu, 2001), which included judgments of single and multi-doc summaries by NIST human assessors on areas of content and quality (including grammaticality, cohesion and coherence), were used for the human evaluation. The summaries were also scored with the ROUGE metric for each $n\text{-}gram(1,4)_n$, and with different summary sizes (50, 100, 200 and 400 words). The authors computed the Pearson r and Spearman ρ (Siegel and Castellan, 1988)

[3]Note that ROUGE-L and ROUGE-W are a combination of precision and recall based metrics.

[4]A skip-bigram is any pair of words in the sentence order, ignoring gaps between words.

correlation values for the comparison of the human judgments and the ROUGE scores, and reported a range from 0.84 to 0.97 for Pearson's r and 0.88 to 0.99 for Spearman's ρ (with ROUGE unigrams at the various summary sizes).

Both BLEU and ROUGE use reference summaries, and base their techniques on the idea that the closer an automatic summary is to a human reference summary, the better it is. However, it is possible for an automatic summary to be of good quality (as determined by a human in an intrinsic evaluation or relevance assessment task) and not use the same words that appear in the reference summary. This would pose a challenge for either metric, in that their scoring methods rely completely on overlap with the reference summaries. Because of the challenges with the use of reference summaries, the Pyramid Method was introduced by researchers at the University of Columbia.

2.2.2.3 Pyramid Method

The Pyramid Method is a semi-automatic method in that it relies greatly on human labor, but the tallying of scores is done automatically. The method was created with the idea that no single best model summary exists. Information is ordered within reference texts by level of importance to the overall idea of the text and assigned a weight, with the most important items receiving the highest weight. The summaries would then be compared against the list of prioritized information, and assigned scores based on the appearance of the important items and the summation of their weights.

Central to the Pyramid Method is that information should not be compared on a sentence level, but on a smaller, clausal level termed Semantic Content Units (SCUs) (Nenkova and Passonneau, 2004; Passonneau and Nenkova, 2003). SCUs are not formally defined, but can be understood more through the example below.

> Reference 1 - In 1998 two Libyans indicted in 1991 for the Lockerbie bombing were still in Libya.
>
> Reference 2 - Two Libyans were indicted in 1991 for blowing up a Pan Am jumbo jet over Lockerbie, Scotland in 1988.
>
> Reference 3 - Two Libyans, accused by the United States and Britain of bombing a New York bound Pan Am jet over Lockerbie, Scotland in 1988, killing 270 people, for 10 years were harbored by Libya who claimed the suspects could not get a fair trial in America or Britain.
>
> Reference 4 - Two Libyan suspects were indicted in 1991.

The previous four reference summaries produce two SCUs[5] denoted by the underlining. The first SCU communicates that *two Libyans were accused/indicted* (of the Lockerbie bombing) and the second SCU communicates that this indictment occurred *in 1991*.

Once the SCUs have been identified, a weighted inventory—a pyramid—is created based on the appearance of the SCUs in the reference summaries. If a SCU appears in all reference summaries, it is given the highest weight, equal to

[5]More SCUs can be found, but two are used for illustrative purposes here.

the total number of reference summaries. If a SCU appears in only one reference summary, it is given the lowest weight of 1. Therefore, the pyramid has layers (or tiers) equal to the number of reference summaries.

For example, if there are four reference summaries, an SCU appearing in all four summaries can be thought of as one of the most important ideas (since all the summarizers include them in their summaries) and would receive a weight of 4. An SCU appearing in only three reference summaries would receive a weight of 3; it is still an important concept, but probably not as important as an SCU with weight of four since only three out of the four human summarizers agree on its inclusion. For the example showing the discovery of SCUs above, the first SCU *two Libyans were accused/indicted* would receive a weight of 4 since it appears in all four references. The second SCU *in 1991* would receive a weight of 3, having appeared in three of the four references.

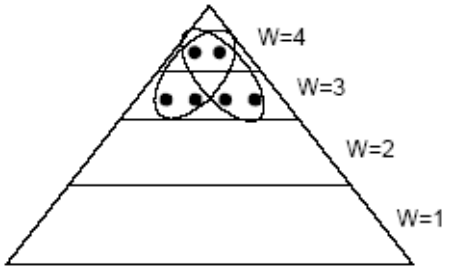

Figure 2.1: Example of a Pyramid with SCUs Identified and Marked for the Top Two Tiers from Nenkova and Passonneau (2004). W indicates the number of references associated with the SCUs at each level.

A "pyramid" is formed because the tiers descend with the SCUs assigned the highest weight at the top, and the SCUs with the lowest weight appearing in the bottom-most tiers. The fewest SCUs would appear in the topmost tier since

fewer concepts would be present in all reference summaries. In general, each tier contains fewer concepts than the tier at the next level down—because fewer SCUs are associated with n references than with $n-1$ references—as shown in Figure 2.1.

The Pyramid score is a ratio of the sum of the weights of the SCUs to the sum of the weights of an optimal summary with the same number of SCUs. A summary is considered optimal if it contains more (or all) SCUs from the top tiers and less from the lower tiers, as long as length permits. The optimal summary would not contain an SCU from tier $(n-1)$ if all the SCUs in tier n are not included because SCUs from top tiers can be thought of as the most salient information from the text because all (or most) of the reference summaries contain this information.

The formal equation for the pyramid SCU weight D:

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

where the pyramid has n tiers, with tier T_n on the top, and T_1 on the bottom. i is the weight and D_i is the number of SCUs in the candidate summary that appear in T_i. The final Pyramid score P is the ratio of D to the maximum optimal content score. For a summary with four SCUs, the maximum optimal content can be seen in Figure 2.1 with one of the circled examples. The score for this example is computed as $2 \times 3 + 2 \times 4$, for a total of 14.

Although ROUGE and BLEU can also use multiple reference summaries, an advantage of the Pyramid Method is that it relies on semantic matching for scoring rather than exact string matching; meaning the information conveyed (its ideas or

concepts) are matched rather than the exact words. However, a major drawback of the Pyramid Method is that it is not automatic and, thus, requires a lot of human effort. The creation of reference summaries, the SCU annotation, and the comparison of reference SCUs with candidate summaries are all completed through human labor. Therefore, the method becomes time consuming, labor intensive and expensive (if human laborers are financially compensated for their work).

The implementation of semantic comparisons in a fully automatic evaluation method would address many of the shortcomings of the BLEU, ROUGE and Pyramid methods. A new method, Basic Elements, was created with this intent and is described in the next section.

2.2.2.4 Basic Elements

Although the approach of the Pyramid Method with the focus on semantic-level comparisons rather than explicit term matching seems promising, the method is still primarily manual and relies heavily on clause-level meaning units. In an effort to explore the evaluation of summaries with smaller meaningful units of information while providing a foundation for a fully automatic method, the researchers at ISI (and creators of the ROUGE metric) developed a new evaluation metric, Basic Elements (BE) (Hovy et al., 2005, 2006). The BE metric uses minimal semantic units, also termed BE(s), which are defined as a triple: the head of a major syntactic constituent (noun, verb, adjective or adverbial phrase) and two arguments (or "dependents) of that head (Hovy et al., 2005). Examples of BEs

include “United States of America” (where the triple is “OF(United States, America)”), “coffee mug” (where the triple is “HOLDS(mug, coffee)”), “the/a plane landed” (where the triple is “LAND(Plane,_)”), and “the landing was safe”(where the triple is “BE(Landing, safe)”).

This predicate-argument structure allows BE to focus on the semantic relationship of terms within a sentence and makes it easier to match information in two sentences that have the same meaning but are expressed differently. The sentences “Bob hit Sue” and “Sue was struck by Bob” both identify that bob did the action of hitting, and that an action of hitting was done to Sue, and the BE’s of the two sentences would reflect this association.

The BE method extracts semantic units automatically using four modules:

- BE Breakers which create individual BE units, given a text.
- BE Scorers that assign scores to each BE unit individually.
- BE Matcher that rates the similarity of two BE units.
- BE Score Integrators which produce a score given a list of rated BE units.

The first three modules, BE Breakers, Scorers and Matcher are automatic and are currently implemented as part of the BE system. The fourth module, BE Score Integrators, is suggested as a part of the package, but has not been implemented yet.

Reference summaries are submitted as input to the system and the BE Breakers creates a preferred list of BEs, ranked from the most important to the least

important. The candidate summary is also submitted to the BE Breakers, and the BEs created from the candidate are compared against the reference BEs for scoring.

The BE matcher module currently allows matching of BEs based on exact words or the root forms of words ('introduces' will match with 'introduced' since the root form for both words are 'introduce'), but extensions to include synonym matches and phrasal paraphrase matching are also being implemented.

For correlations, the authors compare BE, ROUGE (which they state is an instance of the BE method in which the BEs are unigrams), the Pyramid Method, and a responsiveness score[6] from NIST's 2005 Document Understanding Conference (Dang, 2005). Their results suggested that BE correlated more highly with the human responsiveness measure than the Pyramid Method using both the Spearman rank coefficient and the Pearson coefficient. They also suggest that BE has a slightly higher Pearson correlation than ROUGE, yet ROUGE has a slightly higher Spearman correlation.

2.2.2.5 GTM

An issue with the BLEU metric is the inability to make definitive conclusions about a system based solely on its BLEU score. BLEU produces scores that allow

[6]The responsiveness score is a coarse ranking of the summaries for each topic, based on the amount of information given in the summary. The NIST assessors assigned these scores, ranging from 1 to 5, with 1 being least responsive and 5 being most responsive (Dang, 2005).

systems to be ranked, but it is difficult to determine exactly what a particular score means. For example, one cannot say that translation or summarization system$_a$ with a BLEU score of 0.5 is only half as good as an ideal or human translation or summary. Although one could say that system$_b$ scoring 0.6 performed better than than system$_a$, one must wonder, how much better? Is system system$_b$ an acceptable translator/summarizer whereas system$_a$ produces poor quality output; or are both system$_a$ and system$_b$ of poor quality?

To address some of these issues noted with BLEU, the General Text Matcher (GTM) machine translation evaluation metric was proposed (Melamed et al., 2003). GTM bases its scoring on the common natural language processing measures of precision and recall. For the base measures, given a set of candidate translations/summaries, Y, and a set of reference translations/summaries, X,

$$Precision(Y|X) = \frac{|X \cap Y|}{|Y|}$$

and

$$Recall(Y|X) = \frac{|X \cap Y|}{|X|}.$$

An important concept of the GTM method is the notion of coordination of words in the reference and candidate texts, which can be projected onto a bitext grid. A visual example of this matching can be seen with the bitext grid of Figure 2.2, in which the reference text is represented on the X axis and the candidate text is represented on the Y axis. The cells in the grid represent the coordination of some word in the reference text with some word in the candidate

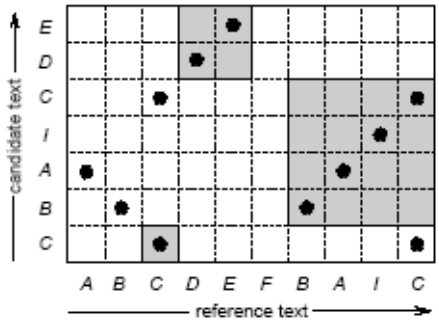

Figure 2.2: Bitext Grid Example from Melamed et al. (2003)

text. If the two words denoted by a cell match, then that is considered a hit.

The GTM method introduces a new concept called Maximum Matching Size (MMS). MMS of a bitext is the size of the largest number of hits in a subset containing only one hit per row or column (so that word matches are not counted more than once in the subset). The MMS as seen in Figure 2.2 is 7. This definition produces an MMS between 0 and the length of the shortest text (candidate or reference). The GTM Recall and Precision scores given a set of candidate translations/summaries, Y, and a set of reference translations/summaries, X,

$$Precision(Y|X) = \frac{MMS(X, Y)}{|Y|}$$

and

$$Recall(Y|X) = \frac{MMS(X, Y)}{|X|}.$$

As with other evaluation metrics, words that occur in the same sequence in both texts are scored more highly (the longer the co-occuring sequence, the higher the scoring bonus). In the bitext grid, these sequences are diagonally adjacent hits, as seen in Figure 2.3.

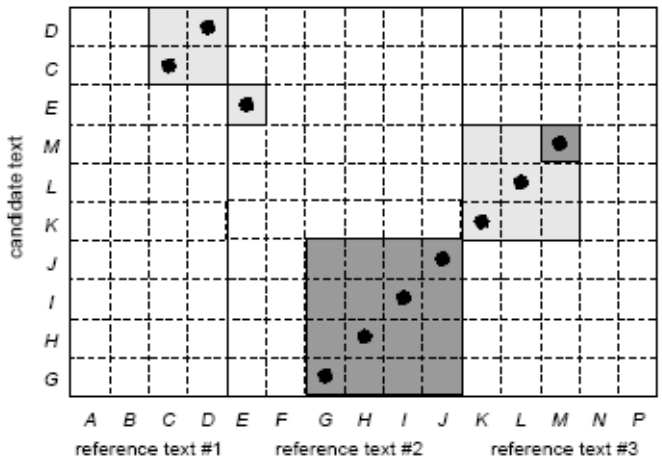

Figure 2.3: Bitext Grid with Multiple Reference Text and Co-occuring Word Matches from Melamed et al. (2003)

The authors show that their F-measure (combination Precision and Recall) scoring correlated more highly with adequacy than BLEU scores. However, their initial claim of producing a measure whose scores are more easily interpretable than BLEU scores is not supported in the paper.

2.2.2.6 Meteor

In an attempt to also address perceived issues with the BLEU metric, the METEOR metric was developed and tested for machine translation evaluation (Banerjee and Lavie, 2005). The authors state that recall measures obtain a higher correlation with human judgments than measures of precision, the basis for BLEU scoring (Lavie et al., 2004), and combination recall and precision measures obtain higher correlations than either alone. The METEOR metric builds upon the

success of base precision and recall measures (as seen in the first two equations in Section 2.2.2.5) and produces a score based on the harmonic mean of precision and recall (with more weight on recall). This measure, the Fmean (van Rijsbergen, 1979), for Meteor is computed as:

$$Fmean = \frac{10PR}{R + 9P}$$

Recall is more heavily weighted because the correlation of pure recall and human MT evaluation is much higher than that of precision and the equally weighted harmonic mean produces an even higher correlation. The authors show that the heavily recall-weighted harmonic mean Meteor scoring produces the highest correlations with human evaluation than an equally weighted harmonic mean or any of the other measures.

The METEOR metric provides flexibility in its unigram matching. The method incorporates a three-stage matching process. Stage 1 maps each candidate word with its exact reference match. Stage 2 incorporates a Porter Stemmer[7] (Porter, 1980), and matches the stemmed form of the candidate words with the stemmed form of the reference words. In Stage 3, a "WN synonymy" module is used to map a candidate and reference word if the are synonyms of each other. The METEOR package allows users to specify the order in which the stages are run or if stages are omitted, and the default order is as described here.

[7]The Porter Stemming Algorithm is a process for removing the common morphological and inflexional endings from words in English. Its main use is as part of a term normalisation process that is usually done when setting up Information Retrieval systems (Porter, 2006).

METEOR, like other evaluation metrics, incorporates a method to penalize shorter n-gram matches (some of the other methods offer a 'reward' for longer n-gram matches). The matching unigrams in the candidate translation/summary are grouped into chunks, with each chunk containing adjacent terms that exact match the ordering of terms in the reference translation/summary (discovery of n-gram matches). Longer n-grams produce fewer total chunks, and if the candidate translation/summary and the reference translation/summary exactly match, then only one chunk is produced. The penalty is then computed as:

$$Penalty = 0.5 \times \left(\frac{\text{number of chunks}}{\text{number of unigrams matched}} \right)^3$$

Thus, with the combination of the harmonic mean (Fmean) and penalty equations, METEOR scores are calculated as:

$$Score = Fmean \times (1 - Penalty)$$

A shortcoming of the method is seen in cases where more than one reference translation/summary is utilized. Instead of using a combinatory or averaging technique to produce scores based on comparison with all three references, the candidate translation/summary is scored against each reference individually and only the highest score is used.

2.2.2.7 (H)TER

Recently the GALE (Global Autonomous Language Exploitation) research program introduced a new method for machine translation evaluation called Trans-

lation Error Rate (TER). TER was originally designed to count the number of edits (including phrasal shifts) performed by a human to change a hypothesis so that it is both fluent and has the correct meaning. This was then decomposed into two steps: defining a new reference and finding the minimum number of edits so that the hypothesis exactly matches one of the references. This method is semi-automatic in that it requires human annotation for scoring, and it is also expensive, in that it requires approximately 3 to 7 minutes per sentence for the annotation.

TER is defined as the minimum number of edits needed to change a hypothesis (candidate translation) so that it exactly matches one of the reference translations, normalized by the average length of the references. Since the concern is the minimum number of edits needed to modify the hypothesis, only the number of edits to the closest reference is measured (by the TER score). Specifically:

$$\text{TER} = \frac{\text{\# of edits}}{\text{average \# of reference words}}$$

Possible edits include the insertion, deletion, and substitution of single words as well as shifts of word sequences. A shift moves a contiguous sequence of words within the hypothesis to another location within the hypothesis. All edits, including shifts of any number of words, by any distance, have equal cost. In addition, punctuation tokens are treated as normal words and mis-capitalization is counted as an edit.

The Human-targeted Translation Error Rate (HTER) involves a procedure for creating targeted references. In order to accurately measure the number of edits

necessary to transform the hypothesis into a fluent target language (often English) sentence with the same meaning as the references, one must do more than measure the distance between the hypothesis and the current references. Specifically, a more successful approach is one that finds the closest possible reference to the hypothesis from the space of all possible fluent references that have the same meaning as the original references.

To approximate this, human annotators who are fluent speakers of the target language are used to generate a new targeted reference. This process is started with automatic system output (hypothesis) and one or more pre-determined, or *untargeted*, reference translations. They could generate the targeted reference by editing the system hypothesis or the original reference translation. It is found that most editors edit the hypothesis until it is fluent and has the same meaning as the untargeted reference(s). The minimum TER is computed using this single targeted reference as a new human reference. The targeted reference is the only human reference used for the purpose of measuring HTER. However, this reference is not used for computing the average reference length.[8]

HTER has been shown in studies to correlate more highly with human judgments than the BLEU, and METEOR metrics.

[8]The targeted reference is not used to compute the average reference length, as this would change the denominator in the TER calculation, and crafty annotators could favor long targeted references in order to minimize HTER.

2.2.2.8 Pourpre

In addition to the above metrics for summarization and machine translation, a new metric, Pourpre (Lin and Demner-Fushman, 2005) has been suggested for question-answering task evaluations. Question-answering tasks concentrate on determining whether a specific (candidate) response to a presented question contains information representing the correct answer to the question. The tasks are more closely related to the relevance assessment task of text summarization evaluation than machine translation tasks. Therefore, it is possible for question-answering metrics to likewise be used for text summarization evaluation and this possibility has now been suggested in the context of the GALE Distillation initiative (DARPA GALE BAA, 2005).

POURPRE is a technique for automatically evaluating answers to definition questions (Lin and Demner-Fushman, 2005) based on n-gram co-occurences like the BLEU and ROUGE metrics.

Definition questions are slightly different from factoid questions that were previously the focus of question answering tasks. Factoid questions would include "What city is the capital of New York?" or "What is the name of the 40th President of the United States?" where definition questions could include "Who is Bill Clinton?" The answers to factoid question could be singular names, places or very short, specifically defined responses (typically noun phrases), while the answers to definition questions could be what the authors term "nuggets" of information;

relevant information about the entity defined in the question. The "nuggets" used as the "answer key" to the questions are produced by a human assessor from research done during the original creation of the questions and from a compiled list of all the output produced by the question answering systems. A human assessor uses the nuggets in the answer key in comparison against the output of a question answering system to determine whether the important nuggets are contained within the system response.

Unlike ROUGE, unigram matching is preferred over bigram or longer n-gram matches with POURPRE in that the authors believe that longer n-grams are related more to the fluency of the candidate responses which would be important in machine translation or text summarization, but is less important for answers to definition questions. For scoring, POURPRE matches nuggets by summing unigram co-occurences between the (reference) nuggets and the candidate response. POURPRE also uses a harmonic mean of precision and recall for their scoring (and like METEOR, recall is weighed more heavily than precision).

Let

r # of *vital* nuggets returned in a response
a # of *okay* nuggets returned in a response
R # of *vital* nuggets in the answer key
l # of non-whitespace characters in the entire answer string

Then

$$\text{recall}(\mathcal{R}) = r/R$$

$$\text{allowance}(\alpha) = 100 \times (r + a)$$

$$\text{precision}(\mathcal{P}) = \begin{cases} 1 & \text{if } l < \alpha \\ 1 - \frac{l-\alpha}{l} & \text{otherwise} \end{cases}$$

Finally, the $F(\beta) = \frac{(\beta^2+1)\times\mathcal{P}\times\mathcal{R}}{\beta^2\times\mathcal{P}+\mathcal{R}}$
$\beta = 5$ in TREC 2003, $\beta = 3$ in TREC 2004.

Official definition of F-measure from Lin and Demner-Fushman (2005)

POURPRE calculates the F-measure using the sum of the match scores for the nuggets divided by the total number of nuggets for nugget recall. They also allow alternatives when some of the reference nuggets are deemed more important than others; listing the preferred nuggets as "vital" versus "okay" for the information that is relevant to answering the question but is not the most critical information. Incorporation of "vital" and "okay" terms change the scoring mechanism such that the recall only counts matches for vital information. Finally, POURPRE incorporates inverse document frequency[9] (*idf*) sums as replacements

[9]Inverse document frequency (*idf*) is a commonly used measure in information retrieval based on the observation that the more specific, i.e., low-frequency terms are likely to be of particular

for the match score. *idf* is defined as $log(\frac{N}{c_i})$, where N is the number of documents in the collection and c_i is the number of documents within that set that contain the term t_i. The match score of a particular nugget becomes the sum of the *idf*s of matching terms in the candidate response divided by the sum of all term *idf*s in the reference nugget.

2.2.3 Human Extrinsic Evaluations

Common human extrinsic tasks are question-answering, information retrieval, and relevance assessments. In selecting the extrinsic task it is important that the task be unambiguous enough that users can perform it with a high level of agreement. If the task is so difficult that users cannot perform it with a high level of agreement—even when they are shown the entire document—it will not be possible to detect significant differences among different summarization methods because the amount of variation due to noise will overshadow the variation due to summarization method.

Relevance assessments are often used as an extrinsic task-based evaluation method and and can be equated to the real-world task of web searching and information retrieval. Relevance assessment tasks measure the impact of summarization on determining the relevance of a document to a topic (Brandow et al., 1995; Jing

importance in identifying relevant material. The number of documents relevant to a query is generally small, frequently occurring terms occur in many irrelevant documents; infrequently occurring terms have a greater probability of occurring in relevant documents (Jones, 1980).

et al., 1998; Tombros and Sanderson, 1998). These tasks can be executed in numerous ways. In one study (Tombros and Sanderson, 1998), participants were given five minutes to find as many relevant documents as possible for a query. Another type of relevance assessment task requires users to determine whether a given document, based on a summary of the document or the full text, is related to a specified event or topic.[10]

In the relevance assessment task, a user is given a topic or event description and has to judge whether or not a document is related to the specified topic/event based solely on the provided summary or the entire text. The base agreement measure, accuracy, is the most commonly used measure for analysis of relevance assessment tasks. To produce the accuracy measure, human judgments were usually compared to a *gold standard* judgment to produce a measure of the quality of the summary or summarizing system. Higher agreement percentages were supposed to denote a better quality summary.

Chapter 3 introduces a new method of comparison called **Relevance Prediction** that compares a human's judgment on a summary with his or her own judgment on the full text document instead of relying on external *gold standard* judgments. The chapter also describes additional measures from the information retrieval and signal detection fields that can be produced with the Relevance-Prediction and LDC-Agreement methods and used to provide additional state-

[10]A topic is an event or activity, along with all other related events or activities. An event is something that happens at some specific time or place, and the unavoidable consequences.

ments about the trends of task data.

The Relevance-Prediction approach addresses some of the shortcomings of the SUMMAC studies (Mani et al., 2002) in that the use of user-centric judgments—rather than an external gold standard—yields higher agreement rates. In addition, the goals of this work are broader than those of the SUMMAC studies, where the focus was on extrinsic evaluations: here, both extrinsic and intrinsic measures are explored to determine whether there is a correlation between them.

2.2.4 Automatic Extrinsic Evaluations

Currently, there are no automatic extrinsic evaluators for text summarization, but systems can be designed that judge summarizers based on their ability to allow the completion of tasks such as question-answering or categorization. For question-answering,[11] an automatic system would search a summary for answers to specified questions. Since the answers would be present in the source document, this task would involve determining whether the summary retained the important information from the source that could help a user complete this task. Similarly, an automatic system can be used to complete a categorization task, to determine

[11]Latent Semantic Analysis (LSA) (Landauer et al., 1998) has been used in the general question-answering domain to determine how well students learn information and correctly answer questions (Kanejiya et al., 2003), but has not specifically been used for task-based evaluation of text summarization. Therefore, LSA could be considered as an automatic extrinsic evaluator for general question-answering tasks.

how well a summary can help categorize a document into a set of topics (Jing et al., 1998). The automatic system may search the summary for topical keywords or clues to then make the topic categorization or association.

2.3 Summary

This chapter described numerous intrinsic and extrinsic metrics that have been created for use in text summarization evaluations. Newly proposed methods build upon the successes and shortcomings of previous methods and aim to be as reliable in measuring summary quality as humans. The BLEU and ROUGE methods are used as part of the correlation studies in Chapters 4 and 5. The Basic Elements method is evaluated along with the ROUGE method in the study of Chapter 6.

Relevance Assessment tasks have been used in many large-scale extrinsic evaluations, e.g., the Tipster SUMMAC evaluation (Mani et al., 2002) and the Document Understanding Conference (DUC) (Harman and Over, 2004). The usefulness of summaries for the task of relevance assessment is often assessed through gold-standard based judgments, and an existing gold-standard measure (LDC Agreement) is used as part of the experimental studies of Chapters 4 through 7. The next chapter describes the LDC-Agreement method in more detail, and introduces a new method, Relevance Prediction, that is shown to be a more reliable measure of individual task performance than LDC Agreement.

Chapter 3

Toward a New Agreement Measure: Relevance Prediction

In the past, human judgments in task-based evaluations were compared against *gold standards* to build a measure of agreement. Gold standards are thought to be the *correct* answers in reference to a specific task. In the case of relevance assessments, the gold standard judgments of *relevant* or *not relevant* are thought to reflect the true relevance level of the document. Agreement is measured by comparing the judgments made by users on a text to the gold standard judgment for the same text. For gold-standard based agreement, if a user makes a judgment on a summary consistent with the gold standard judgment this is thought to indicate that the summary is good in that it gave the users enough information to make the *correct* judgment. If a user makes a judgment on a summary that is inconsistent with the gold standard, this is thought to be an indicator of a low-quality summary that did not provide the user with the most salient informa-

tion or that provided the user with too little information and led to an *incorrect* judgment.

One variant of gold-standard judgment, LDC Agreement, uses LDC-commissioned judgments for relevance assessment (see Section 3.1). However, this thesis argues that gold-standards are unreliable and, as stated in other work, (Edmundson, 1969; Paice, 1990; Hand, 1997; Jing et al., 1998; Ahmad et al., 2003), there is no 'correct' judgment—judgments of relevance vary and are based on each user's beliefs. Therefore, a new measure, Relevance Prediction is proposed in Section 3.2. This measure assesses relevance based on each user's own judgments. In experiments described in Chapters 5, 6, and 7, LDC Agreement and Relevance Prediction are compared for their correlation with human judgments.

In the sections below, LDC Agreement and Relevance Prediction will be examined in more detail, common alternative measures to agreement will be introduced, and the issue of Agreement Measure Validation will be discussed.

3.1 LDC Agreement

The University of Pennsylvania's Linguistic Data Consortium (LDC) is an open consortium of universities, companies and government research laboratories whose goal is to create, collect and distribute speech and text databases, lexicons, and other resources for research and development purposes (LDC, 2006). The LDC trains their employees in a variety of corpus data annotation tasks. With the Topic

Detection and Tracking version 3 (TDT-3) corpus, the trained annotators judged all of the documents as relevant or not relevant to a list of topics and/or events. These document annotations were intended as the correct relevance representation of each of the individual documents. Other researchers and institutions could then use the documents contained within the corpus for relevance assessment tasks, and compare the results of the users to that of the LDC annotators.

LDC Agreement compares the gold-standard judgments produced by the LDC annotators with the judgments made by each individual user. The user's judgment is assigned a value of *1* if it equals the judgments made by the LDC annotators, and a value of *0* if they do not match. Because the LDC judgments are considered "correct," they are considered the "gold standard" against which other judgments should be compared. Futhermore, it is thought that if a summary gives a user enough information to make the "correct" judgment (the judgment consistent with the gold-standard), then it is a good summary. Likewise, if the summary does not give enough information to make the "correct" judgment, then it is a bad summary.

An issue with the LDC-Agreement method is the use of external gold-standard judgments and the resulting low interannotator agreement rates as seen in the LDC General and LDC Event Tracking experiments (described in detail in Sections 4.1 and 4.2). The way agreement is measured with the LDC-Agreement method may be useful in some contexts, e.g., for group-oriented or group-consensus tasks. However, this work focuses on tasks tailored for the individual user, specifically web

browsing and information retrieval tasks. Thus, a new method that eliminates external gold-standard judgments and is thought to be more reliable for evaluation of browsing and retrieval tasks than the LDC-Agreement method is proposed in the next section. In Section 3.4, factors that affect human judgment are identified and problems associated with external gold-standard measurements are described.

3.2 Relevance Prediction

I define an alternative to LDC Agreement—an extrinsic measure called *Relevance Prediction*—where each user builds their own "gold standard" based on the full-text documents. Agreement is measured by comparing a user's surrogate-based judgment against his/her own judgment on the corresponding text. If a user makes a judgment on a summary consistent with the judgment made on a corresponding full text document, this signifies that the summary has provided enough information to make a reliable judgment. Therefore, the summary should receive a high score. If the user makes a judgment on a summary that is inconsistent with the full text judgment, this implies that the summary is lacking in some way; that it did not provide key information to make a reliable judgment, and that it should receive a low score.

To calculate the Relevance-Prediction score, a user's judgment is assigned a value of *1* if his/her surrogate judgment is the same as the corresponding full-text judgment, and *0* otherwise. These values are summed over all judgments

for a surrogate type and are divided by the total number of judgments for that surrogate type to determine the effectiveness of the associated summary method.

Formally, given a summary/document pair (s, d), if users make the same judgment on s that they did on d, we say $j(s, d) = 1$. If users change their judgment between s and d, we say $j(s, d) = 0$. Given a set of summary/document pairs DS_i associated with event i, the Relevance-Prediction score is computed as follows:[1]

$$\textit{Relevance-Prediction}(i) = \frac{\sum_{s,d \in DS_i} j(s, d)}{|DS_i|}$$

In the experiments discussed in Chapter 4 through Chapter 7, users make relevance judgments on a subset of all the summaries produced a given system and then they make judgments on the corresponding full texts. This ordering ensures that the user does not make a judgment on an individual summary immediately before seeing the corresponding document. In cases where more than one summary system is used, the users make judgments on a subset of summaries produced by a given system prior to judging the summaries produced by another system until all summary systems are exhausted, prior to judging the corresponding texts.[2]

[1]This definition for Relevance Prediction is described for use with binary judgments, as seen in the single-document relevance assessment tasks of Chapters 4 through 6. The judgments in the multi-document relevance assessment task of Chapter 7 are non-binary and a description of their use with Relevance Prediction is presented in that chapter.

[2]In the tasks, users make judgments on hundreds of summaries and documents and the time delays are great enough that the user is unable to associate a specific summary with its corresponding document. This belief is detailed and tested in the Memory and Priming study in

The results of the experiments described in Chapters 5, 6, and 7 demonstrate that this approach yields a more reliable comparison mechanism than that of LDC Agreement because it does not rely on gold-standard judgments provided by other individuals. Moreover, Relevance Prediction can be more helpful in illuminating the usefulness of summaries for a real-world scenario, e.g., a browsing environment, where credit is given when an individual user would choose (or reject) a document under both conditions.

The studies in this research will focus primarily on the Accuracy (agreement) measure produced with the Relevance-Prediction and LDC-Agreement methods. However, other measures are sometimes used in analyzing experimental results including recall, precision and f-score from information retrieval, and sensitivity and specificity from Signal Detection Theory. These measures will be discussed in more detail in the section below, and used in the experiments of Chapters 4 through 7.

3.3 Other Measures

Although accuracy is the primary metric used to investigate the correlations between intrinsic and extrinsic measures in Chapters 4 through 7, other measures are also discussed and, in some cases, used in later experiments. Signal Detection Theory (SDT) is a method used to study decision making under uncertainty. Rel-

Section 4.3.

	Judged Relevant	**Judged Not-Relevant**
Relevant is True	TP	FN
Relevant is False	FP	TN

Table 3.1: Contingency Table for Extrinsic Task

evance assessments fall into this category in that they are decision making tasks that involve humans making judgments under uncertain conditions. The measures of specificity and sensitivity are used in Signal Detection to analyze uncertain judgments. These measures and SDT will be described in more detail in Section 3.3.2.

Relevance assessments are also similar to the real-world task of browsing information retrieval. Therefore, measures from the IR field are also incorporated in the analysis and are introduced in the next section.

3.3.1 Information Retrieval Measures

In the SUMMAC study (Mani et al., 2002), in addition to the accuracy measure the IR measures of precision, recall, and f-score are used in the analysis of the experimental results. Following the lead of the SUMMAC researchers, these additional measures are also included in the analysis of the experiments in Chapters 4 through 7.

The contingency table for the extrinsic task is shown in Table 3.1, where **TP** (*true positives*), **TN** (*true negatives*), **FP** (*false positives*), and **FN** (*false negatives*) are taken as percentage of totals observed in all four categories.

Using this contingency table, the extrinsic measures are given here:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F\text{-}Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

3.3.2 Signal Detection Theory

Signal Detection Theory is a method of studying decision making in situations where uncertainty exists. This is applicable to relevance assessments in that a user has to make a decision about the relevance of a summary to a specified topic or event while being unsure of the contents and main focus of the source document. The only "clues" to the information contained in the document is the summary which, through compression, has lost a lot of the information that a user will have to mentally guess at or reconstruct. In Signal Detection Theory, the "truth" is thought to be whether or not there is a signal, and the judgment made by the user is compared to this "truth." In this work, the "truth" or presence of a signal is whether or not the full text document is considered to be relevant to the specified topic/event. For LDC Agreement, this truth is given by the external gold standard

judgments. For Relevance Prediction, this truth is given by the judgment a user makes on the full text document.

Signal Detection Theory also enables the analysis of decision making tasks, in this case, the relevance assessment task. Similar to information retrieval metrics, the metrics used in Signal Detection Theory are described below. The metrics are given as the judgment on the summary divided by the measure of "truth" for LDC Agreement and Relevance Prediction (described in the paragraph above), where "rel" corresponds to a "relevant" judgment, and "notrel" corresponds to a "not relevant judgment":

$$Hit = \frac{rel}{rel} = True\ Positive\ (TP)$$

$$Miss = \frac{notrel}{rel} = False\ Negative\ (FN)$$

$$False\ Alarm = \frac{rel}{notrel} = False\ Positve\ (FP)$$

$$Correct\ Rejection = \frac{notrel}{notrel} = True\ Negative\ (TN)$$

This is equivalent to the contingency table used in information retrieval, seen in Table 3.1.

For Signal Detection Theory, the extrinsic measures are sensitivity and specifity defined as:

$$sensitivity = \frac{total\ hits}{total\ hits + total\ misses} = \frac{TP}{TP + FN}$$

$$specificity = \frac{total\ false\ alarms}{total\ false\ alarms + total\ correct\ rejections} = \frac{FP}{FP + TN}$$

The sensitivity and specificity measures can be used to approximate how well a person can identify the relevance of a document. This measure is given by the discriminability index, d'. When the distributions for the signal (representing the *relevant* cases) and noise (representing the *not relevant* cases) are overlapping equal-variance normal distributions (consistent with the experimental data described in this thesis), d' is defined as:

$$d' = Z_{(1-specificity)} - Z_{sensitivity}$$

For relevance assessments, the d' value for a given system will estimate how well that system allowed the users to distinguish between the relevant and not relevant documents. A system with a higher d' value means it is better at helping the user correctly identify relevant documents than a system with a lower d' value.

The full set of metrics used in the experimental studies are accuracy, precision, f-score, recall/sensitivity, specificity, d', and time (the average amount of seconds it takes each user to make judgments for a specific system). Although the metrics used in the studies were inspired by the SUMMAC study, an issue with SUMMAC was the low agreement scores, associated with the LDC-Agreement measurement. In developing the Relevance-Prediction measure, it was surmised that the LDC-Agreement method produced low scores because all users cannot be held to a single standard; the relevance level of a document is determined by each

individual user's beliefs. One must then wonder why users judge documents differently, or more specifically, what factors affect a user's judgment. Some background on human judgments and factors that could influence the variance of judgments are described in the next section.

3.4 Agreement Measure Validation

For the relevance assessment task, it is important to note that there is no right or wrong answer. Whether a document is relevant to a specified topic or event is central to each individual's beliefs. Gold-standard based measures try to impose 'correct' answers and judge the performance of other individuals by those criteria. A key factor in the creation of the Relevance-Prediction method is the accommodation of the variance of human judgments.

Relevance Assessments are decision-making tasks in which items are classified into one of two categories—relevant to the topic or not-relevant to the topic—based on the information present and personal beliefs. Important to the relevance assessment task is the notion of inference. Since summaries are significantly shorter than the full text documents, users should be able to infer information about the full text from just a few words in the summary.

The relevance assessment task also shares properties of simple categorization tasks studied by a number of researchers (Sloutsky and Fisher, 2004). In a simple categorization task, a child is told to put an animal in one of two categories, such

as *bird* or *mammal*. The child uses a number of clues to achieve this: whether or not the animal has feathers, wings, lays eggs, or has hair. For relevance assessment, one would also use clues such as whether words or similar concepts from the event or topic description appear in the summary or full text document. These clues can reflect reference words or words within the summary that exactly match words present in the document itself (indicating more of an extractive summary).

There are also cases where abstractive summaries are used which do not include words found in the document but which do include similar words or synonyms of words from the document (or are somehow able to convey some level of meaning of the document). One would expect inferences made from extractive summaries to be more accurate than inferences made from abstractive summaries. The performance may be the same, but the psychological inferences about the document's contents are likely to be more accurate if one assumes that the user tries to determine the informational content of the full text document by reconstructing the inferred text around the information and words presented in the summary. If this is the case, automatic summaries—which are usually extractive—should help a user make better relevance judgments than they would be with abstractive, human-generated summaries. However, as will be discussed in more depth in Chapter 4, this is not yet the case. Various factors influence the variance in human judgment on the relevance assessment tasks. These factors are discussed below.

Prior Knowledge – In work by Goldstein and Hogarth (1997), it is stated that users rely on prior knowledge to guide the encoding, organization, and ma-

nipulation of information. Therefore, the prior knowledge that each individual has about a specific event or topic can affect his or her relevance decisions for that event or topic. An example of this can be imagined with a summary "the collapse of the Alfred P. Murrah building" being judged against the "The Oklahoma City bombing" event. Users that are very familiar with the details of the Oklahoma city bombing would probably know that the Alfred P. Murrah building was the target of the bombing, and would likely mark this document as "relevant" to the event. However, users that are unfamiliar with that particular event may not know details about the bombing, nor its association with the "Alfred P. Murrah building," and would likely mark this document as "not relevant."

Topic Difficulty – Topics can be perceived by users as difficult if they have no prior knowledge of the event, or in cases where the topic or event description lists items that are not easily reflected in the summary or document. An example of this would include a topic or event description "The Palestinian Government given new powers and responsibility" and a document or summary beginning with:

> "The Israel Declaration of Principles (the DOP), provided for a five year transitional period of self-rule in the Gaza Strip and the Jericho Area and in additional areas of the West Bank pursuant to the Israel-PLO 28 September 1995 Interim Agreement, the Israel-PLO 15 January 1997 Protocol Concerning Redeployment in Hebron, the Israel-PLO 23 October 1998 Wye River Memorandum, and the 4 September 1999 Sharm el-Sheikh Agreement. The DOP provides that Israel will retain

> responsibility during the transitional period for external and internal security and for public order of settlements and Israeli citizens. Direct negotiations to determine the permanent status of Gaza and West Bank began in September 1999 after a three-year hiatus, but were derailed by a second intifadah that broke out in September 2000."

If a user is not familiar with the geographic locations of Israel, Palestine, the Gaza Strip and the West Bank, or the details of the Israel DOP, he or she may feel uncertain about making judgments for this topic and may conclude that the decision-making task is harder than that of other topics. The difficulty of topics and the manipulation of task complexity is known to affect other factors such as attention, accuracy, and the time needed to complete a trial (Gonzalez, 2005) and therefore is considered a factor that contributes to variance in judgments.

Saliency of Information – The most important part of text summarization is determining what information in a text document is the most important. For human-generated text summaries, people would read or skim a text document, find the sentences or concepts that are central to the document's meaning and either use these exact terms to produce an extractive summary, or re-write the identified information and produce an abstractive summary. Of course, the key to this is the determination of the most important concepts. In a study by Salton et al. (1997), two users were asked to identify the most important paragraphs within a text document and to use these as the basis of their summary of the document. The authors planned to compare the automatically generated summaries against the

human-generated summaries to evaluate the quality of the automatic summarizers. However, they found that the humans often did not agree on which paragraphs were most important—the agreement rate was less than 50%.[3] The implication is that a person reading a text document or news article will consider certain sentences or concepts to be most important to him or her, while another user may find a different set of sentences or concepts to be most important. Since summaries omit information that is present in the source document, a summarizer may extract the sentences or concepts that are deemed important by one person but not those deemed important another.

Imagine that two people, A and B, consider a specific text document to be relevant to a topic/event description. Given a task to choose the most important information in the text, as described above, A and B may choose different concepts or sentences to be indicative of what counts as important. If a summarizer contains the information important to A but not to B, A may mark that summary as *relevant* to the topic/event, while B marks the same summary as *not relevant*. Therefore, A would determine that the summarizer produced a good summary, while B would think otherwise. If we then imagine that A is our experimental participant and B is an LDC annotator (or vice-versa), the LDC-Agreement method would assign the summarizer a score of 0. This would not accurately reflect that person A liked the summarizer's output. However, Relevance-Prediction scoring

[3]The agreement rate was 46%, i.e., an extract generated by one user was likely to cover 46% of the information regarded to be most important by the other user (Salton et al., 1997).

would compare the judgments of each person on the summary with his/her own judgments on the document (producing a score of 1 for person A and 0 for person 2) and produce a 50% score for the system, reflective of the fact that one person liked the summarizer, while the other did not.

Employing Heuristics and Cognitive Biases – In relevance assessment tasks, some users may develop heuristics for their judgments, i.e., if the summary contains the specific words from the topic or event description, then they consider it to be relevant; otherwise they consider it not to be relevant. The heuristics that individuals create or use may lead way to personal biases for their relevance decisions. For decision making, participants try to comprehend the summary or document, and then "accept" or "reject" it, in terms of relevance to the topic or event (Descartes, 1984; Mutz and Chanin, 2004). If a summary contains information that would suggest that it is relevant to the topic or event, but is viewed as incoherent or does not seem to be fluent, a participant can decide that this is "not relevant" to the specified topic or event. This would reflect a bias of the participant towards very coherent and fluent summaries. Although the coherence and fluency of a summary would usually influence the perception of its quality, in the case of relevance assessments, users are instructed to base their determination of relevance on the presence of information related to the topic or event description—not on factors pertaining to coherence or fluency (see Appendices A and C).

Other biases are possible, such as ones based on "anchoring" heuristics, where a participant may rely on a single piece of information too much for their decisions.

An example of this can be seen if a participant marks any summary containing the word "Oklahoma" as relevant to the event description "Oklahoma City Bombing trial."

The cognitive factors listed above are the bases for individual-level differences in relevance judgment and perception of the relevance assessment task. Since the LDC-Agreement method compares the judgments of one participant against judgments of another as the basis for scoring, the method is not sensitive to the individual differences and does not produce scores that are reflective of each user's preferences. However, by comparing each individual's summary judgments against his/her **own** judgments on the corresponding full text document, the Relevance-Prediction method is sensitive to the individual differences, and therefore produces a more reliable evaluation method that is more consistent with the individual preferences.

Chapter 4

Initial Studies: Correlation of Intrinsic and Extrinsic Measures

This chapter describes the first three experiments investigating the level of correlation of the various intrinsic measures (i.e. BLEU, ROUGE, BE) to human performance on a document relevance assessment task. The findings from these first three studies have encouraged modifications to the design, hypotheses and methods of the subsequent experiments, to be described in Chapters 5, 6, and 7.

The three experiments discussed in this chapter are referred to as LDC General, LDC Event Tracking and Memory and Priming. These experiments examine the use of the LDC-Agreement method in determining the effectiveness of summaries with respect to relevance assessment tasks. The existence of correlations between automatic metrics and human task performance with LDC Agreement is also examined. Details of each experiment are given below.

4.1 LDC General: Correlation of BLEU and ROUGE and Extrinsic Task Performance

This initial experiment, LDC General, investigates the correlation of BLEU and ROUGE scoring for summaries with the performance of humans on an extrinsic relevance assessment task using the agreement method based on LDC's human judgments. The goal is to determine if a correlation exists and to study how different types of summaries affect human performance.

4.1.1 Hypotheses

The main hypothesis is that the full text would be an upper bound on performance because the summaries represent compressed data and omit a great deal of information that was present in the source document. The omission of data would result in lower precision and recall scores than that of the uncompressed source document.

A second hypothesis is that, out of all the summaries, the human-generated summaries would perform the best. The human summarizers would know how to easily identify the most important information in a document, which is often a problem for automatic summarizers. Also, human generated summaries are often formatted in a more fluent and easily readable manner than automatic summaries.

A third hypothesis is that the Keyword in Context (KWIC) system would have the worst performance results because this system uses single topical-like key-

words and does not format the results as fluent sentences. Although the KWIC system can give users an idea of some of the topics in the text, it does not include the relationship between the identified topics or keywords, nor identifies one keyword or topic as the main focus of the text.

The fourth hypothesis is that one or both of the intrinsic metrics would generate a high (>0.8) positive correlation with a measure of human performance. A primary goal of this experiment is to determine whether correlations exist between intrinsic and extrinsic measures. The determination of the level of correlation is expected to aid the summarization community in identifying an intrinsic measure for evaluation rather than the using the current method of laborious and costly human assessments.

For this experiment, the lengths of the summaries are about an order of magnitude shorter than the length of the full text document. Therefore, the final hypothesis is that the time for making judgments on summaries would be an order of magnitude faster than judgments on the full text documents—the reduction in time being the main possible benefit for summarization.

4.1.2 Experiment Details

The experiment uses four types of automatically generated document surrogates; two types of manually generated surrogates; and, as a control, the entire document. The automatically generated surrogates are:

- **HMM** – a statistical summarization system developed by UMD and BBN;
- **Trimmer** – Fluent headline based on a linguistically-motivated parse-and-trim approach (Dorr et al., 2003);
- **ISI Keywords** (ISIKWD) – Topic independent keyword summary (Hovy and Lin, 1997);
- **Keywords in Context** (KWIC)– two 10-word selections containing query words from the document. This KWIC system was developed by Jun Luo at the University of Maryland as part of the MIRACLE system.

The manual surrogates are:

- **Headline** – the original human-generated headline associated with the source document;
- **Human** – a generic summary written by a human, in the range of 10-15 words. These summaries were commissioned from University of Maryland students for use in this experiment.

Finally, the "Full Text" was added to the experiment, for determining an upper-bound on extrinsic measures and a lower bound on the speed measurements. The average lengths of the surrogates and Full Text used are shown in Table 4.1.

The National Institute of Standards and Technology (NIST) has provided 16 topics and a search set of 50 documents for each topic, including human relevance

System	Avg Word Count	Avg Char Count
HMM	14.76	88
Trimmer	15.18	97
ISIKWD	9.99	71
KWIC	19.32	107
Headline	13.14	73
Human Summary	12.15	76
Full Text	1232.54	5561

Table 4.1: Experiment 1: Average Word and Character Counts for Each Surrogate

assessments (produced by LDC) for each document with respect to its topic. Because the automatic summarization systems were designed for prose articles, the experiments described herein are limited to this type of input, thus reducing the set of viable topics and documents. In particular, (non-prose) transcripts and tables of content have been eliminated. Lengthy documents, e.g., treaties, have been eliminated to attempt to limit the amount of time required from participants to a reasonable duration. Finally, the document set is reduced further so as to induce a comparable proportion of relevant-to-non-relevant documents across topics. The total number of documents in the reduced set is 20 per topic, using 14 topics. Within each topic, 6 documents are selected randomly from those assessed to be relevant and 14 documents are selected from those assessed to be non-relevant. (In two cases, this is not possible because they do not have 14 or more non-relevant documents: Topic 403: 8 relevant, 12 non-relevant; and Topic 415: 10 relevant and 10 non-relevant.)

4.1.3 Experiment Design

In this study, 5 undergraduate and 9 graduate students were recruited at the University of Maryland at College Park through posted experiment advertisements to participate in the experiment. Participants were asked to provide information about their educational background and experience (Appendix B). All participants had previous online search experience and their fields of study included physyics, biology, engineering, government, economics, communications, and psychology. The instructions for the task (taken from the TDT-3 corpus instruction set that were given to document annotators) are shown in Appendix C.

Each participant was asked to perform 14 document selection tasks. Each task consisted of reading a topic description and making relevance judgments about 20 documents with respect to the topic. Participants were allowed to choose among three relevance levels: *Highly Relevant, Somewhat Relevant* or *Not Relevant*.

The experiment required exactly one judgment per document. No time limit was imposed, however the participants were timed to determine how long it took them to make each judgment. Each participant saw each topic once, and each system twice. The order of presentation of the topics and systems was varied according to a Latin Square as seen in Table 4.2. Each of the 14 topics, T_1 through T_{14}, consists of 20 documents corresponding to one event. The fourteen human users were divided into seven user groups (A through G), each consisting of two users who saw the same two topics for each system (not necessarily in the

System	T_1T_2	T_3T_4	T_5T_6	T_7T_8	T_9T_{10}	$T_{11}T_{12}$	$T_{13}T_{14}$
Full Text	A	B	C	D	E	F	G
Headline	B	C	D	E	F	G	A
Human	C	D	E	F	G	A	B
HMM	D	E	F	G	A	B	C
Trimmer	E	F	G	A	B	C	D
ISIKWD	F	G	A	B	C	D	E
KWIC	G	A	B	C	D	E	F

Table 4.2: LDC General Latin Square Experiment Design

same order). By establishing these user groups, it was possible to collect data for an analysis of within-group judgment agreement.

This experimental design ensures that each user group (two participants) saw a distinct combination of system and event. The system/event pairs were presented in a random order (both across user groups and within user groups), to reduce the impact of topic-ordering and fatigue effects.

4.1.4 Results and Analysis

The relevance assessments were binary judgments (relevant, non-relevant) commissioned by LDC. Thus, it was necessary to map the three-way relevance judgments of the participants to the LDC judgments. In the following analysis, the term **strict relevance** indicates that a document is considered to have been judged relevant only if the participant selected highly relevant. The term **non-strict relevance** indicates that a document is considered to have been judged relevant if the participant selected highly or somewhat relevant. Thus, under strict relevance, a "somewhat relevant" judgment in this experiment would match a

"non-relevant" LDC judgment, whereas under non-strict relevance, it would match a "relevant" LDC judgment.

Table 4.3 shows **TP**, **FP**, **FN**, **TN**, **P**recision, **R**ecall, **F**-score, and **A**ccuracy for each of the seven systems using Strict Relevance. The values for Non-Strict Relevance are shown in Table 4.4. In addition, the tables give the average **T**(ime) it took users to make a judgment—in seconds per document—for each system. The rows are sorted by accuracy (the same as LDC Agreement), and the boldfaced text highlights the highest result for each column.

One-factor repeated-measures ANOVA (with 97 degrees of freedom) was computed to determine if the differences among the systems were statistically significant for the five measures: precision, recall, f-score, accuracy, and time. Each user saw each system twice during the experiment, so each sample consisted of a user's judgments on the 40 documents that comprised the two times the user saw the output of a particular system. Precision, recall, f-score, accuracy and time were calculated on each sample of 40 judgments.

The ANOVA test indicates that for Strict Relevance, the differences are significant for recall and time measures, with $p<0.01$. However, the differences are not significant for the measures of accuracy, precision or f-score. For Non-Strict Relevance, the ANOVA indicates that the differences are significant for recall, f-score and time measures, with $p<0.01$; the differences are not significant for the accuracy and precision measures.

The ANOVA test only determines if one pair of systems is significantly differ-

System	**TP**	**FP**	**FN**	**TN**	**A**	**P**	**R**	**F**	**T** (s)
Human	62	45	118	335	**0.709**	0.579	0.344	0.432	8.56
HMM	43	26	**137**	**354**	**0.709**	**0.623**	0.239	0.345	9.73
Headline	49	34	131	346	0.705	0.590	0.272	0.373	9.40
Full Text	**94**	**81**	86	299	0.702	0.537	**0.522**	**0.530**	**33.15**
ISIKWD	45	39	135	341	0.689	0.536	0.250	0.341	9.23
Trimmer	46	48	134	332	0.675	0.489	0.256	0.336	10.08
KWIC	51	57	129	323	0.668	0.472	0.283	0.354	10.91
HSD, p<0.05	–	–	–	–	0.070	0.247	0.142	0.143	4.086

Table 4.3: Results of Extrinsic Task Measures on Seven Systems with Strict Relevance, sorted by **A**ccuracy

System	**TP**	**FP**	**FN**	**TN**	**A**	**P**	**R**	**F**	**T** (s)
Full Text	**145**	**143**	35	237	**0.682**	**0.503**	**0.806**	**0.620**	**34.33**
Human	114	130	66	250	0.650	0.467	0.633	0.538	8.86
Headline	104	123	76	257	0.645	0.458	0.578	0.511	9.73
HMM	84	112	**96**	**268**	0.629	0.429	0.467	0.447	10.08
ISIKWD	106	140	74	240	0.618	0.431	0.589	0.498	9.56
KWIC	113	154	67	226	0.605	0.423	0.628	0.506	11.30
Trimmer	89	144	91	236	0.580	0.382	0.494	0.431	10.44
HSD, p<0.05	–	–	–	–	0.101	0.152	0.139	0.105	4.086

Table 4.4: Results of Extrinsic Task Measures on Seven Systems with Non-Strict Relevance, sorted by **A**ccuracy

ent. In order to determine exactly which pairs of system are significantly different, Tukey's Studentized Range criterion, called the Honestly Significant Difference (HSD) (for a description, see Hinton (1995)) is used. The HSD results are shown in the bottom row of Tables 4.3 and 4.4 with $p<0.05$.

If the difference in measures between two systems is greater then the HSD, then a significant difference between the systems can be claimed. For example, the automatic system with the highest accuracy for Strict Relevance is HMM (0.709) and the lowest was KWIC (0.668). The difference between them is 0.041, which is less than the HSD for accuracy (0.070), so a significant difference cannot be claimed between HMM and KWIC. A significant difference between automatic systems can only be claimed for the recall measure with Non-Strict Relevance (Table 4.4), between KWIC (0.628) and HMM (0.467) resulting in a difference of 0.161, which is greater than the recall HSD (0.139).

The automatic summarization systems were reanalyzed without the human-generated summaries or the full text to determine whether significant differences with $p<0.05$ could be claimed among the automatic systems using an ANOVA and the Tukey (HSD) Test. This analysis showed that no significant differences between the automatic systems were found with either test.

The HSD value at $p<0.05$ is 0.142 for recall and 0.143 for the f-score for Strict Relevance. This allows the automatic systems to be grouped into two sets, A and B, each of which contains members that are not significantly distinct according to the Tukey test. This is shown in Tables 4.5 and 4.6.

Full Text	A	
Human		B
Headline		B
HMM		B
ISIKWD		B
KWIC		B
Trimmer		B

Table 4.5: Equivalence Classes of Automatic Summarization Systems with respect to Recall for Strict Relevance

Full Text	A	
Human	A	B
Headline		B
HMM		B
ISIKWD		B
KWIC		B
Trimmer		B

Table 4.6: Equivalence Classes of Automatic Summarization Systems with respect to F-Score for Strict Relevance

For Non-Strict Relevance (as seen in Table 4.4), the HSD at $p<0.05$ is 0.142 for recall and 0.143 for the f-score. The systems are grouped into three non-distinct overlapping sets, A, B, and C, as shown in the equivalence class tables 4.7 and 4.8.

The results of the signal detection measures, sensitivity, specificity and d' are shown in Tables 4.9 and 4.10 for strict and non-strict relevance, respectively. The

Full Text	A		
Human		B	
Headline		B	
ISIKWD		B	C
KWIC		B	C
HMM			C
Trimmer			C

Table 4.7: Equivalence Classes of Automatic Summarization Systems with respect to Recall for Non-Strict Relevance

Full Text	A		
Human	A	B	
Headline		B	C
HMM		B	C
ISIKWD		B	C
KWIC		B	C
Trimmer			C

Table 4.8: Equivalence Classes of Automatic Summarization Systems with respect to F-Score for Non-Strict Relevance

System	**Sensitivity**	**Specificity**	**discriminability index (d')**
Human	**0.344**	0.882	**0.783**
HMM	0.239	**0.932**	0.778
Headline	0.272	0.911	0.738
ISIKWD	0.250	0.897	0.592
Trimmer	0.256	0.874	0.487
KWIC	0.283	0.850	0.463

Table 4.9: Results of the Signal Detection Measures Using Strict Relevance, sorted by d'

sensitivity measure is the probability that users judge summaries for a given system "relevant" when the ground truths (the LDC judgments for the corresponding documents) are also relevant. The specificity measure is the probability that users judge summaries for a given system "not relevant" when the ground truths are also not-relevant. Again, when the distributions for the relevant and not-relevant cases (based on the gold standard) are overlapping equal-variance normal distributions, the d' measure can be calculated and determines how well the summaries from a particular system allow the users to correctly differentiate between relevant and not-relevant documents.

Additional measures were also calculated:

System	**Sensitivity**	**Specificity**	**discriminability index (d')**
Human	**0.633**	0.658	**0.747**
Headline	0.578	0.676	0.654
KWIC	0.628	0.595	0.566
ISIKWD	0.589	0.632	0.561
HMM	0.467	**0.705**	0.456
Trimmer	0.494	0.621	0.294

Table 4.10: Results of the Signal Detection Measures Using Non-Strict Relevance, sorted by d'

- Kappa with respect to LDC Agreement (Accuracy). The Kappa score is calculated as:

$$\frac{P_A - P_E}{1 - P_E}$$

with P_A equaling the agreement, and P_E equaling the expected agreement by chance (Carletta, 1996) and (Eugenio and Glass, 2004). It is assumed that the expected agreement will be 0.7 because 30% of the documents are actually relevant, so if you were to guess non-relevant for each document, you would expect agreement of 0.7.

- Between-Participant Agreement:

$$\frac{\text{total number of times two participants made same judgment on same doc,sys}}{\text{total times two participants judged same doc,sys}}$$

- Kappa between participants. Again expected agreement of 0.7 is used.

System	LDC Agreement	Kappa wrt LDC	Between Participant Agreement	Kappa Between Participants
Human	**0.709**	**0.030**	0.835	0.450
HMM	**0.709**	**0.030**	**0.878**	**0.594**
Headline	0.705	0.017	0.821	0.403
Full Text	0.702	0.007	0.749	0.164
ISIKWD	0.689	-0.037	0.853	0.510
Trimmer	0.675	-0.083	0.853	0.510
KWIC	0.668	-0.107	0.760	0.200

Table 4.11: Results Using Strict Relevance, sorted by LDC Agreement (Accuracy)

System	LDC Agreement	Kappa wrt LDC	Between Participant Agreement	Kappa Between Participants
Full Text	**0.682**	**-0.060**	0.681	-0.063
Human	0.650	-0.167	**0.703**	0.008
Headline	0.645	-0.183	0.670	-0.099
HMM	0.629	-0.237	**0.703**	0.008
ISIKWD	0.618	-0.273	0.602	-0.326
KWIC	0.605	-0.317	0.656	-0.147
Trimmer	0.580	-0.400	0.699	**0.004**

Table 4.12: Results Using Non-Strict Relevance, sorted by LDC Agreement (Accuracy)

4.1.5 Discussion

The experiment above yielded a number of interesting results. First, the participants performed surprisingly poorly with respect to LDC Agreement (between 0.67 and 0.71), even when exposed to the Full Text document. Agreement scores for full text would be expected to be in the 80% range. Similarly, the resulting Kappa scores shown in Tables 4.11 and 4.12 were also low, indicating that most of the reported user agreement stemmed from chance. These results are consistent with the low scores of the Summac experiments (Mani et al., 2002) which reported an agreement range of 16% to 69%. As seen in Table 4.3, the Full Text system ranked in the middle for the measures of accuracy, and precision with Strict Relevance. This did not support the main hypothesis that the Full Text would provide an upper bound for all performance measures. The Full Text did perform the best of all the systems with Non-Strict Relevance (Table 4.4 supporting the main hypothesis). Also, for both Strict and Non-Strict Relevance (Tables 4.3 and 4.4), the full text produced substantially higher recall than the surrogates, because the task was to determine whether the document contained any information relevant to the topic, and a summary will, necessarily, omit some content.

Similar to the results of the Full Text system, the human-generated systems, Headline and Human, did not consistently generate the highest performance scores of the summaries with Strict Relevance for the accuracy, precision, recall and d' measures. The HMM system tied with the Human system for the highest accuracy

result, achieved a higher result than both the Headline and Human systems for precision, and ranked in between Human and Headline for the d' measure.

By contrast, the Non-Strict Relevance results indicated that the Headline and Human systems ranked highest for the accuracy, precision, f-score measures and d', supporting the second hypothesis.

The KWIC system had the lowest performance results with strict relevance for accuracy and precision (Table 4.3) and also for d' (Table 4.9). However, KWIC generated mid-range scores with non-strict relevance for recall and f-score (Table 4.4) and d' (Table 4.10), which does not support the third hypothesis that the KWIC system would perform the worst of all systems on the performance measures.

The inter-annotator agreement among the participants as seen in Tables 4.11 and 4.12 was higher with non-strict relevance than with strict relevance. However, the full text performed in the mid-range of the systems for both relevance levels (strict and non-strict). This also did not support the main hypothesis.

Tables 4.3 and 4.4 show that the full text documents were processed by participants at a substantially slower speed—although not as slow as was anticipated. This may indicate that the participants were very good at skimming documents quickly. Also the processing speed of the summaries is higher than the full text documents, but the speed improvement factor of approximately 3 seems low. Therefore, this did not support the fifth experimental hypothesis, that the speed on summaries would be an order of magnitude greater than on the full text.

It does not appear that any system is performing at the level of chance, yet the differences are not statistically significant for all the measures. Suppose the participants randomly selected relevant or non-relevant, accepting an average of 10 out of the 20 documents. Given that there were 6 relevant documents and 14 non-relevant, one would expect precision of 0.3 and recall of 0.5. However, it could be that there is so much noise inherent in the task of document selection that this experiment design was not adequate to detect any substantial differences among the systems. If human generic summaries are considered to be an upper-bound on usefulness for this task, then the automatic systems are not performing far below that upper-bound.

4.1.6 Automatic Intrinsic Evaluation

In contrast to SUMMAC which focused on an extrinsic task evaluation, the problem of intrinsic evaluation using automatic metrics has also been examined. BLEU (Papineni et al., 2002) and ROUGE (Lin and Hovy, 2003) will be used as intrinsic measures, because they are based directly on the output of the systems. Both ROUGE and BLEU require reference summaries for the input documents to the summarization systems. For an in-depth description of the intrinsic metrics, refer back to Section 2.2.1.

Three additional short human summaries were commissioned for use as references in the automatic testing. BLEU was used with 1-grams through 4-grams and the results are shown in Figure 4.1. Unsurprisingly, the human summary

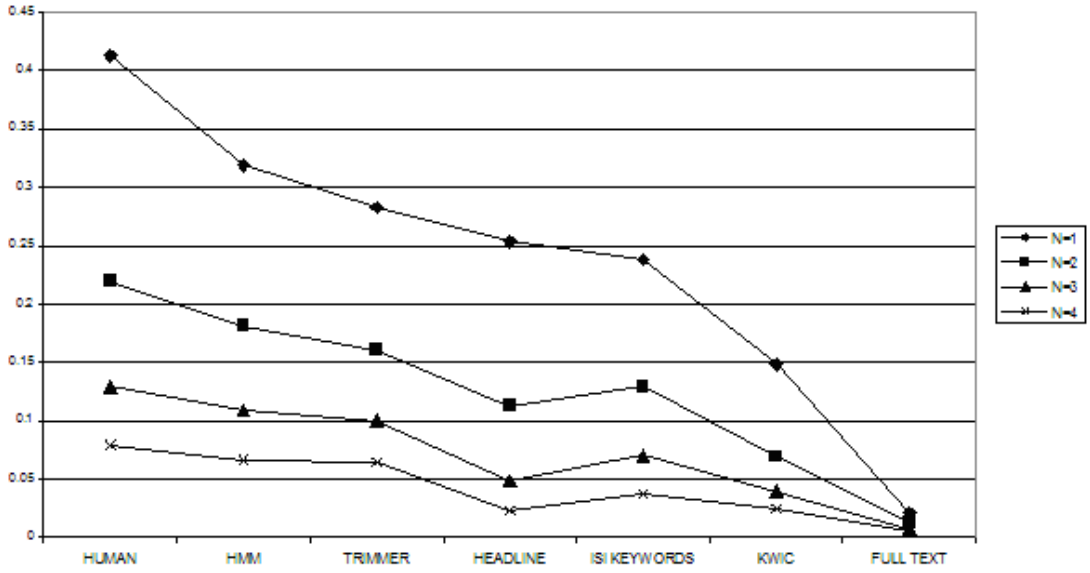

Figure 4.1: BLEU Scores

is automatically evaluated as being most like the reference summaries. The Full Text document, although probably most useful for evaluating relevance, scores very poorly by this automatic metric because so much of the content of the document does not appear in the summaries.

Using ROUGE scoring as seen in Figure 4.2, the entire document scores highest at all values of N, and differences among the summary systems is not very pronounced.

The scores of both the BLEU and ROUGE metrics are shown in Table 4.13. The ANOVA test was performed to determine if there are differences between the systems for each intrinsic evaluation method. The test did not show statistically significant differences with all systems included or with the exclusion of the three

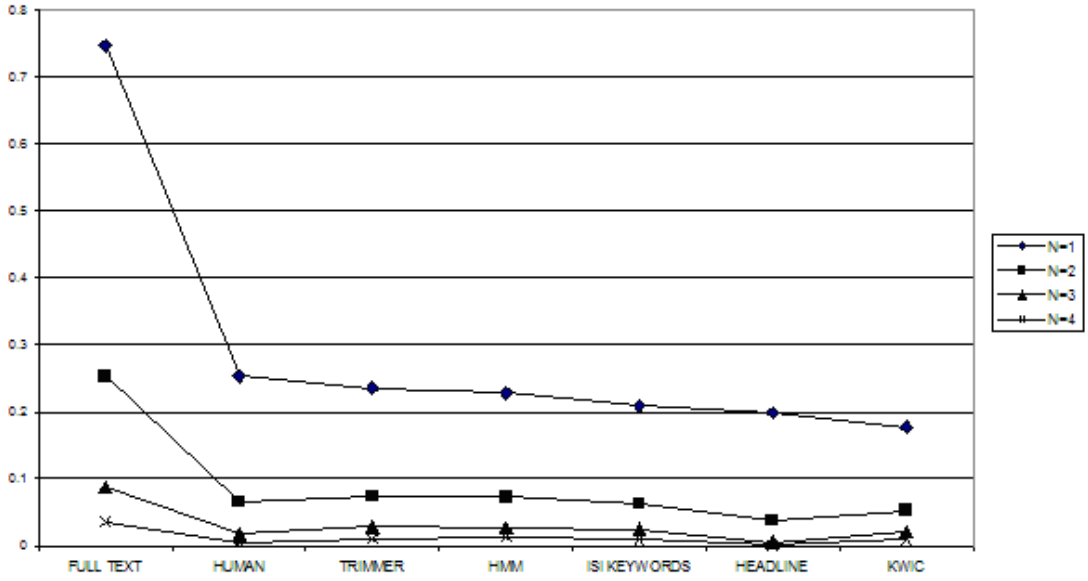

Figure 4.2: ROUGE Scores

System	B1	B2	B3	B4	R1	R2	R3	R4
Human	**0.4129**	**0.2199**	**0.1291**	**0.0795**	0.2531	0.0655	0.0178	0.00527
HMM	0.3192	0.1815	0.1090	0.0666	0.2278	0.0730	0.0270	0.01275
Trimmer	0.2830	0.1606	0.0998	0.0648	0.2354	0.0738	0.0282	0.01015
Headline	0.2536	0.1130	0.0486	0.0229	0.1985	0.0375	0.0054	0.00035
ISIKWD	0.2383	0.1292	0.0703	0.0374	0.2090	0.0624	0.0242	0.00879
KWIC	0.1485	0.0696	0.0396	0.0246	0.1766	0.0525	0.0207	0.00945
Full Text	0.0212	0.0129	0.0077	0.0048	**0.7473**	**0.2528**	**0.0876**	**0.03576**

Table 4.13: BLEU and ROUGE Scores on Seven Systems, sorted by BLEU-1

human-generated outputs (Full Text, Human and Headline).

4.1.7 Correlation of Intrinsic and Extrinsic Measures

First, the correlation is computed on the basis of the average performance of a system for all topics. Table 4.14 and Table 4.15 show the correlations—using Pearson r (Siegel and Castellan, 1988)—between the average system scores assigned by the task-based metrics from Table 4.3 and the automatic metrics from Table 4.13. Pearson's statistics are commonly used in summarization and machine translation evaluation (see e.g. (Lin, 2004; Lin and Och, 2004)). Pearson r is computed as:

$$\frac{\sum_{i=1}^{n}(r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum_{i=1}^{n}(r_i - \bar{r})^2}\sqrt{\sum_{i=1}^{n}(s_i - \bar{s})^2}}$$

where s_i is the score of system i with respect to a particular measure (e.g., precision) and $\bar{s}$ is the average score over all systems, including the full text.

The intrinsic and extrinsic scores for each summarization method are computed, averaging over the individual topics. The correlation between an intrinsic and an extrinsic evaluation method is then computed by pairwise comparing the intrinsic score and the extrinsic score of each summarization system.

Table 4.14 shows that the ROUGE 1-gram and 2-gram results have a very high, positive correlation with the recall measure—the strongest correlation being between ROUGE-1 and recall. When the full text system is excluded (as shown in Table 4.15), this correlation decreases dramatically, and the BLEU-1 measure

	Accuracy	**Precision**	**Recall**	**F-Score**
BLEU-1	**0.299**	**0.457**	-0.617	-0.488
BLEU-2	0.284	0.438	-0.592	-0.475
BLEU-3	0.222	0.364	-0.541	-0.404
BLEU-4	0.166	0.294	-0.490	-0.404
ROUGE-1	0.266	-0.028	**0.945**	**0.904**
ROUGE-2	0.197	-0.080	0.915	0.859
ROUGE-3	0.092	-0.161	0.859	0.783
ROUGE-4	0.070	-0.153	0.820	0.738

Table 4.14: Pearson r Correlation between Intrinsic and Extrinsic Scores Grouped by System (including Full Text) for Strict Relevance

	Accuracy	**Precision**	**Recall**	**F-Score**
BLEU-1	**0.731**	**0.621**	0.464	**0.646**
BLEU-2	0.636	0.551	0.341	0.502
BLEU-3	0.486	0.416	0.295	0.366
BLEU-4	0.376	0.312	0.292	0.366
ROUGE-1	0.495	0.395	0.320	0.435
ROUGE-2	-0.040	-0.018	-0.157	-0.173
ROUGE-3	-0.375	-0.300	-0.391	-0.491
ROUGE-4	-0.357	-0.229	**-0.471**	-0.551

Table 4.15: Pearson r Correlation between Intrinsic and Extrinsic Scores Grouped by System (excluding Full Text) for Strict Relevance

exhibits the highest correlation with accuracy. The results with the Full Text system included supports the fourth hypothesis, that an intrinsic measure would have a high (>0.8) correlation with human performance. It must be noted that without the inclusion of the Full Text system, this hypothesis is not supported.

4.1.8 Experimental Findings

The first hypothesis, that the Full Text would provide an upper bound for all performance measures was supported for Non-Strict relevance, but was not supported for Strict relevance.

The second hypothesis, was that out of all the systems, the human-generated systems (Headline and Human) would perform the best. Again, for Non-strict relevance this hypothesis was supported but was not supported for Strict relevance.

The third hypothesis, that the KWIC system would perform lowest of all systems was supported with the results of Strict relevance for the accuracy, precision, and signal detection d' measures. This hypothesis was not supported for Non-strict relevance for the recall, f-score, and d' measures.

The fourth hypothesis, that one or both of the intrinsic measures would generate a high (> 0.8) correlation with a measure of human performance. When Full Text was included as a system, this hypothesis was supported by the correlation results with the extrinsic measures and ROUGE. However, when Full Text was excluded as a system, this hypothesis was not supported.

The final hypothesis was that the judgment speed for summaries would be an

order of magnitude greater than that of the full text. The observed improvement speed with the summaries was a factor of three over the full text; not supporting this hypothesis.

None of the hypotheses for the experiment were fully supported by the results. A concern with this experiment was the low individual performance, low interannotator agreement, low Kappa scores and inability to show statistically significant differences for most of the measures. This was thought to be related to the type of relevance assessment task used. In the next section, an event-based task instead of a topic-based task is suggested to encourage more reliable results for the next experiment.

Also, the results with Strict and Non-Strict Relevance were not consistent. In some cases, a given system was ranked highly (the highest or second highest scoring system) with Strict Relevance but then ranked poorly (ranking as the lowest or second lowest system) with Non-Strict Relevance. For the next experiment, the users are constrained to making only a "Relevant" or "Not Relevant" judgment. The elimination of "Somewhat Relevant" is expected to help minimize the issues produced by Strict and Non-Strict Relevance.

4.2 LDC Event Tracking: Correlation with an Extrinsic Event Tracking Relevance Assessment

A second experiment, LDC Event Tracking, uses a more constrained type of document relevance assessment in an extrinsic task for evaluating human performance using automatic summaries. This task, *event tracking*, has been reported in NIST Topic Detection and Tracking (TDT) evaluations to provide the basis for more reliable results in that this task relates to the real-world activity of an analyst conducting full-text searches using an IR system to quickly determine the relevance of a retrieved document. The choice of a more constrained task for this experiment was motivated by the need to overcome the low interannotator agreement and inconsistencies of the previous experiment.

Users were asked to decide if a document contains information related to a particular event in a specific domain. The user is told about a specific event, such as the bombing of the Murrah Federal Building in Oklahoma City. A detailed description is given about what information is considered relevant to an event in the given domain. For instance, in the criminal case domain, information about the crime, the investigation, the arrest, the trial and the sentence are considered relevant.

4.2.1 Hypotheses

The initial hypothesis is that it is possible to save time using summaries for relevance assessment without adversely impacting the degree of accuracy that would be possible with full documents. This is similar to the "summarization condition test" used in SUMMAC (Mani et al., 2002), with the following differences: (1) the lower baseline is fixed to be the first 75 characters (instead of 10% of the original document size); and (2) all other summaries are also fixed-length (no more than 75 characters), following the NIST Document Understanding Conference (DUC) guidelines.

A second hypothesis is that this task supports a very high degree of interannotator agreement, i.e., consistent relevance decisions across users. This is similar to the "consistency test" applied in SUMMAC, except that it is applied not just to the full-text versions of the documents, but also to all types of summaries. In addition, to validate the hypothesis, a much higher degree of agreement was required—e.g., a 0.67 Kappa score as opposed to the .38 Kappa score achieved in the SUMMAC experiments. According to Krippendorf (1980), Kappa scores should be at least 0.67 to allow conclusions to be drawn. (The reader is also referred to (Carletta, 1996) and (Eugenio and Glass, 2004) for further details on Kappa agreement.)

A third hypothesis is that it is possible to demonstrate a correlation between automatic intrinsic measures and extrinsic task-based measures-most notably, a

correlation between ROUGE (the automatic intrinsic measure) and recall (the extrinsic measure)—in order to establish an automatic and inexpensive predictor of human performance. In the previous experiment, a high correlation was seen with ROUGE and accuracy in Table 4.14, so the aim here is to determine if this correlation is consistent.

Crucially, the validation of this third hypothesis—i.e., finding a positive correlation between the intrinsic and extrinsic measures—will result in the ability to estimate the usefulness of different summarization methods for an extrinsic task in a repeatable fashion without the need to conduct user studies. This is an important because, as pointed out by (Mani, 2002), conducting a user study is extremely labor intensive and requires a large number of human users in order to establish statistical significance.

4.2.2 Experiment Details

This experiment uses seven types of automatically generated document surrogates; two types of manually generated surrogates; and, as a control, the entire document. The automatically generated surrogates are:

- **KWIC** – Keywords in Context (Monz, 2004);
- **GOSP** – Global word selection with localized phrase clusters (Zhou and Hovy, 2003);
- **ISIKWD** – Topic independent keyword summary (Hovy and Lin, 1997);

- **UTD** – Unsupervised Topic Discovery (Schwartz et al., 2001);
- **Trimmer** – Fluent headline based on a linguistically-motivated parse-and-trim approach (Dorr et al., 2003);
- **Topiary** – Hybrid topic list and fluent headline based on integration of UTD and Trimmer (Zajic et al., 2004a);
- **First75** – the first 75 characters of the document; used as the lower baseline summary.

The manual surrogates are:

- **Human** – a human-generated 75 character summary (commissioned for this experiment);
- **Headline** – a human-generated headline associated with the original document.

Finally, as before, the "Full Text" document was included as a system and was expected to serve as an upper baseline.

This experiment includes some additional systems that were not available for the previous experiment. The First75 system was added as a lower baseline measure. It was expected that all systems would generate performance measures between that of the Full Text (upper baseline) and First75 (lower baseline).

The average lengths of the surrogates in this experiment are shown in Table 4.16. In this experiment, the outputs of each of the experimental systems

System	Avg Word Count	Avg Char Count
TEXT	594	3696
Headline	9	54
Human	11	71
First75	12	75
KWIC	11	71
GOSP	11	75
ISIKWD	11	75
UTD	9	71
TRIMMER	8	56
TOPIARY	10	73

Table 4.16: LDC Event Tracking Experiment: Average Word and Character Counts for Each Surrogate

were constrained to 75 characters, a guideline used by the current DUC evaluation (Harman and Over, 2004). This constraint was imposed to encourage consistency amongst the size of the system output and to make the evaluation process more fair. Systems with longer summary output may have an unfair scoring advantage over systems with shorter output since the longer output means that more information is retained from the original text.

In this experiment, 20 topics are selected from the Topic Detection and Tracking version 3 (TDT-3) corpus (Allan et al., 1999). For each topic, a 20-document subset has been created from the top 100 ranked documents retrieved by the FlexIR information retrieval system (Monz and de Rijke, 2001). Crucially, each subset has been constructed such that exactly 50% of the documents are relevant to the topic. The full-text documents range in length from 42 to 3083 words. The documents are long enough to be worth summarizing, but short enough to be read within a reasonably short amount of time. The documents consist of a combination of news stories

stemming from the Associated Press newswire and the New York Times. The topics include Elections, Scandals/Hearings, Legal/Criminal Cases, Natural Disasters, Accidents, Ongoing violence or war, Science and Discovery News, Finances, New Laws, Sport News, and miscellaneous news (see Appendix A for details). Each topic includes an event description and a set of 20 documents. An example of an event description is shown in Table 4.17. The *Rules of Interpretation* (Appendix A) are used as part of the instructions to users on how to determine whether or not a document should be judged relevant or not relevant.

The TDT-3 data also provides 'gold-standard' judgments—what are thought to be the correct relevance level of the documents as decided by the LDC annotators. (This is referred to as "LDC Agreement" in the later experiments.) Each document is marked *relevant* or *not relevant* with respect to the associated event. These gold-standard judgments are used in the analysis to produce accuracy, precision, recall, and f-score results.

4.2.3 Experiment Design

In this study, 14 undergraduate and 6 graduate students were recruited at the University of Maryland at College Park through posted experiment advertisements to participate in the experiment. Participants were asked to provide information about their educational background and experience (Appendix B). All participants had extensive online search experience (4+ years) and their fields of study included engineering, psychology, anthropology, biology, communication, American studies,

System	Example Output
Full Text	Ugandan President Yoweri Museveni flew to **Libya**, apparently violating U.N. sanctions, for talks with **Libyan leader Moammar Gadhafi**, the official JANA news agency said Sunday. Egypt's Middle East News Agency said the two met Sunday morning. The JANA report, monitored by the BBC, said the two leaders would discuss the peace process in the Great Lakes region of Africa. Museveni told reporters on arrival in the **Libyan** capital Tripoli on Saturday that he and **Gadhafi** also would discuss "new issues in order to contribute to the solution of the continent's problems," the BBC quoted JANA as saying. African leaders have been flying into **Libya** since the Organization of African Unity announced in June that it would no longer abide by the air embargo against **Libya** when the trips involved official business or humanitarian projects. The **U.N.** Security Council imposed an air travel ban and other sanctions in 1992 to try to force **Gadhafi** to **surrender** two Libyans wanted in the **1988 bombing** of a **Pan Am** jet over **Lockerbie**, Scotland, that **killed 270** people.
Headline	Museveni in **Libya** for talks on Africa
Human	Ugandan president flew to **Libya** to meet **Libyan leader**, violating **UN** sanctions
First75	Ugandan President Yoweri Museveni flew to **Libya**, apparently violating **U.N.**
KWIC	**Gadhafi** to surrender two Libyans wanted in the **1988** bombing of a **PanAm**
GOSP	ugandan president yoweri museveni flew **libya** apparently violating un sancti
ISIKWD	**gadhafi libya** un sanctions ugandan talks **libyan** museveni **leader** agency pres
UTD	**LIBYA** KABILA **SUSPECTS** NEWS CONGO IRAQ FRANCE NATO PARTY **BOMBING** WEAPONS
TRIMMER	Ugandan President Yoweri Museveni flew apparently violating **U.N.** sanctions
TOPIARY	NEWS **LIBYA** Ugandan President Yoweri Museveni flew violating **U.N.** sanctions

Table 4.17: Example Output From Each Experimental System

and economics. The instructions for the task (taken from the TDT-3 corpus instruction set that were given to document annotators) are shown in Appendix C.

Each of the 20 topics, T_1 through T_{20}, consisted of 20 documents corresponding to one event. The twenty human users were divided into ten user groups (A through J), each consisting of two users who saw the same two topics for each system (not necessarily in the same order). By establishing these user groups, it was possible to collect data for an analysis of within-group judgment agreement.

Each human user was asked to evaluate 22 topics (including two practice event topics not included in this analysis). Their task was to specify whether each displayed document was "relevant" or "not relevant" with respect to the associated event. Because two users saw each system/topic pair, there were a total of $20 \times 2 = 40$ judgments made for each system/topic pair, or 800 total judgments per system (across 20 topics). Thus, the total number of judgments, across 10 systems, was 8000.

A Latin square design (Table 4.18) was used to ensure that each user group viewed output from each summarization method and made judgments for all twenty event sets (two event sets per summarization system), while also ensuring that each user group saw a distinct combination of system and event. The system/event pairs were presented in a random order (both across user groups and within user groups), to reduce the impact of topic-ordering and fatigue effects.

The users performed the experiment on a Windows or Unix workstation,

System	T_1T_2	T_3T_4	T_5T_6	T_7T_8	T_9T_{10}	$T_{11}T_{12}$	$T_{13}T_{14}$	$T_{15}T_{16}$	$T_{17}T_{18}$	$T_{19}T_{20}$
Full Text	A	B	C	D	E	F	G	H	I	J
Headline	B	C	D	E	F	G	H	I	J	A
Human	C	D	E	F	G	H	I	J	A	B
First75	D	E	F	G	H	I	J	A	B	C
KWIC	E	F	G	H	I	J	A	B	C	D
GOSP	F	G	H	I	J	A	B	C	D	E
ISIKWD	G	H	I	J	A	B	C	D	E	F
UTD	H	I	J	A	B	C	D	E	F	G
Trimmer	I	J	A	B	C	D	E	F	G	H
Topiary	J	A	B	C	D	E	F	G	H	I

Table 4.18: LDC Event Tracking Latin Square Experiment Design

using a web-based interface that was developed to display the event, document descriptions and to record the judgments. The users were timed to determine how long it took him/her to make all judgments on an event. Although the judgments were timed, the users were not confined to a specific time limit for each event but were allowed unlimited time to complete each event and the experiment.

4.2.4 Results and Analysis

Two main measures of human performance were used in the extrinsic evaluation: time and accuracy. The time of each individual's decision was measured from a set of log files and is reported in minutes per document.

The LDC 'gold-standard' relevance judgments associated with each event were used to compute accuracy. Based on these judgments, accuracy was computed as the sum of the correct hits (true positives, i.e., those correctly judged relevant) and the correct misses (true negatives, i.e., those correctly judged irrelevant) over the total number of judgments. The motivation for using accuracy to assess the human's performance is that, unlike the more general task of IR, a

System	**TP**	**FP**	**FN**	**TN**	**A**	**P**	**R**	**F**	**T** (s)
Full Text	**328**	55	68	349	**0.851**	**0.856**	**0.828**	**0.842**	**23.00**
Human	302	54	94	350	0.815	0.848	0.763	0.803	7.38
Headline	278	52	118	**652**	0.787	0.842	0.702	0.766	6.34
ISIKWD	254	60	142	344	0.748	0.809	0.641	0.715	7.59
GOSP	244	57	152	347	0.739	0.811	0.616	0.700	6.77
Topiary	272	88	124	316	0.735	0.756	0.687	0.720	7.60
First75	253	59	143	345	0.748	0.811	0.639	0.715	6.58
Trimmer	235	76	**161**	328	0.704	0.756	0.593	0.665	6.67
KWIC	297	**155**	99	249	0.683	0.657	0.750	0.700	6.41
UTD	271	135	125	269	0.675	0.667	0.684	0.676	6.52
HSD, $p<0.05$					0.099	0.121	0.180	0.147	4.783

Table 4.19: Results of Extrinsic Task Measures on Ten Systems, sorted by **A**ccuracy

50% relevant/irrelevant split has been enforced across each document set. This balanced split justifies the inclusion of true negatives in the performance assessment. (This would not be true in the general case of IR, where the vast majority of documents in the full search space are cases of true negatives.)

Again using the contingency table, Table 3.1, the extrinsic measures used for this experiment are: accuracy, precision, recall/sensitivity, f-score, specificity and d'. The Tukey Honestly Significant Difference (HSD) is also computed to determine whether differences found between groups of systems are statistically significant.

Table 4.19 shows **TP**, **FP**, **FN**, **TN**, **P**recision, **R**ecall, **F**-score, and **A**ccuracy for each of the 10 systems. In addition, the table gives the average **T**(ime) it took users to make a judgment-in seconds per document-for each system. The rows are sorted by accuracy, which is the focus for the remainder of this discussion.

One-factor repeated-measures ANOVA (with 97 degrees of freedom) was com-

puted to determine if the differences among the systems were statistically significant for five measures: precision, recall, f-score, accuracy, and time. Each user saw each system twice during the experiment, so each sample consisted of a user's judgments on the 40 documents that comprised the two times the user saw the output of a particular system. Precision, recall, f-score, accuracy and time were calculated on each sample of 40 judgments.

The HSD is shown for each measure in the bottom row of Table 4.19 with $p<0.05$. If the difference in measures between two systems is greater then the HSD, then a significant difference between the systems can be claimed. Unfortunately, significant differences with $p<0.05$ cannot be claimed between any of automatic systems for precision, recall, f-score or accuracy using the Tukey Test.

Using the mean scores from Table 4.19, the results of the ANOVA were tested for significant differences among the extrinsic measures with just the seven automatic systems. In this analysis, only precision was found to have significant differences due to system. The HSD value at $p<0.05$ is 0.117 for precision, which allows the automatic systems to be grouped into two overlapping sets, A and B, the members of which are not significantly distinct according to the Tukey test. This is shown in Table 4.20.

Although the accuracy differences are insignificant across systems, the decision-making was sped up significantly—3 times as much (e.g., 7.38 seconds/summary for HUMAN compared to 23 seconds/document for the TEXT)—by using summaries instead of the full text document. In fact, it is possible that the summaries

First75	A	
GOSP	A	
ISIKWD	A	
TOPIARY	A	B
TRIMMER	A	B
UTD		B
KWIC		B

Table 4.20: Equivalence Classes of Automatic Summarization Systems with respect to Precision

provide even more of a timing benefit than is revealed by these results. Because the full texts are significantly longer than 3 times the length of the summaries, it is likely that the human users were able to use the bold-faced descriptor words to skim the texts—whereas skimming is less likely for a one-line summary. However, even with skimming, the timing differences are very clear.

Note that the human-generated systems—Text, Human and Headline—performed best with respect to accuracy, with the Text system as the upper baseline, consistent with the initial expectations. However, the tests of significance indicate the many of the differences in the values assigned by extrinsic measures are small enough to support the use of machine-generated summaries for relevance assessment. For example, four of the seven automatic summarization systems show about a 5% or less decrease in accuracy in comparison with the performance of the Headline system. This validates the first hypothesis: that reading document summaries saves time over reading the entire document text without an adverse impact on accuracy. This finding is consistent with the results obtained further in the previous SUMMAC experiments.

System	Sensitivity	Specificity	discriminability index (d')
Headline	0.702	**0.926**	**1.978**
Human	**0.763**	0.866	1.824
ISIKWD	0.641	0.851	1.405
GOSP	0.616	0.859	1.371
Trimmer	0.593	0.812	1.121
Topiary	0.687	0.718	1.064
KWIC	0.750	0.616	0.970
UTD	0.684	0.666	0.908
First75	0.639	0.707	0.900

Table 4.21: Results of the Signal Detection Measures, sorted by d'

The results for the sensitivity, specificity, and d' measures are given in Table 4.21. Here, the human-generated systems, Headline and Human, achieve the highest and second highest d' score, respectively. These results support the claim that human-generated summaries are more helpful to users than automatic summaries.

4.2.5 Discussion

Recall that the second hypothesis is that this task supports a very high degree of interannotator agreement-beyond the low rate of agreement (16-69%) achieved in the SUMMAC experiments. Table 4.22 shows "User Agreement," i.e., agreement of both relevant and irrelevant judgments of users within a group, and the kappa score based on user agreement.

Again, the Kappa score is calculated as:

$$\frac{P_A - P_E}{1 - P_E}$$

System	User Agreement	Kappa Score
Full Text	**0.840**	**0.670**
Human	0.815	0.630
Headline	0.800	0.600
ISIKWD	0.746	0.492
Topiary	0.735	0.470
GOSP	0.785	0.570
First75	0.778	0.556
Trimmer	0.805	0.610
KWIC	0.721	0.442
UTD	0.680	0.350

Table 4.22: User Agreement and Kappa Score

with P_A equaling the agreement, and P_E equaling the expected agreement by chance, which in this case is 0.5. As shown in the table, the kappa scores for all systems except UTD are well above the kappa scores computed in the SUMMAC experiment (0.38), thus supporting the hypothesis that this task that is unambiguous enough that users can perform it with a high level of agreement.

4.2.6 Automatic Intrinsic Evaluation

Three 75-character summaries were commissioned (in addition to the summaries in the HUMAN system) to use as references for BLEU and ROUGE. As before, BLEU and ROUGE were run with 1-grams through 4-grams, and two new variants of ROUGE (that were not previously available), ROUGE-L and ROUGE-W-1.2, were run. The results are shown in Table 4.23.

Analogously to the extrinsic evaluation measures discussed above, the ANOVA values were computed to see whether there are differences between the systems for

System	R1	R2	R3	R4	RL	RW	B1	B2	B3	B4
Full Text	**0.8181**	**0.3510**	**0.1678**	**0.1001**	**0.7012**	**0.3866**	0.030	0.020	0.014	0.010
First75	0.2600	0.0982	0.0513	0.0312	0.2289	0.1384	0.389	**0.256**	**0.186**	**0.142**
ISIKWD	0.2419	0.0087	0.0003	0.0000	0.1623	0.0946	0.404	0.074	0.017	0.000
Topiary	0.2248	0.0699	0.0296	0.0137	0.1931	0.1158	0.360	0.207	0.133	0.090
KWIC	0.2027	0.0609	0.0281	0.0169	0.1731	0.1048	0.331	0.191	0.129	0.095
Headline	0.2008	0.0474	0.0128	0.0030	0.1767	0.1040	0.349	0.186	0.102	0.057
GOSP	0.2004	0.0629	0.0211	0.0084	0.1810	0.1080	0.307	0.186	0.112	0.069
Trimmer	0.1890	0.0710	0.0335	0.0163	0.1745	0.1055	0.341	0.228	0.160	0.115
Human	0.1684	0.0387	0.0118	0.0046	0.1451	0.0857	**0.433**	0.254	0.154	0.096
UTD	0.1280	0.0144	0.0013	0.0000	0.1068	0.0654	0.191	0.023	0.000	0.000
HSD, $p < 0.05$	0.05	0.0289	0.02	0.013	0.0429	0.0246	0.0826	0.0659	0.0568	0.0492

Table 4.23: ROUGE and BLEU Scores on Ten Systems, sorted by ROUGE-1

each evaluation method. For each case, ANOVA showed that there are statistically significant differences with $p<0.05$ and the last row shows the honestly significant differences for each measure.

The ROUGE and BLEU results are shown graphically in Figures 4.3 and 4.4, respectively. In both graphic representations, the 95% confidence interval is shown by the error bars on each line.

In Figure 4.3, it can be seen that the full text performs much better than some of the summarization methods, e.g. ISIKWD and Topiary for ROUGE-1. This is to be expected because the full text contains almost all n-grams that appear in the reference summaries. In figure 4.4, the full document representation performs rather poorly. This is also an expected result because the full document contains a large number of n-grams, only a small fraction of which occur in the reference summarizations.

The ANOVA test was also performed on the seven automatic systems with respect to the different intrinsic measures. The ANOVA test showed that all intrinsic measures resulted in statistically significant differences between the systems,

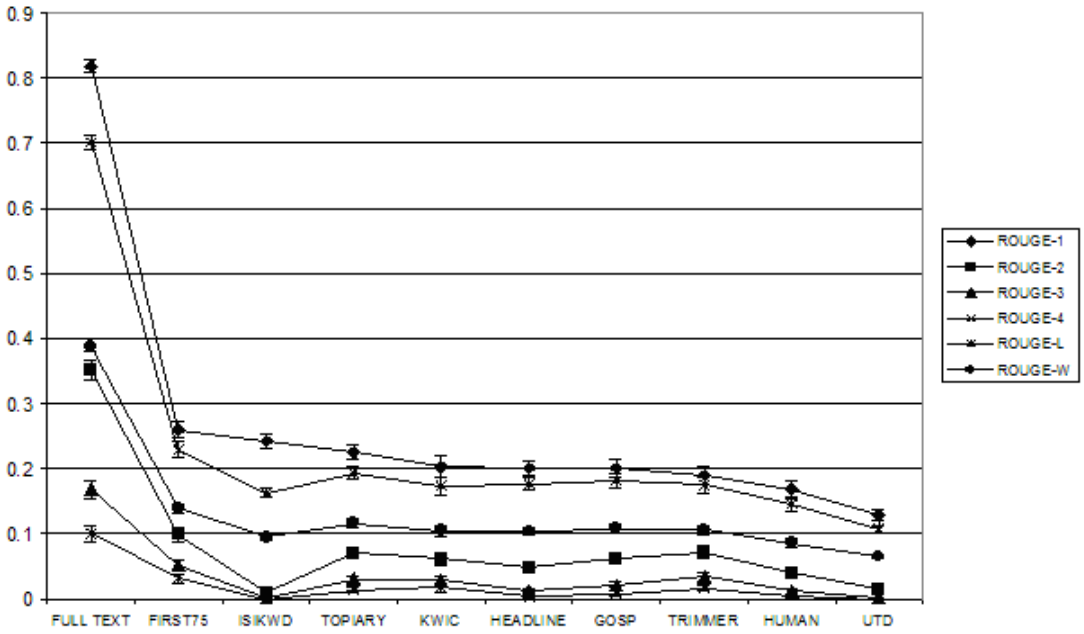

Figure 4.3: ROUGE Results for Ten Systems, (X axis ordered by ROUGE-1)

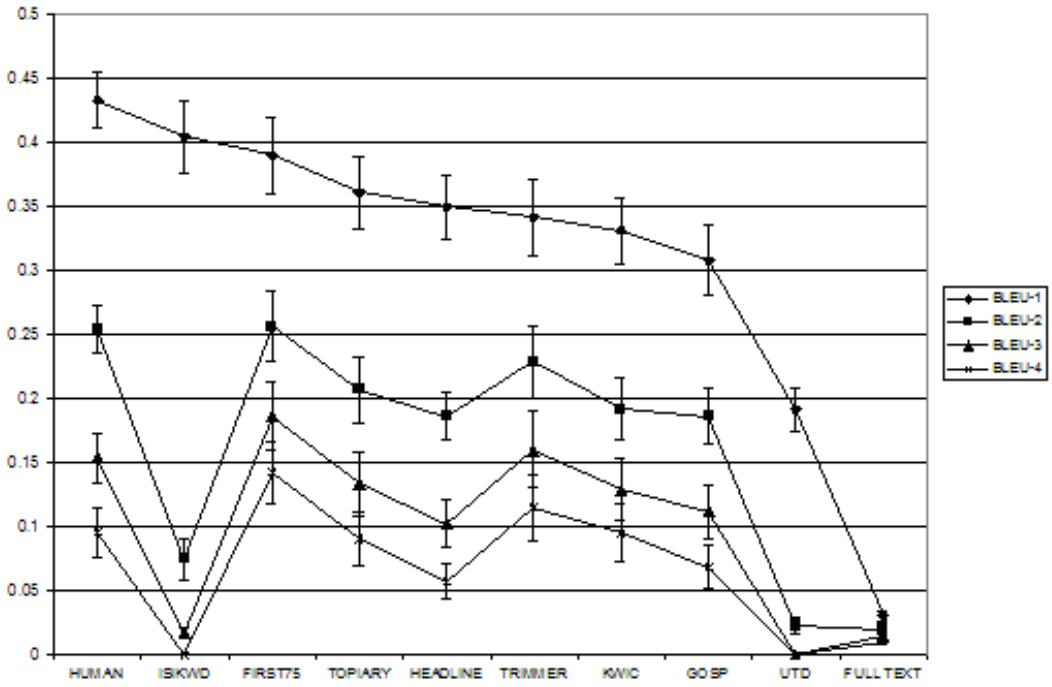

Figure 4.4: BLEU Results for Ten Systems, (X axis ordered by BLEU-1)

	R1	**R2**	**R3**	**R4**	**RL**	**RW**	**B1**	**B2**	**B3**	**B4**
HSD, $p < 0.05$	0.04	0.03	0.02	0.01	0.04	0.02	0.09	0.09	0.06	0.05

Table 4.24: Honestly Significant Differences for Automatic Summarization Methods Using ROUGE and BLEU

First75	A			
ISIKWD	A	B		
Topiary	A	B	C	
KWIC		B	C	
GOSP		B	C	
Trimmer			C	
UTD				D

Table 4.25: Equivalence Classes of Automatic Summarization Systems with respect to ROUGE-1

which allows the honestly significant differences (HSD) to be computed for each measure, which is shown in Table 4.24.

As was done for the extrinsic measures above, the different summarization systems can be grouped, based on the honestly significant difference. For illustration purposes the groupings are shown for ROUGE-1 and BLEU-1 in Tables 4.25 and 4.26.

ISIKWD	A	
First75	A	
Topiary	A	
Trimmer	A	
KWIC	A	
GOSP	A	
UTD		B

Table 4.26: Equivalence Classes of Automatic Summarization Systems with respect to BLEU-1

Evaluation with ROUGE-1 allows for a differentiated grouping with the sys-

tems being separated into four groups, A, B, C and D, while evaluation with BLEU-1 only resulted in two groups, A and B.

4.2.7 Correlation of Intrinsic and Extrinsic Measures

To test the third hypothesis, the results of the automatic metrics were compared to those of the human system performance and it was shown that there is a statistically significant correlation between different intrinsic evaluation measures and common measures used for evaluating performance in an extrinsic task, such as accuracy, precision, recall, and f-score. In particular, the automatic intrinsic measure ROUGE-1 is significantly correlated with accuracy and precision. However, as will be seen shortly, this correlation is low when the summaries are considered alone (i.e., if the full text is excluded).

First, the correlation is computed on the basis of the average performance of a system for all topics. As was seen above, there are significant differences between human performance measures and the scoring by the automatic evaluation systems. Table 4.27 through Table 4.29 below show the rank correlations between the average system scores assigned by the task-based metrics from Table 4.19 and the automatic metrics from Table 4.23. Two methods were used for computing this correlation: Pearson r as used for comparison with the previous experiment, and also Spearman ρ (Siegel and Castellan, 1988) is introduced in this experiment to produce correlation results more suitable for this task.

The intrinsic and extrinsic scores for each summarization method are com-

	Accuracy	**Precision**	**Recall**	**F-Score**
ROUGE-1	**0.647***	**0.441**	0.619	**0.717***
ROUGE-2	0.603	0.382	0.602	0.673*
ROUGE-3	0.571	0.362	0.585	0.649*
ROUGE-4	0.552	0.342	0.590	0.639*
ROUGE-L	0.643*	0.429	0.619	0.710*
ROUGE-W	0.636*	0.424	0.613	0.703*
BLEU-1	-0.404	-0.082	**-0.683***	-0.517
BLEU-2	-0.211	-0.017	-0.475	-0.305
BLEU-3	-0.231	-0.064	-0.418	-0.297
BLEU-4	-0.302	-0.137	-0.417	-0.339

Table 4.27: Pearson r Correlation between Intrinsic and Extrinsic Scores Grouped by System (including Full Text)

puted, averaging over the individual topics. Then, the correlation between an intrinsic and an extrinsic evaluation method is computed by pairwise comparing the intrinsic score and the extrinsic score of each summarization system.

Table 4.27 shows the results for Pearson r correlation. Correlations that are statistically significant at the level of $p<0.05$ with respect to one-tailed testing are marked with a single asterisk (*).

Looking back at Figures 4.3 and 4.4, the full text has much higher ROUGE scores than any of the other systems, and also the full text has much lower BLEU scores than any of the other systems. These extremes result in correlation results that are highly distorted. Thus, it is questionable whether the inclusion of full text allows valid statistical inferences to be drawn. If the full text is treated as an outlier, removing it from the set of systems, the correlations are significantly weaker—this point will be described in more depth later. Table 4.28 shows the results for Pearson r over all systems, excluding full text.

	Accuracy	**Precision**	**Recall**	**F-Score**
ROUGE-1	0.229	**0.389**	-0.271	0.171
ROUGE-2	0.000	0.055	-0.222	-0.051
ROUGE-3	-0.111	-0.013	-0.241	-0.128
ROUGE-4	-0.190	-0.083	-0.213	-0.168
ROUGE-L	0.205	0.329	**-0.293**	0.115
ROUGE-W	0.152	0.275	-0.297	0.071
BLEU-1	**0.281**	0.474	-0.305	**0.197**
BLEU-2	0.159	0.224	-0.209	0.089
BLEU-3	0.026	0.104	-0.222	-0.022
BLEU-4	-0.129	-0.012	-0.280	-0.159

Table 4.28: Pearson r Correlation between Intrinsic and Extrinsic Scores Grouped by System (excluding Full Text)

	Accuracy	**Precision**	**Recall**	**F-Score**
ROUGE-1	**0.233**	0.083	-0.116	0.300
ROUGE-2	-0.100	-0.150	-0.350	-0.150
ROUGE-3	-0.133	-0.183	-0.316	-0.200
ROUGE-4	-0.133	**-0.216**	-0.166	-0.066
ROUGE-L	0.100	-0.050	-0.233	0.100
ROUGE-W	0.100	-0.050	-0.233	0.100
BLEU-1	0.300	**0.216**	-0.250	**0.333**
BLEU-2	-0.016	-0.083	**-0.366**	-0.066
BLEU-3	-0.016	-0.083	**-0.366**	-0.066
BLEU-4	-0.133	-0.183	-0.316	-0.200

Table 4.29: Spearman ρ Correlation between Intrinsic and Extrinsic Scores Grouped by System (excluding Full Text)

Spearman ρ is computed exactly like the Pearson r correlation, but instead of comparing actual scores, one compares the system ranking based on an intrinsic measure with the system ranking based on an extrinsic measure. The Spearman ρ correlation between intrinsic and extrinsic scores is shown-excluding the full text-in Table 4.29 below.

Tables 4.28 and 4.29 show that there is a positive correlation in some cases,

	Accuracy	**Precision**	**Recall**	**F-Score**
ROUGE-1	**0.306***	**0.208***	**0.246***	**0.283***
ROUGE-2	0.279*	0.169*	0.227*	0.250*
ROUGE-3	0.245*	0.134	0.207*	0.217*
ROUGE-4	0.212*	0.106	0.188*	0.189*
ROUGE-L	0.303*	0.199*	0.244*	0.278*
ROUGE-W	0.299*	0.197*	0.243*	0.274*
BLEU-1	-0.080	0.016	-0.152	-0.106
BLEU-2	-0.048	0.012	-0.133	-0.088
BLEU-3	-0.063	-0.032	-0.116	-0.096
BLEU-4	-0.082	-0.076	-0.104	-0.095

Table 4.30: Pearson r Correlation between Intrinsic and Extrinsic Scores Grouped by System-Topic Pair-200 Data Points (including Full Text)

but it also shows that all positive correlations are rather low. Tests of statistical significance indicate that none of the Pearson r and Spearman ρ correlations is statistically significant at level $p<0.05$.

Computing correlation on the basis of the average performance of a system for all topics has the disadvantage that there are only 10 data points which leads to rather unstable statistical conclusions. In order to increase the number of data points a data point is redefined here as a system-topic pair, e.g., First75/topic3001 and Topiary/topic3004 are two different data points. In general a data point is defined as system-i/topic-n, where $i = 1...10$ (ten summarization systems are compared) and $n = 1...20$ (20 topics are being used). This new definition of a data point will result in 200 data points for the current experiment.

The Pearson r correlation between extrinsic and intrinsic evaluation measures using all 200 data points—including the full text—is shown in Table 4.30.

Because the primary interest is in the performance with respect to summaries

	Accuracy	**Precision**	**Recall**	**F-Score**
ROUGE-1	**0.181***	**0.178***	**0.108**	**0.170***
ROUGE-2	0.078	0.057	0.034	0.058
ROUGE-3	0.005	-0.007	-0.120	-0.010
ROUGE-4	-0.063	-0.062	-0.051	-0.069
ROUGE-L	0.167*	0.150	0.098	0.151
ROUGE-W	0.149	0.137	0.092	0.135
BLEU-1	0.137	0.171*	-0.005	0.078
BLEU-2	0.065	0.088	-0.051	0.009
BLEU-3	0.014	0.016	-0.057	-0.028
BLEU-4	-0.027	-0.042	-0.057	-0.045

Table 4.31: Pearson r Correlation between Intrinsic and Extrinsic Scores Grouped by System-Topic Pair (excluding Full Text)

only, the 20 data points that use full text will be removed from the data set and the following discussion is based on the remaining 180 data points only. The Pearson r correlation for all pairs of intrinsic and extrinsic measures on all systems, excluding the full text, is shown in Table 4.31.

Overall, the correlation is not very strong, but in some cases, a statistically significant positive correlation can be detected between intrinsic and extrinsic evaluation measures—again, those marked with a single asterisk (*).

Although grouping the individual scores in the form of system-topic pairs results in more data points than using only the systems as data points it introduces another source of noise. In particular, given two data points system-i/topic-n and system-j/topic-m, where the former has a higher ROUGE-1 score than the latter but a lower accuracy score, the two data points are inversely correlated. The problem is that the reordering of this pair with respect to the two evaluation measures may not only be caused by the quality of the summarization method,

but also by the difficulty of the topic. For some topics it is easier to distinguish between relevant and non-relevant documents than for others. Since the main interest here lies in the effect of system performance, the effect of topic difficulty is eliminated while maintaining a reasonable sample size of data points.

In order to eliminate the effect of topic difficulty, each of the original data points are normalized in the following way: For each data point compute the score of the intrinsic measure m_i and the score of the extrinsic measure m_e. Then, for a given data point d, compute the average score of the intrinsic measure m_i for all data points that use the same topic as d and subtract the average score from each original data point on the same topic. The same procedure is applied to the extrinsic measure m_e. This will result in a distribution where the data points belonging to the same topic are normalized with respect to their difference to the average score for that topic. Since absolute values are not being used anymore, the distinction between hard and easy topics disappears.

Table 4.32 shows the adjusted correlation—using Pearson r—for all pairs of intrinsic and extrinsic measures on all systems (excluding the full text).

For completeness, as above, the Spearman ρ rank correlation between intrinsic and extrinsic evaluation measures is computed, for both the non-adjusted and adjusted cases (see Tables 4.33 and 4.34). Unlike Pearson r, the Spearman ρ rank correlation indicates that only one of the pairs shows a statistically significant correlation, viz. ROUGE-1 and precision at a level of $p<0.05$. The fact that Spearman ρ indicates significant differences in fewer cases than Pearson r might

	Accuracy	**Precision**	**Recall**	**F-Score**
ROUGE-1	0.114	**0.195***	-0.038	0.082
ROUGE-2	-0.034	0.015	-0.097	-0.050
ROUGE-3	-0.120	-0.057	**-0.140**	-0.117
ROUGE-4	**-0.195**	-0.126	-0.159	**-0.172**
ROUGE-L	0.092	0.156	-0.046	0.060
ROUGE-W	0.071	0.137	-0.054	0.045
BLEU-1	0.119	0.194*	-0.053	0.074
BLEU-2	0.039	0.093	-0.100	-0.008
BLEU-3	-0.038	0.005	-0.111	-0.063
BLEU-4	-0.107	-0.063	-0.132	-0.108

Table 4.32: Adjusted Pearson r Correlation between Intrinsic and Extrinsic Scores Grouped by System-Topic Pair (excluding Full Text)

	Accuracy	**Precision**	**Recall**	**F-Score**
ROUGE-1	**0.176**	**0.214**	**0.095**	**0.172**
ROUGE-2	0.104	0.093	0.055	0.097
ROUGE-3	0.070	0.064	0.013	0.060
ROUGE-4	0.037	-0.030	0.004	-0.012
ROUGE-L	0.160	0.170	0.089	0.160
ROUGE-W	0.137	0.172	0.083	0.140
BLEU-1	0.119	0.177	-0.006	0.077
BLEU-2	0.080	0.109	-0.019	0.041
BLEU-3	0.052	0.042	0.010	0.026
BLEU-4	-0.003	-0.037	-0.003	-0.021

Table 4.33: Spearman ρ Correlation between Intrinsic and Extrinsic Scores Grouped by System-Topic Pair (excluding Full Text)

be because Spearman ρ is a stricter test that is less likely cause a Type-I error, i.e., to incorrectly reject the null hypothesis.

4.2.8 Experimental Findings

The first hypothesis was that summaries reduce judgment time without greatly impacting the degree of accuracy as seen with the full text. The results

	Accuracy	**Precision**	**Recall**	**F-Score**
ROUGE-1	**0.123**	**0.248***	-0.070	0.064
ROUGE-2	0.022	0.072	-0.073	-0.011
ROUGE-3	-0.010	0.046	**-0.088**	-0.027
ROUGE-4	-0.066	-0.063	-0.084	-0.085
ROUGE-L	0.109	0.203	-0.066	**0.160**
ROUGE-W	0.084	0.201	-0.079	0.035
BLEU-1	0.115	0.229	-0.083	0.050
BLEU-2	0.065	0.135	-0.086	0.007
BLEU-3	0.027	0.057	-0.050	-0.009
BLEU-4	-0.034	-0.008	-0.073	-0.065

Table 4.34: Adjusted Spearman ρ Correlation between Intrinsic and Extrinsic Scores Grouped by System-Topic Pair (excluding Full Text)

supported this hypothesis in that most of the summary systems had less than a 5% reduction in accuracy with a three-fold decrease in time.

The second hypothesis wass that a higher level of interannotator agreement than the score reported in the SUMMAC study (0.38) would be seen and that these scores would also be at least 0.67, the minimum threshold proposed by Krippendorf (1980). The resulting kappa scores were higher than that of the SUMMAC study, but only the Full Text system produced of at least 0.67, thus, only partially supporting this hypothesis.

The third hypothesis was that a correlation between the results intrinsic and extrinsic measures similar to that of the previous experiment (a high correlation with ROUGE and accuracy) would be seen. Although a moderate (0.647) correlation was seen with the Pearson correlation results with the Full Text system was included, the results decreased dramatically when Full Text was excluded—both results not supporting this hypothesis.

The final hypothesis was that the Full Text would serve as an upper baseline for all the systems. Full Text achieved the highest scores for all of the extrinsic measures, supporting this hypothesis.

Although the third hypothesis wasn't supported, these experiments do show that there is a small yet statistically significant correlation between some of the intrinsic measures and a user's performance in an extrinsic task. Unfortunately, the strength of correlation depends heavily on the correlation measure: Although Pearson r shows statistically significant differences in a number of cases, a stricter non-parametric correlation measure such as Spearman ρ only showed a significant correlation in one case.

The overall conclusion that can be drawn at this point is that ROUGE-1 does correlate with precision and to a somewhat lesser degree with accuracy, but that it remains to be investigated how stable these correlations are and how differences in ROUGE-1 translate into significant differences in human performance in an extrinsic task.

4.3 Memory and Priming Study

One concern with the previous evaluation methodology was the issue of possible memory effects or priming: if the same users saw a summary and a full document about the same event, their judgments for the second system may be biased by the information provided by the first system. Thus, the goal of this

study is to determine whether the order in which summaries and corresponding full text documents are displayed can affect user's judgments.

4.3.1 Experiment Details

A small two-part experiment was conducted which explored ten summary and document orderings, further referred to as *document presentation methods*. These presentation methods range from including an extreme form of influence, with the summary and full text being presented in immediate succession, to an information source (e.g. summary) being presented on one week and the alternative source (e.g. full text) presented a week later. 8 topics including news story documents and associated headlines from the TDT-3 corpus (Allan et al., 1999) were used. The topics (termed K, M, N, P, Q, R, S and T below; the lowercase letters denote an individual document within that lettered topic, the uppercase letters denote the entire topic document set) were displayed with 10 documents each.

4.3.2 Experiment Design

Two study participants were recruited through emailed experiment advertisements. The users were given instructions on how to make relevance judgments (Appendix C) and completed a practice set in which they were shown practice summaries and documents to understand the task (the practice judgments were not included in the analysis).

The following methods were tested, ordered as shown:

- **SD1**: ($\text{Summary}_k \rightarrow \text{Document}_k$, $\text{Summary}_{k+1} \rightarrow \text{Document}_{k+1}$ on week 1) - A user is shown and makes a judgment on a document summary and then immediately makes a judgment on the corresponding full text document, for 10 summary-document pairs.

- **SD2**: ($\text{Summary}_m \rightarrow \text{Document}_m$, $\text{Summary}_{m+1} \rightarrow \text{Document}_{m+1}$ on week 2) - A user is shown and makes a judgment on a document summary and then immediately makes a judgment on the corresponding full text document, for 10 summary-document pairs.

- **S1D1**: ($\text{Summary}_n \rightarrow \text{Summary}_{n+1}$...; $\text{Document}_n \rightarrow \text{Document}_{n+1}$... on week 1) - A user is shown and makes a judgment on 10 summaries. The user then is shown and makes a judgment on the corresponding 10 full text documents.

- **S2D2**: ($\text{Summary}_p \rightarrow \text{Summary}_{p+1}$...; $\text{Document}_p \rightarrow \text{Document}_{p+1}$ on week 2) - A user is shown and makes a judgment on 10 summaries. The user then is shown and makes a judgment on the corresponding 10 full text documents.

- **S1S2**: (Summary set Q on week 1, then Summary set Q again on week 2) - A user is shown and makes a judgment on 10 consecutive summaries within a specific topic. On week 2, the user is shown and makes judgments on the same summaries from week 1.

- **D1D2**: (Document set R on week 1, then Document set R again on week 2)

- A user is shown and makes a judgment on 10 consecutive documents within a specific topic. On week 2, the user is shown and makes judgments on the same documents from week 1.

- **S1D2**: (Summary set S on week 1, then Document set S on week 2) - A user is shown and makes a judgment on 10 summaries within a specific topic on week 1. On week 2, the user is shown and makes a judgment on all the 10 corresponding full text documents.

- **D1S2**: (Document set T on week 1 then Summary set T on week 2) - A user is shown and makes a judgment on 10 full text documents within a specific topic on week 1. On week 2, the user is shown and makes a judgment on the 10 corresponding summaries.

- **SD1D2**: ($\text{Summary}_k \rightarrow \text{Document}_k$, $\text{Summary}_{k+1} \rightarrow \text{Document}_{k+1}$ on week 1 AND Document set K on week 2) - A user is shown and makes a judgment on a document summary and then immediately makes a judgment on the corresponding full text document, for 10 summary-document pairs on week 1 (which corresponds to the summary and full text document set used in Method SD1). On week 2, the user is shown and makes a judgment on the 10 corresponding documents from week 1.

- **D1SD2**: (Document set M on week 1 AND $\text{Summary}_m \rightarrow \text{Document}_m$, $\text{Summary}_{m+1} \rightarrow \text{Document}_{m+1}$ on week 2) - A user is shown and makes a judgment on Document set M on week 1. On week 2, a user is shown and

makes a judgment on the corresponding document summary and then immediately makes a judgment on the corresponding full text document (again), for 10 summary-document pairs (which corresponds to the summary and full text document set used in Method SD2).

Multiple methods were tested to determine what differences, if any, existed between the methods that could potentially influence the judgments of a user. Part 2 of the experiment was completed exactly a week after part 1 of the experiment. This was designed to decrease or factor out possible memory effects on making a summary judgment then its full text judgment or vice versa. In Methods SD1 and SD2, memory effects become a concern in that the judgments for the full text are made immediately after the user has judged the summary so the summary judgment could bias the full text judgment (the user could be encouraged to make the same judgment on the document as they did on the summary). Memory effects also become an issue in Methods S1D1 and S2D2. If memory effects are shown to exist, this method should have a lesser memory effect than that of SD1 and SD2, but a greater memory effect than if a user makes summary judgments on one week and the corresponding full text judgments a week later (Method S1D2).

As described in the SUMMAC papers (Mani, 2001; Mani et al., 2002) there were concerns with users changing relevance judgments when being presented the same full text document or summary at a different time. This is investigated with methods S1S2 and D1D2, which are used to determine if there is consistency in the user's judgments from one week to another.

	SD1	**SD2**	**S1D1**	**S2D2**	**S1S2**	**D1D2**	**S1D2**	**D1S2**
User 1	70	70	90	70	80	100	80	80
User 2	60	60	100	80	100	100	60	100

Table 4.35: Comparison of Summary/Document Judgments

	D1**SD2**	**D1**S**D2**	**SD1**D2	S**D1D2**
User 1	70	100	70	100
User 2	60	100	50	90

Table 4.36: Additional Comparison of Summary/Document Judgments

4.3.3 Results and Analysis

Tables 4.35 and 4.36 show the results of this experiment. The percentages are whether judgments remained same either from:

- Summary to corresponding Document,
- Summary week 1 → Summary week 2, or
- Document week 1 → Document week 2

Table 4.36 shows that two comparisons were made for sets D1SD2 and SD1D2. In D1SD2, the judgment made on the summary and corresponding document on week 2 were compared (shown in column one) and the judgment made on the full text documents on week one and the full text documents on week 2 were compared (shown in column two). Similarly, in SD1D2, the judgment made on the summary and corresponding document on week 1 were compared (shown in column three) and the judgment made on the full text documents on week one and the full text documents on week 2 were compared (shown in column four).

4.3.4 Discussion

The main findings of the experiment are as follows:

1. Memory effects were not an issue. This can be seen with the results of Method D1SD2 and SD1D2. The judgments users made on a document after seeing its corresponding summary were the same when they were presented with the document only. If a memory effect existed, the judgments made on full text documents that were seen immediately after a summary would differ from the judgments made when they saw the document only a week later (Method SD1D2). The judgments would also differ when they saw the full documents on week 1 and then saw the documents immediately after a summary on week two (Method D1SD2).

   (a) For example, with method SD1D2 users saw and made judgments on Document set M on week 1 without previously seeing Summary set M. On week 2, the users saw and made judgments on summary_m then the corresponding document_m, and on to $\text{summary}_{m+10} \rightarrow \text{document}_{m+10}$. The judgments made on the document set without having seen the summary and then the document set after seeing the summary were equal for user 1, and differed only by one for user 2.

   (b) Also, on week 1, with D1SD2 users saw and made judgments on summary_k then the corresponding document_k, and on to $\text{summary}_{k+10} \rightarrow \text{document}_{k+10}$. On week 2, users saw and made judgments on Document set K. The

	Summary	Document
User 1	9.2	47.9
User 2	9.6	27.5
Average	9.4	37.7

Table 4.37: Average Timing for Judgments on Summaries and Full Text Documents (in seconds)

judgments made on the document set without having seen the summary in a week and then the document set after seeing the summary were equal for both users.

2. Since memory effects were not seen, the low scoring on Methods SD1 and SD2 can be attributed to a topical effect. The topics were randomly assigned, and it is known that users may find some events more difficult to judge.

3. It is not necessary to have a two part experiment since the memory effects were not seen. Therefore, for further experimentation, any of the presentation ordering methods can be used.

4. It took users 4 times as long to make a judgment on a full document as it took to make a judgment on a summary as can be seen in Table 4.37.

4.3.5 Experimental Findings

This experiment has shown that the order in which the summaries and full text are shown do not bias the user's selections for subsequent judgments. Therefore, any of the types of presentation ordering methods can be used without fear

of a memory effect. For future experiments, method S1D1 is used, where users will make a judgment on a subset of the summaries for a given event (approximately 10 summaries), then will make judgments on the corresponding subset of the full text documents (approximately 10 full text documents). In cases where more than one summary type is used, the user will make judgments on subsets of the summaries for each of the systems, then will make judgments on the corresponding subset of the full text documents.

The concern with the previous experiments was the low agreement results of Tables 4.3, and 4.4, and low Kappa scores shown in Tables 4.11, 4.12, and 4.22. It was hypothesized, on the basis of these earlier experiments, that the order in which the summaries and documents were shown may have biased the users' judgments, but the Memory and Priming study has showed that the ordering did not have an adverse impact on the judgments. It can be concluded that additional research is necessary to determine why the agreement rates are so low and to further investigate the correlations of the human extrinsic and automatic intrinsic measures. Chapters 5 through 7 detail three additional experiments that focus on agreement measurements using the Relevance-Prediction method measure agreement rather than the gold-standard based LDC-Agreement method.

Chapter 5

A New Evaluation Method: Relevance Prediction

One of the primary goals of this research is to determine the level of correlation between current automatic intrinsic measures and human performance on an extrinsic task. At the core of this is the measurement of human performance. It has been shown (in Chapter 4) that a measure that uses low-agreement human-produced annotations does not yield stable results. It has also been argued (in Chapter 3) that this is a significant hurdle in determining the effectiveness of a summarizer for an extrinsic task such as relevance assessment. The key innovation of this thesis is the introduction and use of a new measurement technique, the Relevance-Prediction method, that yields more stable results.

This chapter reports initial findings and lays the groundwork for further experiments using this new measurement technique. The experiment presented here aims to overcome the problem of interannotator inconsistency by measuring

summary effectiveness in an extrinsic task using a much more consistent form of user judgment instead of a gold standard. The user judgments are scored with both the Relevance-Prediction and the LDC-Agreement methods.

For this experiment, only the human-generated summaries are used—the original news story Headline (*Headline*), and human summaries that were commissioned for this experiment[1] (*Human*). Although neither summary is produced automatically, this experiment focuses on the question of summary usefulness and to learn about the differences in presentation style, as a first step toward experimentation with the output of automatic summarization systems.

5.1 Hypotheses

The first hypothesis is that the summaries will allow users to achieve a Relevance-Prediction rate of 70–90%. Since these summaries are significantly shorter than the original document text, it is expected that the rate would not be 100% compared to the judgments made on the full text document. However, a ratio higher than 50% is expected, i.e., higher than that of random judgments on all of the surrogates. High performance is also expected because the meaning of the original document text is best preserved when written by a human (Mani, 2001).

[1]The human summarizers were instructed to create a summary no greater than 75 characters for each specified full text document. The summaries were not compared for writing style or quality.

A second hypothesis is that the Headline surrogates will yield a significantly lower agreement rate than that of the Human surrogates. The commissioned Human surrogates were written to stand in place of the full document, whereas the Headline surrogates were written to catch a reader's interest. This suggests that the Headline surrogates might not provide as informative a description of the original documents as the Human surrogates.

A third hypothesis is also tested: that the Relevance-Prediction measure will be more reliable than that of the *LDC-Agreement* method used for SUMMAC-style evaluations (thus providing a more stable framework for evaluating summarization techniques). LDC Agreement, as described in Section 3.1, compares a user's judgment on a surrogate or full text against the "correct" judgments as assigned by the TDT corpus annotators (Linguistic Data Consortium 2001).

Finally, the hypothesis that using a text summary for judging relevance would take considerably less time than using the corresponding full text document is also tested.

5.2 Experiment Details

Ten human participants were recruited to evaluate full text documents and two summary types.[2] The original text documents were taken from the Topic Detection and Tracking 3 (TDT-3) corpus (Allan et al., 1999) which contains news

[2]All human participants were required to be native-English speakers to ensure that the accuracy of judgments was not degraded by language barriers.

stories and Headlines, topic and event descriptions, and a mapping between news stories and their related topic and/or events. Although the TDT-3 collection contains transcribed speech documents, the investigation was restricted to documents that were originally text, i.e., newspaper or newswire, not broadcast news.

For this experiment, three distinct events were selected and related document sets[3] from TDT-3. For each event, the participants were given a description of the event (pre-written by LDC) and then asked to judge relevance of a set of 20 documents associated with that event (using three different presentation types to be discussed below).

The events used from the TDT data set were worldwide events occurring in 1998. It is possible that the participants had some prior knowledge about the events, yet it is believed that this would not affect their ability to complete the task. Participants' background knowledge of an event can also make this task more similar to real-world browsing tasks, in which participants are often familiar with the event or topic they are searching for.

The 20 documents were taken from a larger set of documents that were automatically retrieved by a search engine. A constrained subset was used where exactly half (10) were judged relevant by the LDC annotators. Because all 20 documents were somewhat similar to the event, this approach ensured that this task would be more difficult than it would be if documents were chosen from

[3]The three event and related document sets contained enough data points to achieve statistically significant results.

completely unrelated events (where the choice of relevance would be obvious even from a poorly written summary). Each document was pre-annotated with the Headline associated with the original newswire source. These Headline surrogates were used as the first summary type and had an average length of 53 characters. In addition, human-generated summaries were commissioned for each document as the second summary type. The average length of these Human surrogates was 75 characters.

Two main factors were measured: (1) differences in judgments for the three presentation types (Headline, Human, and the Full Text document) and (2) judgment time. Each participant made a total of 60 judgments for each presentation type since there were 3 distinct events and 20 documents per event. To facilitate the analysis of the data, the participant's judgments were constrained to two possibilities, *relevant* or *not relevant*.[4]

Although the Headline and Human surrogates were both produced by humans, they differed in style. The Headline surrogates were shorter than the Human surrogates by 26%. Many of these were "eye catchers" designed to compel the reader to examine the entire document (i.e., purchase the newspaper); that is, the Headline surrogates were not intended to stand in the place of the full document.

[4]If participants were allowed to make additional judgments such as *somewhat relevant*, this could possibly encourage participants to always choose this when they were the least bit unsure. Previous experiments indicate that this additional selection method may increase the level of variability in judgments (Zajic et al., 2004b).

By contrast, the writers of the Human surrogates were instructed to write text that conveyed what happened in the full document. It was observed that the Human surrogates used more words and phrases extracted from the full documents than the Headline surrogates.

5.3 Experimental Design

Experiments were conducted using a web browser (Internet Explorer) on a PC in the presence of the experimenter. Participants were given written and verbal instructions for completing their task and were asked to make relevance judgments on a practice event set. The judgments from the practice event set were not included in the experimental results or used in the analyses. The written instructions (see Appendices A and C) were given to aid participants in determining requirements for relevance. For example, in an Election event, documents describing new people in office, new public officials, change in governments or parliaments were suggested as evidence for relevance.

Each of ten participants made judgments on 20 documents for each of three different events. After reading each document or summary, the participants clicked on a radio button corresponding to their judgment and clicked a *submit* button to move to the next document description. Participants were not allowed to move to the next summary/document until a valid selection was made. No backing up was allowed. Judgment time was computed as the number of seconds it took the

System	TP	FP	FN	TN	A	P	R	F	T (s)
Full Text	**226**	**102**	74	198	**0.707**	0.689	**0.753**	**0.720**	**13.38**
Human	196	90	104	210	0.677	0.685	0.653	0.669	4.57
Headline	171	67	**129**	**233**	0.673	**0.718**	0.570	0.636	4.60
HSD, p<0.05					0.037	0.037	0.057	0.045	7.23

Table 5.1: Results of Extrinsic Task Measures on Three Presentation Types, sorted by **A**ccuracy (using LDC Agreement)

System	TP	FP	FN	TN	A	P	R	F	T (s)
Human	**251**	**35**	77	237	**0.813**	0.878	**0.765**	**0.818**	4.57
Headline	211	27	**117**	**245**	0.760	**0.887**	0.643	0.746	**4.60**
HSD, p<0.05					0.038	0.053	0.031	0.037	0.83

Table 5.2: Results of Extrinsic Task Measures on Three Presentation Types, sorted by **A**ccuracy (using Relevance Prediction)

participant to read the full text document or surrogate, comprehend it, compare it to the event description, and make a judgment (timed up until the participant clicked the *submit* button).

5.4 Results and Analysis

Tables 5.1 and 5.2 show the humans' judgments using both Relevance Prediction and LDC Agreement. Using the Relevance-Prediction measure, the Human surrogates yielded an average of 0.813 for accuracy, significantly higher than the rate of 0.707 for LDC Agreement with $p < 0.01$ (using a one-factor repeated-measures ANOVA with 29 degrees of freedom), thus confirming the first hypothesis. The Relevance-Prediction precision and f-score results were also significantly higher than the LDC-Agreement results with $p < 0.01$.

Surrogate	Sensitivity	Specificity	d'
Human (RP)	**0.765**	0.871	**1.856**
Headline (RP)	0.643	**0.901**	1.653
Headline (LDC)	0.570	0.777	0.937
Human (LDC)	0.653	0.700	0.919

Table 5.3: Results with the Signal Detection Measures

However, the second hypothesis was not confirmed. For Relevance Prediction, the Headline system yielded a rate of 0.760, which was lower than the rate for Human (0.813), but the difference was not statistically significant. It appeared that humans were able to make consistent relevance decisions from the non-extractive Headline surrogates, even though these were shorter and less informative than the Human surrogates.

This finding can be further explored using the signal detection measures of sensitivity, specificity, and d', shown in Table 5.3[5]. The LDC-Agreement and Relevance-Prediction results both show that the Human system is more useful in correctly identifying relevant documents (given by a higher sensitivity score), while the Headline system is more useful in correctly identifying not-relevant documents (given by a higher specificity score).

As for the third hypothesis, that the Relevance-Prediction measure would be more reliable than that of LDC Agreement, Tables 5.1 and 5.2 illustrate a substantial difference between the two agreement measures.

[5]The distributions for the relevant and not-relevant cases (based on the gold standards for LDC Agreement and Relevance Prediction) are overlapping equal-variance normal distributions, which allows for the calculation of the d' measure.

System	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	Avg
Headline	.80	.80	.85	.70	.73	.60	.80	.75	.60	.75	.88	.68	.80	.93	.83	**.77**
Human	.83	.88	.85	.68	.75	.75	.93	.75	.98	.90	.75	.70	.80	.90	.78	**.82**

Table 5.4: Relevance-Prediction Rates for Headline and Human Surrogates (Representative Partition of Size 4)

System	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	Avg
Headline	.70	.73	.85	.70	.63	.60	.60	.85	.50	.73	.70	.78	.65	.63	.73	**.69**
Human	.68	.75	.58	.68	.75	.70	.68	.80	.88	.58	.63	.55	.55	.60	.78	**.68**

Table 5.5: LDC-Agreement Rates for Headline and Human Surrogates (Representative Partition of Size 4)

The Relevance-Prediction rate (Accuracy) is 20% higher for the Human summaries and 13% higher for the Headline summaries. These differences are statistically significant for Human summaries (with $p<0.01$) and Headline summaries (with $p<0.05$) using a single-factor ANOVA (with 29 degrees of freedom). The higher Relevance-Prediction rate supports this hypothesis and confirms this approach provides a more stable framework for evaluating different summarization techniques.

Finally, the average timing results confirm the fourth hypothesis. The users took 4–5 seconds (on average) to make judgments on both the Headline and Human summaries, as compared to about 13.4 seconds to make judgments on full text documents. This shows that it takes users almost 3 times longer to make judgments on full text documents as it took to make judgments on the summaries (Headline and Human). This finding is not surprising since text summaries are an order of magnitude shorter than full text documents.

In preparation for the correlation studies (to be presented in Section 5.6)

further analysis was done to reduce the effect of outliers. Specifically, an average was computed over all judgments for each user (20 judgments $\times$ 3 events), thus producing 60 data points. These data points were then partitioned into either 1, 2, or 4 partitions of equal size. (Partitions of size four have 15 data points, partitions of size two have 30 data points, and the partition of size one has 60 data points per user—or a total of 600 datapoints across all 10 users). To ensure that these results did not depend on a specific partition, this same process was repeated using 10,000 different (randomly generated) partitions for partitions of size 2 and 4.

Partitioned data points of size four provided a high degree of noise reduction without compromising the size of the data set (15 points). Larger partition sizes would result in too few data points and compromise the statistical significance of the correlation results. In order to show the variation within a single partition, the partitioning of size 4 with the smallest mean square error on the Headline surrogate compared to the other partitionings was used as a representative partition.

For this representative 15-fold partitioning, the individual data points are shown for each of the two agreement measures in Tables 5.4 and 5.5. This shows that, across partitions, the maximum and minimum Relevance-Prediction rates for Headline (0.93 and 0.60) are higher than the corresponding LDC-Agreement rates (0.85 and 0.50). The same trend is seen with the Human surrogates: Relevance Prediction has a maximum of 0.98 and a minimum of 0.68; and LDC Agreement has a maximum 0.88 and a minimum of 0.55. This provides further support for the hypothesis that Relevance Prediction is more reliable than that LDC Agreement

System	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	Avg
Headline	.10	.23	.13	.27	.20	.24	.26	.22	.13	.08	.30	.16	.26	.27	.30	**.211**
Human	.16	.22	.17	.23	.19	.36	.39	.29	.28	.25	.37	.22	.22	.39	.27	**.269**

Table 5.6: Average ROUGE-1 Scores for Headline and Human Surrogates (Representative Partition of Size 4)

for evaluation of summary usefulness.

5.5 Automatic Intrinsic Evaluation

To correlate the partitioned agreement scores above with the intrinsic measure, ROUGE was first run on all 120 surrogates in the experiment (i.e., the Human and Headline surrogates for each of the 60 event/document pairs) and then the ROUGE scores were averaged for all surrogates belonging to the same partitions (for each of the three partition sizes). These partitioned ROUGE values were then used for detecting correlations with the corresponding partitioned agreement scores described above.

Table 5.6 shows the ROUGE scores, based on 3 reference summaries per document, for partitions P1–P15 used in the previous tables.[6] The ROUGE 1-gram measurement (R1) is included here. ROUGE 2-gram, ROUGE L and ROUGE W were also computed, but the trend for these did not differ from ROUGE-1. The ROUGE scores for Headline surrogates were slightly lower than those for Human surrogates. This is consistent with the earlier statements about the difference be-

[6]A total of 180 human-generated reference summaries (3 for each of 60 documents) were commissioned (in addition to the human generated summaries used in the experiment).

tween non-extractive "eye catchers" and informative Headlines. Because ROUGE measures whether a particular summary has the same words (or n-grams) as a reference summary, a more constrained choice of words (as found in the extractive Human surrogates) makes it more likely that the summary would match the reference.

A summary in which the word choice is less constrained—as in the non-extractive Headline surrogates—is less likely to share n-grams with the reference. Thus, non-extractive summaries can be found that have almost identical meanings, but very different words. This raises the concern that ROUGE may be highly sensitive to the style of summarization that is used. Section 5.7 discusses this point further.

5.6 Correlation of Intrinsic and Extrinsic Measures

To test whether ROUGE correlates more highly with Relevance Prediction than with LDC Agreement, the correlation for the results of both techniques were calculated using Pearson's r (for a full definition, refer back to Section 4.1.7).

Table 5.7 shows the Pearson Correlations with ROUGE-1 for Relevance Prediction and LDC Agreement. For Relevance Prediction, a positive correlation for both surrogate types was observed, with a slightly higher correlation for Headline than Human. For LDC Agreement, no correlation (or a minimally negative one) was observed with ROUGE-1 scores, for both the Headline and Human surrogates.

Surrogate	P = 1	P = 2	P = 4
Headline (RP)	**0.1270**	**0.1943**	**0.3140**
Human (RP)	0.0632	0.1096	0.1391
Headline (LDC)	-0.0968	-0.0660	-0.0099
Human (LDC)	-0.0395	-0.0236	-0.0187

Table 5.7: Pearson Correlations with ROUGE-1 for Relevance Prediction (RP) and LDC Agreement (LDC), where Partition size (P) = 1, 2, and 4

Surrogate	P = 1	P = 2	P = 4
Headline (RP)	0.1020	0.2628	**0.3799**
Human (RP)	0.1297	**0.3446**	0.2611
Headline (LDC)	**-0.1520**	-0.2364	-0.1821
Human (LDC)	-0.0669	0.0463	0.1166

Table 5.8: Spearman Correlations with ROUGE-1 for Relevance Prediction (RP) and LDC Agreement (LDC), where Partition size (P) = 1, 2, and 4

The highest correlation was observed for Relevance Prediction with the Headline system.

The Spearman ρ correlations with ROUGE 1-gram for Relevance Prediction and LDC Agreement are shown in Table 5.8. Again, we see positive and higher correlations with Relevance Prediction than with LDC Agreement. The highest correlation is seen with Relevance Prediction and the Headline system for partition of size 4.

These results show that the ROUGE 1-gram measure correlates more highly with the Relevance-Prediction measurement than the LDC-Agreement measurement, although it must be noted that none of the correlations in Tables 5.7 and 5.8 were statistically significant at $p < 0.05$. The low LDC-Agreement scores are consistent with previous studies where poor correlations were attributed to low

interannotator agreement rates.

5.7 Discussion

These results suggest that ROUGE may be sensitive to the style of summarization that is used. As observed above, many of the Headline surrogates were not actually summaries of the full text, but were eye-catchers. Often, these surrogates did not allow the user to judge relevance correctly, resulting in lower agreement. In addition, these same surrogates often did not use a high percentage of words that were actually from the story, resulting in low ROUGE scores. (It was noticed that most words in the Human surrogates appeared in the corresponding stories.) There were three consequences of this difference between Headline and Human: (1) The rate of agreement was lower for Headline than for Human; (2) The average ROUGE score was lower for Headline than for Human; and (3) The correlation of ROUGE scores with agreement was higher for HEAD than for Human.

A further analysis supports the (somewhat counterintuitive) third point above. Although the ROUGE scores of true positives (and true negatives) were significantly lower for Headline surrogates (0.2127 and 0.2162) than for Human surrogates (0.2696 and 0.2715), the number of false negatives was substantially higher for Headline surrogates than for Human surrogates. These cases corresponded to much lower ROUGE scores for Headline surrogates (0.1996) than for Human (0.2586) surrogates.

Judgment (Surr/Doc)	Headline			Human		
	Raw	R1-Avg	Time (s)	Raw	R1-Avg	Time (s)
Rel/Rel	211 (35%)	0.2127 (±0.120)	4.6	**251 (42%)**	0.2696 (±0.130)	4.2
Rel/NonRel	27 (5%)	0.2115 (±0.110)	7.1	35 (6%)	**0.2725 (±0.131)**	4.6
NonRel/Rel	117 (19%)	0.1996 (±0.127)	**8.5**	77 (13%)	0.2586 (±0.120)	**13.8**
NonRel/NonRel	**245 (41%)**	**0.2162 (±0.126)**	2.5	237 (39%)	0.2715 (±0.131)	1.9
TOTAL	600 (100%)	0.2115 (±0.124)	4.6	600 (100%)	0.2691 (±0.129)	4.6

Table 5.9: Users' Judgments and Corresponding Average ROUGE-1 Scores

A more detailed analysis of the users' judgments and the corresponding ROUGE-1 scores is given in Table 5.9, where true positives and negatives are indicated by Rel/Rel and NonRel/NonRel, respectively, and false positives and negatives are indicated by Rel/NonRel and NonRel/Rel, respectively. The (average) elapsed times for summary judgments in each of the four categories are also included. One might expect a "relevant" judgment to be much quicker than a "non-relevant" judgment (since the latter might require reading the full summary). However, it turned out non-relevant judgments did not always take longer. In fact, the NonRel/NonRel cases took considerably less time than the Rel/Rel and Rel/NonRel cases. On the other hand, the NonRel/Rel cases took considerably more time—almost as much time as reading the full text documents—an indication that the users may have re-read the summary a number of times, perhaps vacillating back and forth. Still, the overall time savings was significant, given that the vast majority of the non-relevant judgments were in the NonRel/NonRel category.

In Table 5.9 the numbers in parentheses after each ROUGE score refer to

the standard deviation for that score. This was computed as follows:

$$Std.\text{-}Dev. = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}}$$

where N is the number of surrogates in a particular judgment category (e.g., $N = 245$ for the Headline-based NonRel/Rel judgments), x_i is the ROUGE score for the i^{th} surrogate, and $\bar{r}$ is the average of all ROUGE scores in that category.

Although there were very few false positives (less than 6% for both Headline and Human), the number of false negatives (NonRel/Rel) was particularly high for Headline (50% higher than for Human). This difference was statistically significant at $p<0.01$ using the t-test. The large number of false negatives with Headline may be attributed to the eye-catching nature of these surrogates. A user may be misled into thinking that this surrogate is not related to an event because the surrogate does not contain words from the event description and is too broad for the user to extract definitive information (e.g., the surrogate *There he goes again!*). Because the false negatives were associated with the lowest average ROUGE score (0.1996), it is speculated that, if a correlation exists between Relevance Prediction and ROUGE, the false negatives may be a major contributing factor.

Based on this experiment, it is conjectured that ROUGE may not be a good method for measuring the usefulness of summaries when the summaries are not extractive. That is, if someone intentionally writes summaries that contain different words than the story, the summaries will also likely contain different words than a reference summary, resulting in low ROUGE scores. However, the summaries, if

well-written, could still result in high agreement with the judgments made on the full text.

5.8 Evaluation of Experimental Hypotheses

The first hypothesis was that the summaries would allow the users to achieve a Relevance Prediction rate of 70–90%. The resulting Relevance-Prediction scores were in this range and were significantly higher than the scores of the LDC-Agreement method, supporting this hypothesis.

The second hypothesis, that the Headline system would yield a significantly lower agreement rate than the Human system, was not supported by the data. The accuracy score of the Headline system (0.760) was lower than that of the Human system (0.813), but this result was not statistically significant with $p < 0.05$.

The third hypothesis was that the Relevance-Prediction method would be more reliable than the LDC-Agreement method. The results showed that there were statistically significant differences between the results of the two methods, and that the results of Relevance-Prediction were consistenly higher than those of the LDC-Agreement method, supporting this hypothesis.

The final hypothesis was that the judgment time with summaries would be much less than that of the full text. The results displayed almost a two-thirds reduction in time with the summaries, supporting this hypothesis.

Since this experiment focused only on evaluations with human-generated

summaries, the findings above will be further explored with both human-generated and automatic summaries—the focus of the experiment described in the next chapter.

Chapter 6

Relevance Prediction with Human and Automatic Summaries

In the previous experiment, the Relevance-Prediction method was introduced, tested and compared against the LDC-Agreement method for evaluation of text summarization. That experiment used only human summaries—the original document headline and human-generated summaries that were commissioned for the study. The experimental results showed that Relevance Prediction (RP) was more reliable than LDC Agreement for human summaries and that ROUGE may be sensitive to the type of summarization used (abstractive or extractive).

This chapter introduces a new study, RP Dual Summary that uses both human and automatic summaries to further compare the Relevance-Prediction and LDC-Agreement methods and explore how these correlated with automatic intrinsic metrics. In addition to the two human-generated systems of the previous experiment, four automatic summarizers are included as part of the evaluation.

6.1 Hypotheses

The first hypothesis is that the Relevance-Prediction method will achieve significantly higher results than the LDC-Agreement method, consistent with the findings of the previous experiment. The RP with Human Summaries experiment determined that the Relevance-Prediction method yielded scores that were at least 8.7% higher than the LDC-Agreement scores, and that these results were statistically significant with $p < 0.01$. Although this experiment includes four automatic systems that were not present in the previous experiment, it is expected that the trend in scoring of the two methods will be similar.

The second hypothesis is that the human-generated summaries (Headline and Human systems) will achieve at least a 5% higher Relevance-Prediction rate than the automatic summaries. As can be seen with the previous experiments, the results of the human-generated systems are consistently higher than that of the automatic generated summaries. It is believed that human summarizers know how to easily identify and extrapolate the most important information in a document, which is often a problem for automatic summarizers. Also, human generated summaries are often formatted in a more fluent and easily readable manner than automatic summaries.

The third hypothesis is that the First75 system will produce results that are slightly lower than that of the human-generated summaries, but will be the highest of the four automatic systems. This system was initially thought to serve

as a lower baseline for the automatic systems, but has been shown in a previous experiment (in Table 4.19) to produce much higher agreement scores than were expected.

The final hypothesis is that the Relevance-Prediction accuracy measure will generate moderate correlations (0.4 or higher) with the ROUGE metric (specifically the ROUGE-1 measure) and that these correlations will be significantly higher than those of the LDC-Agreement measure. This hypothesis stems from the results of the previous experiment, RP with Human Summaries, in which small positive correlations were found with the ROUGE-1 and the Headline and Human Relevance-Prediction scores (refer to Table 5.7).

6.2 Experiment Details

This experiment tests the Relevance-Prediction method using both human and automatic summaries. The experiment uses four types of automatically generated document surrogates; two types of manually generated surrogates; and, as a control, the entire text document. The systems are:

- **Full Text** – the full document itself (used as the upper baseline);
- **Headline** – a human-generated headline associated with the original document;
- **Human** – a human-generated 75 character generic summary written by a human (commissioned from University of Maryland students for this exper-

iment);

- **First75** – an automatic summary that uses the first 75 characters of the document;
- **HMM** – a statistical summarization system developed by UMD and BBN;
- **Trimmer** – Fluent headline based on a linguistically-motivated parse-and-trim approach (Dorr et al., 2003);
- **Topiary** – a system for generating short summaries by combining Trimmer or HMM sentence compressions with statistically generated topic terms from Unsupervised Topic Detection (Zajic et al., 2004a).

The data for the experiment were taken from the Topic Detection and Tracking 3 corpus (Allan et al., 1999) and consisted of three event descriptions and associated full text documents. Twenty documents were used for each event description, and contained an even split of relevant/nonrelevant documents as judged by the annotators at the Linguistic Data Consortium (LDC, 2006).

6.3 Experiment Design

For this study, six participants were recruited through email and posted advertisements at the University of Maryland. The participants were tasked with evaluating summaries produced by two human systems and four automatic systems and the corresponding source text documents.

	Topic 1	**Topic 2**	**Topic 3**
Headline, HMM	Users 1 & 2	Users 3 & 4	Users 5 & 6
Human, Trimmer	Users 3 & 4	Users 5 & 6	Users 1 & 2
First75, Topiary	Users 5 & 6	Users 1 & 2	Users 3 & 4

Table 6.1: RP Dual Experiment Design

The experiment consisted of 3 event and document sets. Each event set contained 20 documents to be judged *relevant* or *not relevant* to the specified event. For each event set, the users were asked to make judgments from 3 types of systems; a manual system (either Headline, Human, or First75), an automatic system (either HMM, Topiary, or Trimmer) and the text document (Full Text). Once a user made judgments with a particular system and event set (with the exception of the Full Text), that system was exhausted and not used for the remaining events. This way, all 6 summary systems would be judged (2 with each of the 3 events), and the full text was judged for all events. An example of this design is shown in Table 6.1. The order in which the events were displayed were randomized for each user, and the 40 total summaries presented within each event (20 summaries for each of two summary systems) were also randomized. The full text documents were always displayed last, so to not bias the user's judgments on the summaries.

The users made one judgment per document. The three events in the experiment contained 20 documents each, for a total of 60 distinct documents in the experiment. Since each user made judgments with three systems per event

(a manual system, an automatic system, and the full text), this gave a total of 180 judgments per user. The users were allowed as much time as they needed to make a judgment, and the judgment time for each document was recorded to allow comparisons of average judgment time for the systems.

The gold standard judgments produced by the LDC Annotators and included as part of the Topic Detection and Tracking 3 (TDT-3) corpus are used to compute LDC Agreement. For Relevance Prediction, the judgments each user made on a summary is compared against his/her judgment on the corresponding full text document.

6.4 Results and Analysis

The users made binary judgments, deciding whether the information presented in each summary was "relevant" or "not relevant" to the specified event. Using the contingency table, Table 3.1, the extrinsic measures used for this experiment are: accuracy, precision, f-score, recall/sensitivity, specificity and d'. The results with the first four measures using LDC Agreement are shown in Table 6.2 and using Relevance Prediction are displayed in Table 6.3. The results for sensitivity, specificity, and d' using LDC Agreement are in Table 6.7 and using Relevance Prediction are in Table 6.8.

One-factor repeated-measures ANOVA tests (with 35 degrees of freedom) were computed to determine if the differences among the systems were statistically

System	TP	FP	FN	TN	A	P	R	F	T (s)
Full Text	**161**	**27**	19	**153**	**0.872**	**0.856**	**0.894**	**0.875**	**159.88**
Headline	50	10	10	50	0.833	0.833	0.833	0.833	7.03
Human	45	11	15	49	0.783	0.804	0.750	0.776	6.87
First75	38	7	22	53	0.758	0.844	0.633	0.724	7.60
Topiary	38	7	22	53	0.758	0.844	0.633	0.724	5.88
Trimmer	39	8	21	52	0.758	0.830	0.650	0.729	5.73
HMM	34	13	**26**	47	0.675	0.723	0.567	0.636	6.93
HSD, p<0.05					——	——	——	——	19.37

Table 6.2: Results of Extrinsic Task Measures sorted by **A**ccuracy (using LDC Agreement)

System	TP	FP	FN	TN	A	P	R	F	T (s)
Human	49	7	8	**56**	**0.875**	0.875	**0.860**	**0.867**	6.87
Headline	**52**	8	14	46	0.817	0.867	0.788	0.825	7.03
First75	43	2	23	52	0.792	**0.956**	0.652	0.775	**7.60**
Trimmer	39	8	17	**56**	0.792	0.830	0.696	0.757	5.73
Topiary	41	4	25	50	0.758	0.911	0.621	0.739	5.88
HMM	34	**13**	**32**	41	0.625	0.723	0.515	0.602	6.93
HSD, p<0.05					0.121	——	0.232	0.219	—

Table 6.3: Results of Extrinsic Task Measures sorted by **A**ccuracy (using Relevance Prediction)

significant for the five measures: precision, recall, f-score, accuracy, and time. If significant differences were found with $p < 0.05$ using the ANOVA, the Tukey Honestly Significant Difference (HSD) was then computed to determine exactly which systems display significant differences. The results of the Tukey HSD test is displayed in the last row of each table, only for measures that have statistically significant differences.

Table 6.2 shows no significant differences between the systems with any of the extrinsic measures for LDC Agreement using the Tukey test. However, Table 6.3 shows significant differences with $p < 0.05$ between at least two systems for

Human	A	
Headline	A	
First75	A	
Trimmer	A	
Topiary	A	
HMM		B

Table 6.4: Equivalence Classes for Relevance Prediction with the Accuracy measure

Human	A		
Headline	A	B	
Trimmer	A	B	C
First75	A	B	C
Topiary		B	C
HMM			C

Table 6.5: Equivalence Classes for Relevance Prediction with the Recall measure

the accuracy, recall, and f-score measures. For accuracy, statistically significant differences were found between the HMM system and the other systems. HMM is a sole member of set B and the remaining systems are members of set A in the associated equivalence class listing of Table 6.4.

For recall, significant differences are seen between Human and Topiary, Human and HMM, and Headline and HMM. These equivalence class results are displayed in Table 6.5 where the systems are grouped into three sets, A, B, and C. For

Human	A	
Headline	A	
First75	A	B
Trimmer	A	B
Topiary	A	B
HMM		B

Table 6.6: Equivalence Classes for Relevance Prediction with the F-Score measure (for Non-Strict Scoring)

System	Sensitivity	Specificity	discriminability index (d')
Headline	**0.833**	0.833	**1.935**
Human	0.750	0.817	1.577
First75	0.633	**0.883**	1.533
Topiary	0.633	**0.883**	1.533
Trimmer	0.650	0.867	1.496
HMM	0.567	0.783	0.951

Table 6.7: Results for Signal Detection Measures Using LDC Agreement, sorted by d'

f-score, the equivalence class results in Table 6.6 include only two overlapping sets, A and B, which show significant differences with $p < 0.05$ between Human and HMM, and Headline and HMM. Therefore, the Relevance-Prediction method again appears to be more reliable in that it produces results with significant differences between systems while LDC Agreement does not.

Significant differences are seen in Table 6.3 between systems with the Relevance-Prediction method, but the resulting scores are not consistently higher than that of LDC Agreement. We see that the Human and First75 scores are higher with Relevance Prediction, yet the Headline scores are lower. For the automatic measures, Trimmer achieves a higher score with Relevance Prediction, the Topiary score remains the same, and the HMM score is lower than that of LDC Agreement. Therefore the first hypothesis, that the Relevance-Prediction method would achieve significantly higher results than LDC Agreement was not confirmed.

Recall that the second hypothesis is that the human-generated summaries will achieve at least a 5% higher Relevance-Prediction rate than the automatic

System	Sensitivity	Specificity	discriminability index (d')
Human	**0.860**	0.889	**2.299**
First75	0.652	**0.963**	2.176
Headline	0.788	0.852	1.843
Trimmer	0.621	0.926	1.755
Topiary	0.696	0.875	1.665
HMM	0.515	0.759	0.742

Table 6.8: Results for Signal Detection Measures Using Relevance Prediction, sorted by d'

summaries. Table 6.3 displays a Relevance-Prediction rate of 0.875 for the Human summaries, which reflects more than a 5% rate increase than the highest automatic system score of 0.792 (the score of the First75 and Topiary systems). However, the Headline rate of 0.817 achieves a rate increase of only 2.5% above the highest-scoring automatic systems.

The third hypothesis was that the First75 system would rank third for the measures, in that it will produce slightly lower results than that of the human-generated summaries but the highest results of the four automatic systems. The LDC-Agreement results in Table 6.2 and the Relevance-Prediction results in Table 6.3 both show the First75 system tied for a third place ranking with other automatic systems for the accuracy measure. Table 6.7 also has First75 tied in third place with another automatic system for LDC Agreement with the d' measure, while the Relevance-Prediction results in Table 6.8 has First75 ranked second. These results did not support the third hypothesis.

System	R1	R2	RL	RW
Full Text	**0.76145**	**0.33990**	**0.64880**	**0.34598**
Human	0.24613	0.07288	0.21061	0.11990
First75	0.23047	0.08953	0.20277	0.11860
Topiary	0.22901	0.08783	0.19791	0.11236
Trimmer	0.22532	0.08420	0.19833	0.11366
HMM	0.20650	0.06588	0.18331	0.10401
Headline	0.18540	0.05477	0.16674	0.09461
HSD, $p<0.05$	0.1517	0.1810	0.1259	0.0634

Table 6.9: ROUGE Recall Results on the Seven Systems, Sorted by ROUGE-1

6.5 Automatic Intrinsic Evaluation

A newer version of the ROUGE metric that was not previously available, ROUGE 1.5.4, was used as the first intrinsic metric for this experiment. Since ROUGE is normally more heavily weighted towards recall, this new version offered three different intrinsic measures; ROUGE-Recall (the same measure of the previous experiments), ROUGE-Precision (more heavily weighted towards precision) and ROUGE-F-Score (heavily weighted towards the harmonic mean of precision and recall). ROUGE-Precision is most useful when there is no limit on summary length. Since this is not the case in this experiment, the results of this measure are not included here. The results using ROUGE-Recall are presented in Table 6.9 and Figure 6.1. The results using ROUGE-F-Score are presented in Table 6.10 and Figure 6.2.

For ROUGE-Recall 1-gram (Table 6.9) we see that the Full Text and Human systems are ranked as the two highest systems, as would be expected. However, the Headline system is ranked last, receiving a lower score than each of the automatic

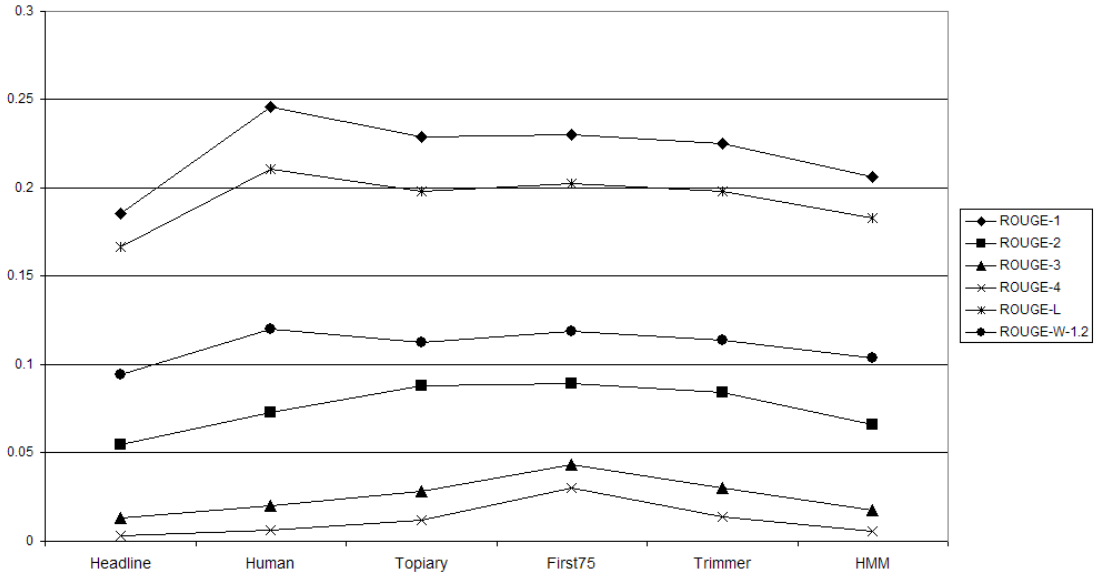

Figure 6.1: ROUGE Recall Results

systems. As was discussed in the previous experiment, this can be attributed to the abstractive, eye-catching nature of the news story headlines. Words that do not appear in the source text may be used in the headlines, with the intent of inciting an observer's interest in reading the entire article. Since the automatic systems are generally extractive, using words and phrases directly from the source text, it is more likely that these summaries would match words in the reference summaries.

Similarly, the results for ROUGE-F-Score 1-gram (Table 6.10) shows that the Headline system receives a lower score than most of the other systems, outperforming only HMM and the Full Text. Note that although Full Text was the highest performing system for ROUGE-Recall, it is the lowest performing system with ROUGE-F-Score. This is unsurprising in that scoring for recall is based

System	R1	R2	RL	RW
Human	**0.24672**	0.07195	**0.21173**	**0.14831**
Topiary	0.22088	**0.08627**	0.19039	0.13518
Trimmer	0.21987	0.08429	0.19415	0.13831
First75	0.21690	0.08572	0.19188	0.14027
Headline	0.21321	0.06449	0.19255	0.13023
HMM	0.19841	0.06279	0.17707	0.12516
Full Text	0.04457	0.02022	0.03750	0.03264
HSD, p<0.05	0.1138	0.0753	0.0871	0.0632

Table 6.10: ROUGE F-Score Results on the Seven Systems, Sorted by ROUGE-1

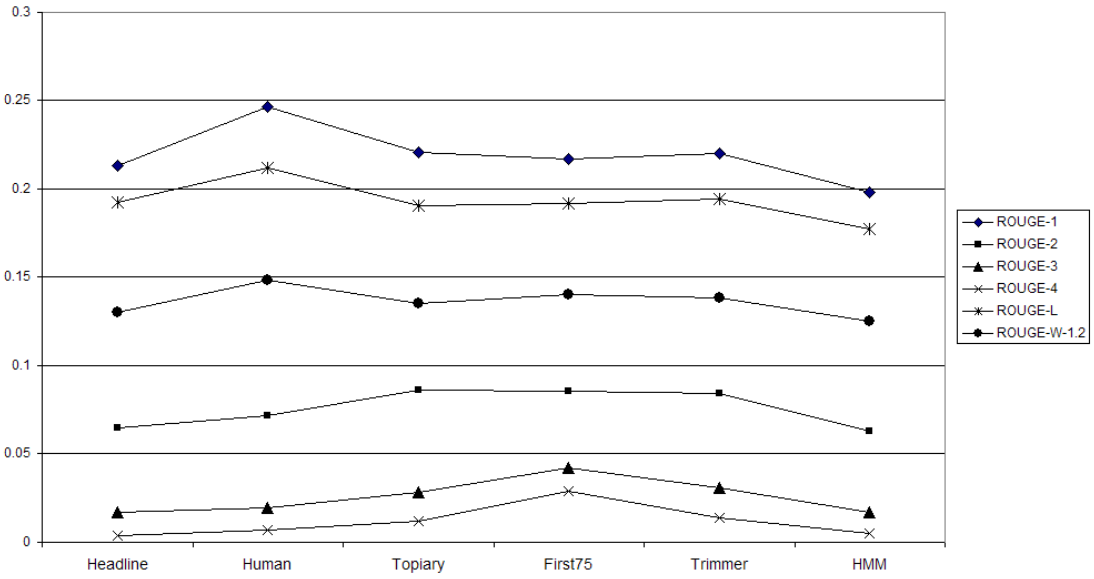

Figure 6.2: ROUGE F-Score Results

System	Score
Trimmer	**0.133**
First75	0.115
Topiary	0.113
Human	0.096
Headline	0.083
HMM	0.065

Table 6.11: Basic Elements Results on the Seven Systems

on dividing the term matching score with the number of words in the reference summary. Scoring for precision is based on the dividing the term matching score with the number of words in the candidate summary (meaning that a candidate summary is penalized for being longer than the reference). Since the Full Text is always longer than the reference summary (often longer by more than an order of magnitude), and because f-score is the harmonic mean of precision and recall, the understandably low precision score for Full Text drastically reduces the resulting f-score.

Statistically significant differences with $p < 0.05$ are seen with all the measures of ROUGE-Recall and ROUGE-F-Score, but the Tukey HSD test results show that these differences are mainly seen between the Full Text and the summary systems. The only exception is the ROUGE-F-Score 2-gram measure, where significant differences are not found between any of the systems using the Tukey test.

Recall that both of the intrinsic metrics that were previously investigated, BLEU and ROUGE, rely on exact term matching for scoring. One of the issues of

	ROUGE-Recall				ROUGE-F-Score			
	R1	**R2**	**RL**	**RW**	**R1**	**R2**	**RL**	**RW**
Basic Elements	0.537	0.805	0.570	0.603	0.373	0.922	0.346	0.559

Table 6.12: Pearson Correlations for the results of Basic Elements and ROUGE

	ROUGE-Recall				ROUGE-F-Score			
	R1	**R2**	**RL**	**RW**	**R1**	**R2**	**RL**	**RW**
Basic Elements	0.429	0.771	0.543	0.543	0.486	0.771	0.371	0.600

Table 6.13: Spearman Correlations for the results of Basic Elements and ROUGE

exact term matching is that credit is not given to summaries that effectively communicate the meaning of the reference texts while using synonymous terms or concepts. Because of this issue, two new content-based methods, the Pyramid Method (described in Section 2.2.2.3) and Basic Elements (described in Section 2.2.2.4), were created. The Pyramid Method is semi-automatic in that it relies heavily on human labor for the identification and creation of the clauses used as the basis for scoring. The Basic Elements method also uses the idea of semantic comparisons using small phrasal and clausal units, while achieving this through a fully automated process. The systems and summaries of this experiment were also evaluated with Basic Elements version 1.1, and the results are shown in Table 6.11.

These results greatly differ from the ROUGE-Recall and F-Score results. Here we see the Human system being ranked lower than three of the automatic systems, including the baseline First75 system. The developers of the BE method claim high correlations (>0.889) with the ROUGE metric, however, it can be seen in Tables 6.12 and 6.13 that the respective Pearson and Spearman correlation of the BE scores and the ROUGE metric are moderate and much lower than the

claims for most cases. A possible explanation for these results is that the version of Basic Elements used in the developer's and DUC evaluations (Hovy et al., 2006) is not the version of Basic Elements that is commercially available and used in this analysis. Some of the advanced modules in the developer's version are not provided in the public version, and may explain why these results and correlations with the ROUGE metric differ so greatly from the reported results of the developers. Since this version of Basic Elements did not produce the level of results as suggested, the method will not be used for intrinsic evaluations in the next experiment.

6.6 Correlation of Intrinsic and Extrinsic Measures

For correlations, Pearson r is the method most widely used in the summarization evaluation community. Spearman ρ is a more fitting method based on the data (in that the intrinsic measures produce results that are ordinal in nature). A detailed description of Pearson r is given in Section 4.1.7 and Spearman ρ is given in Section 4.2.7. Consistent with the previous experiments, both results will be reported and discussed here. Since the Relevance-Prediction method uses the judgments of the Full Text as the gold-standard for scoring, the results of the Full Text system are excluded from these analyses.

The Pearson r correlations with the human performance scores and the two versions of ROUGE, ROUGE-Recall and ROUGE-F-Score, are shown in Tables 6.14 and 6.15, respectively. Recall that the fourth hypothesis was that the

Relevance-Prediction accuracy measure would generate moderate correlations (0.4 or higher) with the ROUGE metric (specifically the ROUGE-1 measure) and that these correlations would be significantly higher than those of the LDC-Agreement measure. The correlation with Relevance Prediction and the 1-gram measure for ROUGE-F-Score is 0.846, much higher than the hypothesized result. The correlation with the Relevance-Prediction accuracy measure for the 1-gram measure with ROUGE-Recall is 0.370, only slightly lower than the hypothesis. Therefore, the hypothesis is confirmed in terms of ROUGE-F-Score, but not for ROUGE-Recall.

The same trends are seen with the Spearman ρ correlations using LDC Agreement and Relevance Prediction, shown in Table 6.16 for ROUGE-Recall and Table 6.17 for ROUGE-F-Score. The correlation with the accuracy measure for Relevance Prediction and ROUGE-F-Score is 0.459, higher than the hypothesis, but the result of 0.359 for ROUGE-Recall is slightly lower than the hypothesis.

Although the first hypothesis—that the Relevance-Prediction method would consistently produce higher scores than LDC Agreement—was not confirmed, note that the Pearson correlations of Tables 6.14 and 6.15 are significantly higher for Relevance Prediction than for LDC Agreement (with $p < 0.05$, using a paired t-test) in all cases except the ROUGE-F-Score and extrinsic Recall correlations. Similarly, the Spearman correlations using ROUGE Recall are significantly higher (with $p < 0.05$) for Relevance Prediction than for LDC Agreement using the extrinsic Precision, Recall and F-Score measures.

Table 6.15 shows that the Pearson r correlation results with the Relevance-

	LDC Agreement				Relevance Prediction			
	Accuracy	**Precision**	**Recall**	**F-Score**	**Accuracy**	**Precision**	**Recall**	**F-Score**
ROUGE-1	-0.151	0.194	-0.292	-0.178	**0.370**	0.389	0.185	**0.270**
ROUGE-2	**-0.233**	**0.449**	**-0.565**	**-0.340**	0.084	**0.518**	**-0.251**	-0.031
ROUGE-L	-0.180	0.196	-0.334	-0.213	0.347	0.389	0.148	0.241
ROUGE-W	-0.169	0.240	-0.346	-0.211	0.350	0.444	0.128	0.244

Table 6.14: Pearson r Correlation between Intrinsic and Extrinsic Scores using ROUGE Recall

	LDC Agreement				Relevance Prediction			
	Accuracy	**Precision**	**Recall**	**F-Score**	**Accuracy**	**Precision**	**Recall**	**F-Score**
ROUGE-1	0.489	0.404	0.446	0.506	0.846	0.490	0.799	0.807
ROUGE-2	0.049	**0.695**	-0.331	-0.072	0.269	**0.664**	-0.080	0.163
ROUGE-L	**0.620**	0.444	**0.596**	**0.644**	**0.924**	0.495	**0.905**	**0.899**
ROUGE-W	0.374	0.472	0.251	0.362	0.810	0.599	0.672	0.742

Table 6.15: Pearson r Correlation between Intrinsic and Extrinsic Scores using ROUGE F-Score

Prediction accuracy measure and ROUGE-F-Score are very high—approaching an almost perfect correlation—with scores of 0.846 for 1-gram, 0.924 for ROUGE-L, and 0.810 for ROUGE-W-1.2. For the extrinsic recall and f-score measures, correlations of the three intrinsic measures (1-gram, L and W-1.2) are also high, ranging from 0.672 to 0.905. Similarly, the Spearman ρ correlations of Table 6.17 for Relevance Prediction and ROUGE-L approach a perfect correlation, with a resulting score of 0.919 for accuracy, 0.943 for recall and 0.829 for f-score.

Recall that a finding of the previous experiment (described in Section 5.7) was that ROUGE may be sensitive to the type of summarization used (abstractive

	LDC Agreement				Relevance Prediction			
	Accuracy	**Precision**	**Recall**	**F-Score**	**Accuracy**	**Precision**	**Recall**	**F-Score**
ROUGE-1	**0.128**	0.207	0.000	0.000	0.345	0.657	0.257	0.371
ROUGE-2	-0.064	**0.673**	**-0.305**	**-0.305**	-0.115	**0.714**	-0.257	-0.143
ROUGE-L	**0.128**	0.052	0.061	0.061	**0.459**	0.486	**0.371**	**0.429**
ROUGE-W	**0.128**	0.052	0.061	0.061	**0.459**	0.486	**0.371**	**0.429**

Table 6.16: Spearman ρ Correlation between Intrinsic and Extrinsic Scores using ROUGE Recall

	LDC Agreement				Relevance Prediction			
	Accuracy	**Precision**	**Recall**	**F-Score**	**Accuracy**	**Precision**	**Recall**	**F-Score**
ROUGE-1	0.447	0.207	0.366	0.366	0.459	0.486	0.486	0.429
ROUGE-2	0.256	**0.828**	0.000	0.000	0.000	**0.771**	-0.086	-0.029
ROUGE-L	**0.703**	-0.155	**0.794**	**0.794**	**0.919**	0.086	**0.943**	**0.829**
ROUGE-W	0.447	0.207	0.366	0.366	0.689	0.600	0.600	0.657

Table 6.17: Spearman ρ Correlation between Intrinsic and Extrinsic Scores using ROUGE F-Score

	ROUGE-Recall		ROUGE-F-Score	
	LDC (Accuracy)	**RP (Accuracy)**	**LDC (Accuracy)**	**RP (Accuracy)**
First75	0.549	-0.955	**0.990**	-0.413
Headline	0.678	0.518	0.901	0.166
HMM	0.791	**0.998**	0.386	0.896
Human	-0.220	-0.734	-0.442	-0.555
Topiary	**0.780**	0.780	0.782	0.782
Trimmer	0.621	-0.896	0.644	**-0.909**

Table 6.18: Pearson r Correlation between Extrinsic Scores and ROUGE for Each System

or extractive). To examine this possibility further, Pearson correlations for each system were also produced.[1] These new correlations are given in Table 6.18 for ROUGE and the extrinsic accuracy measure (for LDC Agreement and Relevance Prediction).

These results vary widely between positive and negative correlations, and high and low correlation. However, the previous correlation results (for all systems) in Tables 6.14 and 6.15 showed more consistent moderate correlations. These differences suggest that ROUGE can order the systems in a manner that moderately reflects the ordering produced by the extrinsic measures. This also suggests that

[1]These additional correlations partitions the data for each of the six summarization systems into 3 groups containing 20 documents each. The previous ROUGE and extrinsic measure correlations use a single averaged score for each system, reflecting correlations based on system ordering.

	LDC Agreement				Relevance Prediction			
	Accuracy	**Precision**	**Recall**	**F-Score**	**Accuracy**	**Precision**	**Recall**	**F-Score**
First75	-0.228	-0.162	0.071	-0.041	-0.080	0.162	0.152	0.196
Headline	-0.436	-0.244	0.118	0.145	**-0.331**	-0.149	0.004	-0.127
HMM	0.106	-0.162	**0.508**	0.191	-0.074	-0.265	0.326	0.139
Human	**-0.622**	0.210	-0.338	**0.282**	0.000	**-0.286**	**-0.328**	**-0.314**
Topiary	0.063	-0.289	0.233	-0.069	-0.242	-0.230	-0.121	-0.170
Trimmer	-0.229	**-0.313**	0.005	-0.223	-0.165	-0.027	-0.049	-0.015

Table 6.19: Pearson r Correlation between Extrinsic Scores and Results of the BE Method

	LDC Agreement				Relevance Prediction			
	Accuracy	**Precision**	**Recall**	**F-Score**	**Accuracy**	**Precision**	**Recall**	**F-Score**
First75	-0.321	**-0.364**	-0.260	**-0.383**	-0.275	-0.218	-0.241	-0.243
Headline	-0.114	-0.144	0.022	0.029	-0.255	-0.300	0.050	-0.361
HMM	-0.041	0.051	**0.382**	0.016	-0.214	-0.248	0.107	0.063
Human	-0.423	0.288	-0.127	0.288	**-0.483**	**-0.515**	**-0.648**	**-0.566**
Topiary	-0.099	-0.270	0.075	-0.231	-0.273	-0.321	-0.187	-0.289
Trimmer	**-0.459**	-0.320	-0.153	-0.330	-0.184	-0.013	-0.169	-0.133

Table 6.20: Spearman ρ Correlation between Extrinsic Scores and Results of the BE Method

the ROUGE document-level results do not accurately reflect the results produced by the extrinsic measures.

Table 6.19 and 6.20 display the Pearson and Spearman correlations, respectively, for Basic Elements with LDC Agreement and Relevance Prediction. The correlations for most cases are negative, indicating an inverse relationship with the Basic Elements results and the results of the extrinsic measures.

6.7 Experimental Findings

The first hypothesis of this experiment was that Relevance Prediction would achieve significantly higher results than those of the LDC-Agreement method. Although the results for Relevance Prediction for some of the systems were higher than that of LDC Agreement, but these differences were not statistically signifi-

cant, and did not lend support to this hypothesis.

Recall that a finding of the previous experiment and the second hypothesis of this experiment was that the Relevance-Prediction method produced results that were significantly higher (with $p < 0.05$) than the LDC-Agreement results (discussed in Section 5.4). Although the same trend was not discovered in this experiment, the results of the Relevance-Prediction method showed significant differences *between* systems (with $p < 0.05$) when the results with the LDC-Agreement method did not. This finding again supports the claim that Relevance Prediction is more reliable in that it distinguishes between systems better than LDC Agreement.

Another finding of the previous experiment (in Section 5.6) was that ROUGE may be sensitive to summarization style. The findings of this experiment show that the ROUGE and extrinsic correlations for partitioned data (in Table 6.18) vary widely, but the ROUGE and extrinsic correlations for system ordering (in Tables 6.14 through 6.17) are more consistent. This suggests that the system rankings produced by ROUGE may accurately reflect rankings produced by human extrinsic task evaluations, but that the partitioned ROUGE scores are not reflective of partitioned human extrinsic task results. This finding also supports the fourth hypothesis—that moderate correlations would be seen with the ROUGE metric and the extrinsic results.

It appears that the commercially available version of Basic Elements does not produce results that are reflective of human task-based performance. Evaluations using the proprietary software version may produce better correlations but since

use of that version is limited, independent evaluations with it are not currently possible.

This experiment and the previous experiments involved the application of several evaluation measures for summaries spanning a single document. The next chapter will investigate the application of these measures to summaries spanning multiple documents.

Chapter 7

Relevance Prediction with Multi-Document Summaries

The experiments of the previous chapters investigated the LDC-Agreement and Relevance-Prediction methods for single-document summaries. Typically, single-document summarization systems constrain the content of the resulting summaries to the relevant information specified in the source text and may not include all of the information to satisfy the query (especially if a source text does not contain enough relevant information).

More recently, the text summarization community has shifted its focus to multi-document summarization, where the resulting summaries are not as restricted in their coverage. Because multi-document summaries are generated from more than one source document, different types of relevant information that may appear in some documents but not in others may be gathered and used in the final summary. Therefore, multi-document summaries should provide a user with all or

most of the information requested in the query (in this case, the event description) from the source texts.

This chapter describes a new experiment—RP with Multi-Document Summaries—where the previous single-document Relevance-Prediction method is extended to the multi-document case. The evaluation approach in this experiment design is novel and differs from the previous judgment and scoring methods. For the single-document summary experiments, each document was pre-annotated as "relevant", "not relevant" or, in some cases, "maybe relevant." However, multiple documents within each topic may have varying levels of relevance to the query; some topics may contain 5 documents that were pre-annotated as "relevant" and 15 documents pre-annotated as "not relevant". To accommodate these differences, a five point likert scale is created to reflect the multiple relevance levels, four new scoring methods are introduced, and the LDC-Agreement and Relevance-Prediction methods are extended to incorporate the differences in the judgment scale.

7.1 Hypotheses

This experiment again focuses on comparing the results of the Relevance-Prediction method against that of the LDC-Agreement method, now in terms of multi-document summaries. In keeping with the expectations of the previous experiments, the first hypothesis is that the Relevance-Prediction method will produce accuracy scores that are at least 5% higher than those associated with

LDC Agreement and that the resulting differences are statistically significant at $p < 0.05$.

As noted in the previous experiments, the summaries of the human-generated systems, Headline and Human, are expected to produce higher accuracy results than the baseline system and the automatic systems. These results were seen with both methods in the last experiment (refer to Section 6.4 for details). Therefore, the second hypothesis for this experiment is that, consistent with the findings of the previous experiment, the human-generated systems (Headline and Human) will achieve higher accuracy scores than the baseline and automatic systems.

The third hypothesis pertains to the differences between Relevance Prediction and LDC Agreement, specifically that Relevance Prediction will have a higher correlation than LDC Agreement with the intrinsic measure and that these results will be statistically significant with $p < 0.05$.

7.2 Experiment Details

This experiment tests the Relevance-Prediction method using both human and automatic summaries that span 20 source documents. Nine participants were recruited through email advertisements and flyers posted on the University of Maryland College Park campus. The users evaluated three types of automatically generated document surrogates and three types of manually generated surrogates. As a control, the set of 20 source text documents texts were displayed after the

summaries. The systems are:

- **Baseline** – includes the first sentence of each text document, truncated at the 250 word maximum;
- **Headline** – includes the human-generated headline from each original document, truncated at the 250 word maximum;
- **Human** – consists of summaries written by a human focusing on the information requested in the event description (commissioned from University of Maryland students for this experiment);
- **SE** – selects the first five sentences of each document and creates a summary based on factors including the relevance of each sentence to the event description;
- **ISCC** – similar to the SE system but also includes rules for trimming and compressing the data;
- **WTMC** – similar to the ISCC system but includes enhanced methods for assigning weights to the selected sentences and determines the output using the optimized weights and compression rules.

The data for the experiment were taken from the (Topic Detection and Tracking version 3 (TDT-3) and version 4 (TDT-4) corpora (Allan et al., 1999). 81 events were selected from the corpora, with 1 event serving as the practice topic. Each

event set contained an event title, description, and twenty associated full text news documents. The levels of relevant/nonrelevant documents, as judged by the annotators at the Linguistic Data Consortium (LDC, 2006), were equally represented with 16 events assigned to each of the following distributions:

- 20 relevant documents and 0 nonrelevant documents;
- 15 relevant documents and 5 nonrelevant documents;
- 10 relevant documents and 10 nonrelevant documents;
- 5 relevant documents and 15 nonrelevant documents;
- 0 relevant documents and 20 nonrelevant documents.

As noted in the previous section, multi-document summaries capture the content of a large set of source documents, usually in response to a user's query. Since the source data set is so large (more than an order of magnitude greater than that of single-document summaries), it is understandable that the resulting output would be similarly increased. In the single-document experiments described in Chapters 4 through 6, the maximum summary size was 75 characters. For this experiment, the maximum summary size is 250 words, consistent with the guidelines for multi-documents summaries as part of the 2006 Document Understanding Conference (DUC) (Dang, 2006).

7.3 Experiment Design

This task was conducted on a PC using an Internet Explorer or Mozilla web browser in the presence of the experimenter. The participants were given verbal and written instructions (see Appendix D for complete written instructions) and initially made judgments on a practice event set, which were not included in the analyses. The data from the first pilot was excluded from the analysis, since small modifications were made to the system after the user's participation.

The users were first shown each of the summaries for a particular event and asked to determine which of five choices best reflected the information presented in the summary. The choices were:

- Very Well - the summary provided quality information about the topic and the summary contained no or almost no unrelated information.
- Well - the summary provided quality information about the topic, but also contained unrelated information.
- Somewhat - the summary contained some information about the topic, but it also contained almost an equal amount of unrelated information; *OR* the summary contained some information about the topic, but not the information that was outlined in the topic description.
- Very Poorly - the summary contained very little information about the topic and contained mostly unrelated information.

- Not at All - the summary did not contain any information about the topic, and contained all unrelated information.

This judgment served as the "summary" judgment.

After the participant made judgments for all the summaries for an event, he/she was then shown the summary again and the complete list of source documents and asked to determine which of the five choices best reflects how well the summary focused on the required information (as specified by the event title and description) from the source documents. If the source documents did not contain any information related to the event description, the user was asked to determine how well the summary alerted them to the fact that no useful information was present in the source documents (this is described further in the experiment instructions in Appendix D). For Relevance Prediction, this judgment served as the *gold-standard*.[1]

Each of the twenty individual source documents included associated relevance levels as judged by the LDC annotators. For LDC Agreement, these judgments were recalibrated for the likert scale to create an averaged external gold-standard judgment, representing the cumulative relevance level of the set of the documents in terms of the specified event. The LDC-Agreement gold-standard judgments

[1]In previous experiments, the Relevance-Prediction judgment was the judgment the user made on the Full Text document. Since multiple documents are used in this case, the judgment the user made while viewing the summary in comparison to the 20 Full Text documents is used as the *gold standard*.

were created as described below.

For document sets that had:

- 0 related documents, a score of 1 was assigned
- 5 related documents, a score of 2 was assigned
- 10 related documents, a score of 3 was assigned
- 15 related documents, a score of 4 was assigned
- 20 related documents, a score of 5 was assigned

The user each made judgments for 10 event sets, which included a judgment for each summary system and each event (6 summary systems with 10 events, for a total of 60 summary judgments) and a judgment comparing each system summary to the full text source documents (6 summary/document judgments with 10 events, for a total of 60 summary/document judgments). At the end of the experiment, each user had made 120 total judgments (60 summary only judgments and 60 summary and document judgments) plus the judgments for the practice topic (which were not included in the analysis).

7.4 Judgment Scoring

The judgments of the previous experiments were binary, and allowed for ease in assigning scores for the metrics; correct answers were given a score of 1, and

incorrect answers were given a score of 0. The accuracy score was computed as the average of all the scores for a given system.

This new multi-document experiment design included a 5 point likert scale (described in the previous section) for judgments. This scale did not easily conform to the previous binary accuracy scoring method, so three scoring methods are proposed below:

Scale 1: Basic Scoring — The judgment on a summary is assigned a score based on its proximity to the judgment on the full text. If the judgments:

are equal, the assigned score is 1;

differ by one point, the assigned score is 0.75;

differ by two points, the assigned score is 0.5;

differ by three points, the assigned score is 0.25;

differ by four points, the assigned score is 0.

Scale 2: Bonus Scoring — The judgment on a summary is scored based on its perceived ability in providing the user with useful information from the source texts. A positive score is given if the judgment on the text is the same as, or HIGHER[2] than that of the summary (indicating that once the user viewed

[2]The scoring details are actually reversed to make the explanation of the comparison more intuitive. For the experiments, summaries with the most relevant information were assigned a lower score, and those with the least information were assigned a higher score (on a scale of 1 to 5). Therefore, all of the information about scoring is reversed—again, to make it easier to understand. The experiment instructions and exact scoring assignments can be found in

the full text documents, he/she determined that the summary provided them with useful information as requested by the event description and present in the texts). A negative score is given if the judgment on the text is LOWER[2] than that of the summary (indicating that the user found more useful information related to the event description in the texts than what was present in the summary). If the judgment on the text (compared to the judgment on the summary) is:

equal, the assigned score is 1;

1 point higher, the assigned score is 0.75;

2 points higher, the assigned score is 0.5;

3 points higher, the assigned score is 0.25;

4 points higher, the assigned score is 0;

1 point lower, the assigned score is -0.25;

2 points lower, the assigned score is -0.5;

3 points lower, the assigned score is -0.75;

4 points lower, the assigned score is -1.

Scale 3: Forced Binary Scoring — This scoring method is similar to the binary scoring method of previous experiments. In those experiments, users were only given "relevant" or "not relevant" as their judgment to describe the information provided by the summary in relation to the described event. The judgments are divided so that one set is mapped to an overall judgment of 1

Appendix D.

(relevant) and the opposite set is mapped to an overall judgment of 0 (not relevant). Because there were five possible judgments offered to the participants, the middle judgment can be mapped to either 1 or 0. Both options will be investigated here, and they are termed "Strict" scoring and "Non-Strict" scoring. Strict scoring represents judgments of 1 or 2 as relevant (score of 1), and judgments of 3, 4, or 5 as not-relevant (score of 0). Non-Strict scoring represents judgments of 1, 2, or 3 as relevant(score of 1), and judgments of 4 or 5 as not-relevant (score of 0).

7.5 Results and Analysis

The results for the Basic Scale (Scale 1) and Bonus Scale (Scale 2) with LDC Agreement and Relevance Prediction are shown in Table 7.1. Recall that the first hypothesis was that Relevance Prediction would produce accuracy scores that are at least 5% higher than LDC Agreement, and that the results would be significant at $p < 0.05$. Here, the Relevance-Prediction accuracy scores are at least 5% higher than LDC Agreement for all systems except the Baseline system. Using a paired t-test, the differences between LDC Agreement and Relevance Prediction for the Basic scale are significant with $p < 0.05$, but for the Bonus scale the differences are statistically significant with $p < 0.06$.

Significant differences are seen between systems for LDC Agreement (using a one-factor ANOVA with 47 degrees of freedom). Using Tukey's HSD test, differences are only seen between Human (the lowest performing system) and Baseline

System	Basic Scale		Bonus Scale	
	LDC	**RP**	**LDC**	**RP**
Baseline	**0.763**	0.756	**0.563**	0.531
Headline	0.703	**0.828**	0.528	**0.591**
Human	0.644	0.781	0.069	0.581
ISCC	0.713	0.772	0.238	0.447
SE	0.738	0.781	0.313	0.481
WTMC	0.741	0.806	0.366	0.556
HSD, $p<0.05$	0.110	—	0.245	—

Table 7.1: Accuracy Results of Extrinsic Task Measures using Two Scoring Scales

(the highest performing system) for the Basic Scale. For the Bonus Scale (and LDC Agreement), significant differences are only seen between the Human system and the three top performing systems, Baseline, Headline, and WTMC. The associated equivalence class results for Basic Scoring and LDC Agreement can be seen in Table 7.2 (with two overlapping sets, A and B). The equivalence classes for Bonus Scoring and LDC Agreement in Table 7.3 are more distinct—the systems are grouped into four sets, labeled A through D.

For the Basic and Bonus scale of LDC Agreement, the ranking of the human system in last place is very surprising. One would expect the Human system to be the or one of the highest performing systems since humans are known to easily identify and condense the important information in a text. A higher ranking of the Human system is seen with the Relevance-Prediction results. For the Bonus Scale, Relevance Prediction ranks Headline as first and the Human system as second. In the Relevance-Prediction Basic Scale results, the Human system ties for third place and is outranked only by the Headline and WTMC systems.

Baseline	A	
WTMC	A	B
SE	A	B
ISCC	A	B
Headline	A	B
Human		B

Table 7.2: Equivalence Classes for Basic Scale scoring with LDC Agreement

Baseline	A			
Headline	A	B		
WTMC	A	B	C	
SE		B	C	D
ISCC			C	D
Human				D

Table 7.3: Equivalence Classes for Bonus Scale scoring with LDC Agreement

System	**TP**	**FP**	**FN**	**TN**	**A**	**P**	**R**	**F**
Baseline	12	10	20	38	0.625	0.545	0.375	0.444
Headline	7	5	**25**	**43**	0.625	**0.583**	0.219	0.318
Human	**25**	**29**	7	19	0.550	0.463	**0.781**	0.581
ISCC	**25**	26	7	22	0.588	0.490	**0.781**	0.602
SE	24	20	8	28	**0.650**	0.545	0.750	**0.632**
WTMC	17	17	15	31	0.600	0.500	0.531	0.515
HSD, p<0.05	—	—	—	—	—	—	0.347	0.344

Table 7.4: LDC-Agreement Results of Extrinsic Task Measures (using Strict Scoring)

System	**TP**	**FP**	**FN**	**TN**	**A**	**P**	**R**	**F**
Baseline	13	9	**18**	40	0.663	0.591	0.419	0.491
Headline	8	4	7	**61**	**0.863**	0.667	0.533	0.593
Human	**46**	8	16	10	0.700	**0.852**	**0.742**	**0.793**
ISCC	31	**20**	11	18	0.613	0.608	0.738	0.667
SE	32	12	14	22	0.675	0.727	0.696	**0.711**
WTMC	22	12	11	35	0.713	0.647	0.667	0.657
HSD, p<0.05	—	—	—	—	0.212	—	—	—

Table 7.5: Relevance-Prediction Results of Extrinsic Task Measures (using Strict Scoring)

Since the other information retrieval metrics (precision, recall and f-score) do not fit with the two new scoring schemes, the results were reanalyzed using the Forced Binary Scoring method (described in Section 7.4).

The results using the LDC-Agreement method and Strict scoring are shown in Table 7.4. For accuracy, the Human system is ranked as the lowest system, consistent with the Basic and Bonus scale results of Table 7.1, however significant differences are not found between systems for this measure.

The Relevance-Prediction results for Strict scoring are shown in Table 7.5. Although the Human system is not ranked in first place as expected for the accuracy measure, significant differences with $p < 0.05$ are seen between systems. The systems are grouped into two overlapping sets, A and B, and the equivalence class results are displayed in Table 7.7.

For the LDC-Agreement recall and f-score measures (in Table 7.4), significant differences were found between systems (using a one-factor ANOVA with 47 degrees of freedom). However, based on the results of Tukey's HSD testing (shown in the last row of the table) only the recall measure for Strict scoring with LDC Agreement actually had distinguishable differences between systems. The associated equivalence classes with two overlapping groups, A and B, are given in Table 7.6.

The results with Non-Strict agreement are shown in Table 7.8 for the LDC-Agreement measure and in Table 7.9 for Relevance Prediction. Statistically significant differences between systems are seen for with three measures for each of the

Human	A	
ISCC	A	
SE	A	
WTMC	A	B
Baseline		B
Headline		B

Table 7.6: Equivalence Classes for LDC Agreement with the Recall measure (for Strict Scoring)

Headline	A	
WTMC	A	B
Human	A	B
SE	A	B
Baseline	A	B
ISCC		B

Table 7.7: Equivalence Classes for Relevance Prediction with the Recall measure (for Strict Scoring)

System	**TP**	**FP**	**FN**	**TN**	**A**	**P**	**R**	**F**
Baseline	39	8	9	24	**0.788**	**0.830**	0.813	**0.821**
Headline	20	7	**28**	**25**	0.563	0.741	0.417	0.533
Human	44	**23**	4	9	0.663	0.657	0.917	0.765
ISCC	45	20	3	12	0.713	0.692	0.938	0.796
SE	**47**	20	1	12	0.738	0.701	**0.979**	0.817
WTMC	39	13	9	19	0.725	0.750	0.813	0.780
HSD, $p<0.05$	—	—	—	—	0.167	—	0.264	0.219

Table 7.8: LDC-Agreement Results of Extrinsic Task Measures (using Non-Strict Scoring)

System	TP	FP	FN	TN	A	P	R	F
Baseline	43	4	15	18	0.763	0.915	0.741	0.819
Headline	21	**6**	**18**	**35**	0.700	0.778	0.538	0.636
Human	**64**	3	10	3	**0.838**	0.955	0.865	**0.908**
ISCC	61	4	9	6	**0.838**	0.938	**0.871**	0.904
SE	62	5	10	3	0.813	0.925	0.861	0.892
WTMC	50	2	13	15	0.813	**0.962**	0.794	0.870
HSD, $p<0.05$	—	—	—	—	—	0.289	0.276	0.248

Table 7.9: Relevance-Prediction Results of Extrinsic Task Measures (using Non-Strict Scoring)

evaluation methods—accuracy, recall and f-score for LDC Agreement, and precision, recall and f-score for Relevance Prediction. The equivalence class results with two overlapping sets, A and B, for LDC Agreement with Non-Strict scoring are shown in Table 7.14 for the accuracy measure, in Table 7.15 for the recall measure, and Table 7.16 for the f-score measure. Similarly, the equivalence classes for Relevance Prediction (again with overlapping sets A and B) are shown in Table 7.17 for the recall measure, and Table 7.18 for the f-score measure. Although the precision measure with Relevance Prediction and Non-Strict scoring was found to have significant differences using a one-factor ANOVA (with 47 degrees of freedom), the results of the Tukey test found no differences between systems for this measure.

To further explore the ordering of the systems, the results of the signal detection measures, sensitivity, specificity and d', are displayed in Tables 7.10 through 7.13. For LDC Agreement with Strict relevance (in Table 7.10), the d' results for Human and Headline rank the systems as third and fifth, respectively. We see that the Human system ranks highly with the sensitivity score, meaning that

System	Sensitivity	Specificity	discriminability index (d')
Baseline	0.375	0.792	0.494
Headline	0.219	**0.896**	0.482
Human	**0.781**	0.396	0.512
ISCC	**0.781**	0.458	0.672
SE	0.750	0.583	**0.885**
WTMC	0.531	0.646	0.453

Table 7.10: Results for Signal Detection Measures Using LDC Agreement, for Strict Relevance

System	Sensitivity	Specificity	discriminability index (d')
Baseline	0.419	0.816	0.698
Headline	0.533	**0.938**	**1.626**
Human	**0.742**	0.556	0.789
ISCC	0.738	0.474	0.571
SE	0.696	0.647	0.889
WTMC	0.667	0.745	1.089

Table 7.11: Results for Signal Detection Measures Using Relevance Prediction, for Strict Relevance

System	Sensitivity	Specificity	discriminability index (d')
Baseline	0.813	0.750	1.562
Headline	0.417	**0.781**	0.566
Human	0.917	0.281	0.804
ISCC	0.938	0.375	1.215
SE	**0.979**	0.375	**1.718**
WTMC	0.813	0.594	1.124

Table 7.12: Results for Signal Detection Measures Using LDC Agreement, for Non-Strict Relevance

System	Sensitivity	Specificity	discriminability index (d')
Baseline	0.741	0.818	1.556
Headline	0.538	0.854	1.149
Human	0.865	0.500	1.102
ISCC	**0.871**	0.600	1.387
SE	0.861	0.375	0.767
WTMC	0.794	**0.882**	**2.006**

Table 7.13: Results for Signal Detection Measures Using Relevance Prediction, for Non-Strict Relevance

this system helped users to judge "relevant" documents well, but ranks lowest for specificity which indicates that the system did not help users easily identify "not relevant" documents. The opposite is the case for the Headline system—Headline receives a high specificity scoreand a low sensitivity score.

The Relevance-Prediction results for Strict scoring are shown in Table 7.11. Here, the d' score for Headline ranked the system as the highest while the Human system ranked third. A result similar to that of LDC Agreement was seen for the sensitivity and specificity measures, Human ranked highest for sensitivity and lowest for specificity, while Headline ranked highest for specificity and lowest for sensitivity.

For Non-Strict Relevance, the LDC-Agreement results are shown in Table 7.12 and the Relevance-Prediction results are in Table 7.13. For LDC Agreement, Headline and Human rank lowest of all systems for the d' measure. Again, Headline ranks highest for specificity while Human ranks third highest for sensitivity (with less than a 2% difference from the system ranked second, and less

than a 6% difference from the system ranked first). For Relevance Prediction, Human ranks second highest for sensitivity, and Headline ranks second highest for specificity.

These results suggest that for both LDC Agreement and Relevance Prediction, the Human system is helpful to the user in identifying "relevant" documents, while the Headline system is helpful for identifying "not relevant" documents. This is consistent with the findings of the Relevance Prediction with Human Summaries experiment (reported in Section 5.4).

7.6 Discussion

The first hypothesis was that the Relevance-Prediction accuracy results would be at least 5% higher than that of LDC Agreement and that a statistically significant difference with $p < 0.05$ would be found among the systems with Relevance Prediction. As stated previously, the results of the Basic and Bonus scale partially support this hypothesis in that the Relevance-Prediction accuracy results are at least 5% higher than those of LDC Agreement (for all systems except Baseline), and the differences were statistically significant with $p < 0.05$ for the Basic scale and with $p < 0.06$ for the Bonus scale. The results for Relevance Prediction and Non-Strict scoring are similar—the Relevance-Prediction scores are at least 5% higher than LDC Agreement for all systems except for the Baseline system. Referring back to Strict scoring (Tables 7.4 and 7.5), the results for Relevance Prediction

Baseline	A	
SE	A	
WTMC	A	B
ISCC	A	B
Human	A	B
Headline		B

Table 7.14: Equivalence Classes for LDC Agreement with the Accuracy measure (for Non-Strict Scoring)

are 5% higher for four systems (Baseline, Headline, Human and WTMC), and 4% higher for two systems (ISCC and SE). Both the results (for Strict and Non-Strict scoring) results are significant with $p < 0.05$.

The Relevance-Prediction results for the Basic Scale (in Table 7.1), support the second hypothesis that the human-generated systems (Headline and Human) would score higher than the Baseline and automatic systems for the accuracy measure. However, the results for the Bonus Scale (also in Table 7.1), Strict Scoring (in Table 7.5), and Non-Strict scoring (in Table 7.9)) do not support this hypothesis. The LDC-Agreement results for Basic and Bonus Scoring (Table 7.1), Strict (Table 7.4) and Non-Strict scoring (Table 7.8) do not support the second hypothesis, in that all of the scoring methods have one of the human-generated systems being ranked as the lowest. For Basic and Non-Strict scoring, both the Human and Headline systems are ranked the lowest.

SE	A	
ISCC	A	
Human	A	
Baseline	A	
WTMC	A	
Headline		B

Table 7.15: Equivalence Classes for LDC Agreement with the Recall measure (for Non-Strict Scoring)

Baseline	A	
SE	A	
ISCC	A	
WTMC	A	
Human	A	
Headline		B

Table 7.16: Equivalence Classes for LDC Agreement with the F-Score measure (for Non-Strict Scoring)

ISCC	A	
Human	A	
SE	A	
WTMC	A	B
Baseline	A	B
Headline		B

Table 7.17: Equivalence Classes for Relevance Prediction with the Recall measure (for Non-Strict Scoring)

Human	A	
ISCC	A	
SE	A	
WTMC	A	B
Baseline	A	B
Headline		B

Table 7.18: Equivalence Classes for Relevance Prediction with the F-Score measure (for Non-Strict Scoring)

System	R1	R2	RL	RW
SE	**0.4450**	**0.1340**	**0.2266**	**0.0512**
WTMC	0.4040	0.1023	0.2061	0.0457
Baseline	0.3894	0.0953	0.2006	0.0449
ISCC	0.3855	0.1003	0.1911	0.0425
Human	0.3805	0.1057	0.1984	0.0441
Headline	0.1370	0.0174	0.0784	0.0182
HSD, $p<0.01$	0.0275	0.0243	0.0192	0.0018

Table 7.19: ROUGE Recall Results on the Seven Systems, sorted by ROUGE-1

7.7 Automatic Intrinsic Evaluation

ROUGE version 1.5.4 is used for the automatic intrinsic evaluation of the experimental summaries. For reasons described in Section 6.5, only the results of the recall-based and f-score-based ROUGE scoring methods will be discussed here. The ROUGE-Recall results are presented in Table 7.19 and Figure 7.1 and the ROUGE-F-Score results are presented in Table 7.20 and Figure 7.2.

In both tables, we see Headline ranking as the lowest performing system for all four measures (ROUGE 1-gram, 2-gram, ROUGE-L and ROUGE-W-1.2). Since ROUGE uses term matching to produce the results and the news story headlines often use words not present in the document, this result is not surprising and is consistent with the ROUGE results of the previous experiment (given in Section 6.5). For the measures of ROUGE-Recall, Human varies in ranking but with ROUGE-F-Score, Human is consistently rated as the highest or second highest system.

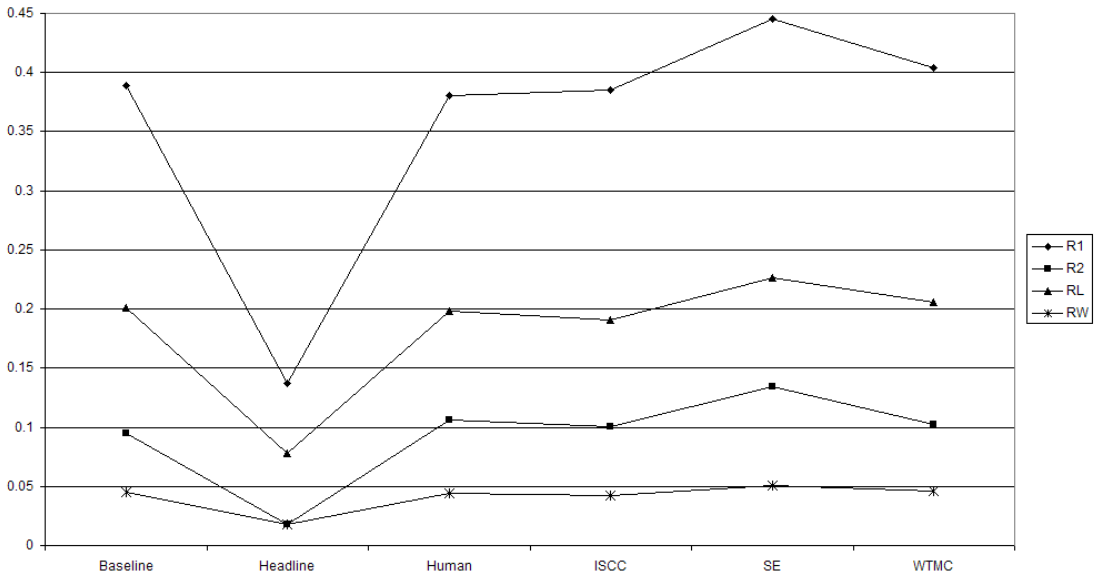

Figure 7.1: ROUGE Recall Results

7.8 Correlation of Intrinsic and Extrinsic Measures

Pearson r and Spearman ρ are the methods used for the correlations of the human extrinsic data and the automatic intrinsic data. For a detailed descrip-

System	**R1**	**R2**	**RL**	**RW**
SE	**0.3826**	**0.1156**	0.1947	**0.0697**
Human	0.3778	0.1044	**0.1975**	0.0646
WTMC	0.3473	0.0881	0.1771	0.0621
Baseline	0.3449	0.0843	0.1775	0.0619
ISCC	0.3323	0.0868	0.1648	0.0579
Headline	0.1817	0.0229	0.1048	0.0310
HSD, p<0.01	0.0239	0.0217	0.0055	0.0026

Table 7.20: ROUGE F-Score Results on the Seven Systems, sorted by ROUGE-1

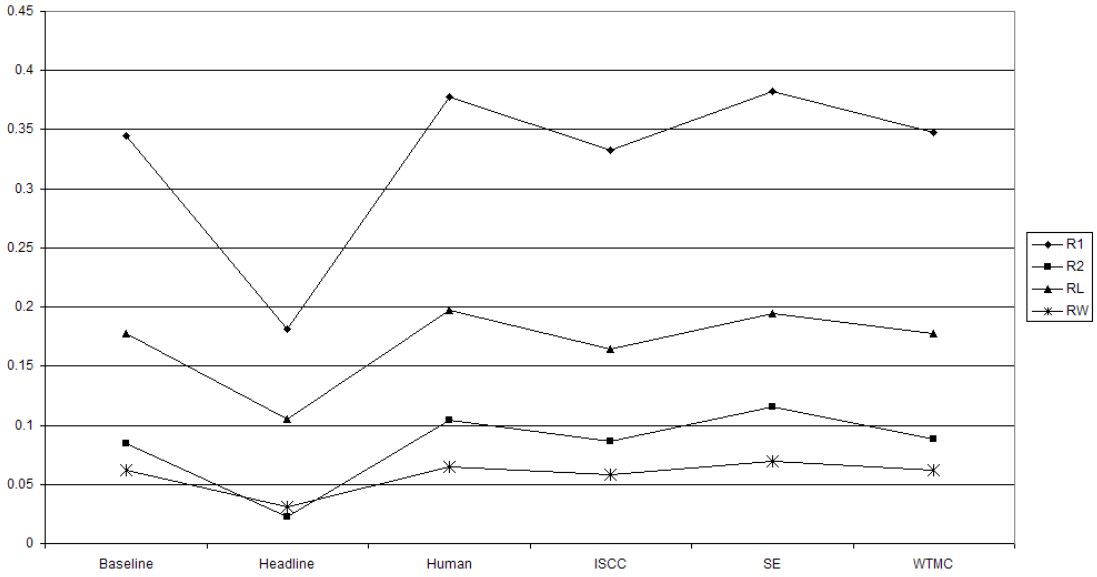

Figure 7.2: ROUGE Recall Results

tion of Pearson r refer back to Section 4.1.7, and for Spearman ρ refer back to Section 4.2.7.

The third hypothesis was that the Relevance-Prediction method would produce higher correlations than the LDC-Agreement method, and that the differences would be statistically significant with $p < 0.05$. Table 7.21 displays the results of the Pearson correlation between ROUGE and the results of the experiment using the Basic and Bonus scales with LDC Agreement and Relevance Prediction. The Spearman correlation results for the Basic and Bonus scoring scales are shown in Table 7.22. The resulting correlations in both table are low, and negative for most cases. This suggests results with ROUGE evaluations are not reflective of the results for either extrinsic measure with the Basic and Bonus scoring scales.

The correlations for the four extrinsic measures (accuracy, precision, recall

	Basic Scoring		Bonus Scoring	
	LDC	**RP**	**LDC**	**RP**
ROUGE-1 Recall	**0.101**	**-0.115**	-0.292	-0.123
ROUGE-2 Recall	0.035	-0.082	-0.374	**-0.130**
ROUGE-1 F-Score	0.066	-0.111	-0.345	-0.089
ROUGE-2 F-Score	0.004	-0.080	**-0.416**	-0.116

Table 7.21: Pearson r Correlation between Intrinsic and Extrinsic Scores for the Basic and Bonus Scoring Scales

	Basic Scoring		Bonus Scoring	
	LDC	**RP**	**LDC**	**RP**
ROUGE-1 Recall	**0.163**	-0.044	-0.142	**-0.139**
ROUGE-2 Recall	-0.019	-0.132	-0.373	-0.107
ROUGE-1 F-Score	0.067	-0.073	-0.283	-0.004
ROUGE-2 F-Score	-0.058	**-0.157**	**-0.439**	-0.093

Table 7.22: Spearman ρ Correlation between Intrinsic and Extrinsic Scores for the Basic and Bonus Scoring Scales

and f-score) with Strict Relevance and Non-Strict Relevance are presented in Tables 7.23 through 7.26. In most cases, the correlations for Strict Relevance with LDC Agreement are higher than those of Relevance Prediction. For Non-Strict Relevance, higher correlations are seen in most cases for Relevance Prediction than LDC Agreement. These results fail to support the third hypothesis.

	LDC Agreement				Relevance Prediction			
	Accuracy	**Precision**	**Recall**	**F-Score**	**Accuracy**	**Precision**	**Recall**	**F-Score**
ROUGE-1 Recall	-0.032	**0.310**	0.465	0.419	**-0.445**	0.163	0.265	0.328
ROUGE-2 Recall	-0.078	0.264	0.500	0.409	-0.406	0.189	0.296	0.373
ROUGE-1 F-Score	-0.043	0.306	0.507	**0.434**	-0.409	0.221	0.266	0.368
ROUGE-2 F-Score	**-0.090**	0.258	**0.527**	0.418	-0.386	**0.224**	**0.304**	**0.402**

Table 7.23: Pearson r Correlation between Intrinsic and Extrinsic Scores for Strict Relevance

	LDC Agreement				Relevance Prediction			
	Accuracy	**Precision**	**Recall**	**F-Score**	**Accuracy**	**Precision**	**Recall**	**F-Score**
ROUGE-1 Recall	0.388	0.077	0.710	0.591	0.338	0.489	0.605	**0.671**
ROUGE-2 Recall	0.311	-0.033	0.738	0.535	**0.360**	0.464	**0.607**	0.661
ROUGE-1 F-Score	**0.415**	**0.106**	0.723	**0.614**	0.337	**0.496**	0.589	0.670
ROUGE-2 F-Score	0.321	-0.028	**0.746**	0.544	0.359	0.469	0.600	0.662

Table 7.24: Pearson r Correlation between Intrinsic and Extrinsic Scores for Non-Strict Relevance

	LDC Agreement				Relevance Prediction			
	Accuracy	**Precision**	**Recall**	**F-Score**	**Accuracy**	**Precision**	**Recall**	**F-Score**
ROUGE-1 Recall	0.029	**0.241**	0.299	0.309	-0.361	0.032	0.176	0.129
ROUGE-2 Recall	**-0.112**	0.116	0.424	0.277	**-0.411**	0.191	0.241	0.304
ROUGE-1 F-Score	0.011	0.230	0.427	**0.372**	-0.297	0.188	0.223	0.290
ROUGE-2 F-Score	-0.101	0.120	**0.499**	0.321	-0.381	**0.238**	**0.261**	**0.368**

Table 7.25: Spearman ρ Correlation between Intrinsic and Extrinsic Scores for Strict Relevance

	LDC Agreement				Relevance Prediction			
	Accuracy	**Precision**	**Recall**	**F-Score**	**Accuracy**	**Precision**	**Recall**	**F-Score**
ROUGE-1 Recall	0.169	-0.042	-0.084	0.528	0.232	0.313	0.438	0.488
ROUGE-2 Recall	0.064	-0.068	**-0.260**	0.637	**0.253**	0.361	**0.475**	**0.558**
ROUGE-1 F-Score	**0.307**	**0.206**	0.027	0.597	0.124	**0.385**	0.370	0.459
ROUGE-2 F-Score	0.102	0.006	-0.236	**0.657**	0.209	0.353	0.463	0.539

Table 7.26: Spearman ρ Correlation between Intrinsic and Extrinsic Scores for Non-Strict Relevance

7.9 Experimental Findings

The Relevance-Prediction method produced accuracy scores that were at least 4% higher than the accuracy scores produced with LDC Agreement for all systems except the Baseline system. These results were shown to be statistically significant at $p < 0.05$ for the Basic scale, Non-Strict relevance and Strict relevance, and $p < 0.06$ for the Bonus scale. These results support the findings of the RP with Human Summaries experiment (described in Section 5.4) and the first hypothesis, that Relevance Prediction produces higher accuracy results than the LDC-Agreement method and that the differences are statistically signficant with $p < 0.05$.

The results did not support the second hypothesis that the Human and Headline systems would produce higher accuracy results than the other systems. However, the Human system consistently generates one of the highest scores for the sensitivity measure (with LDC Agreement and Relevance Prediction), indicating that the summaries are useful in helping a user correctly identify "relevant" documents. The Headline system generated the highest or second highest score for the specificity measure (with LDC Agreement and Relevance Prediction), indicating that the summaries are useful in helping a user correctly identify "not relevant" documents. The findings RP with Human Summaries experiment (in Section 5.4) are also consistent for the Human and Headline systems while the findings for the RP with Dual Summaries experiment (in Section 6.4) are consistent for the Human

system only.[3]

The moderate to high levels of correlations with the ROUGE evaluation and Relevance Prediction found in the previous two experiments in Sections 5.4 and 6.4 (which also served as the third experimental hypothesis) were not consistent with the findings of this experiment. Again, a major difference in this experiment is the use of documents spanning multiple documents and the choice of scoring methods. Here four distinct scoring methods were proposed and produced varying results. Also, the multiple relevance levels of the documents in each topic set had to be combined into a single number. Additional research with multi-document summaries may help to find alternate methods for representing the varied levels of relevance and may produce better results.

[3]The Headline system in the RP with Dual Summaries experiment was ranked third out of six systems for the specificity measure.

Chapter 8

Conclusions and Future Work

This thesis has introduced a new measure for human task-based summarization evaluation, Relevance Prediction, that is a more intuitive measure of human task performance than an external gold-standard based agreement measure, LDC Agreement. For single-document summarization, the results indicated that Relevance Prediction produced results that were significantly higher than those of the LDC-Agreement method or produced results with significant differences between systems when LDC Agreement did not.

The six experimental studies described in the thesis also investigated the claims of automatic intrinsic metrics of correlating highly with human task-based performance evaluation. BLEU, ROUGE, and Basic Elements were used throughout the studies as the intrinsic measures for summarization evaluation. Results of these correlation studies indicated that ROUGE has moderate, yet consistent correlations with the Relevance-Prediction method for single-document summaries. The sections below describe the specific findings, overall contributions of the thesis

System	LDC Event Tracking		RP with Human Summaries		RP Dual Summary	
	Accuracy	**Time**	**Accuracy**	**Time**	**Accuracy**	**Time**
Full Text	0.851	23 s	0.707	13.38 s	0.872	160 s
Headline	0.787	6.34 s	0.673	4.60 s	0.833	7.03 s
Human	0.815	7.38 s	0.677	4.57 s	0.783	6.87 s

Table 8.1: Accuracy and Timing Results for Three Experiments

and directions for future work.

8.1 Overall Findings

The results of the experiments in this thesis show that for single document summarization, Relevance Prediction produces higher scores for the information retrieval measures (accuracy, precision, recall, and f-score) than LDC Agreement in most cases. Specifically, for the accuracy (also known as agreement) measure, these differences are statistically significant with at least $p < 0.06$,[1] confirming that Relevance Prediction provides a more stable framework for evaluating different summarization techniques.

An additional finding is that Human summaries are particularly helpful in allowing a user to correctly identify "relevant" documents while Headline summaries are helpful in allowing a user to correctly identify "not relevant" documents.

Text summarization is shown to be useful for reducing judgment time in the extrinsic task-based evaluations while maintaning a level of accuracy similar to

[1]For most cases, the statistically significant differences were found with $p < 0.05$, but two cases had differences that were significant at the $p < 0.06$ level.

that of the full text judgments. Three of the experimental studies in this work allowed for the comparison of judgment time and accuracy scores for the full text document and summaries.[2] These results (displayed in Table 8.1) show that summaries reduce judgment time by at least 65% while having less than at 10% loss in accuracy.

The ROUGE metric correlates more highly with the results of extrinsic measures (Relevance Prediction and LDC Agreement) than the BLEU or Basic Elements methods. This suggests that out of the three automatic summarization evaluation metrics, the ROUGE metric provides results that are most reflective of of human task-based evaluations.

8.2 Contributions

The thesis yields the following contributions:

- Provision of a means for determining quality of current automatic summarization evaluation methods based on the level of correlation with human judgment measurements.

- Introduction of a new method, Relevance Prediction, that is a more intuitive

[2]Only the results for the LDC Event Tracking, the RP with Human Summaries, and the RP Dual Summaries experiments are shown. The LDC General experiment was shown to have an unsuitable task; the Memory and Priming study did not focus on relevance assessments; and the RP with Multi-Document Summaries experiment judgments were not timed.

measure of individual human task-based evaluation of text summarization than current "gold standard" methods for measuring human performance.

- Development of a methodology for conducting human evaluations to determine the usefulness of text summarization.

- Establishment of the usefulness of text summarization in reducing judgment time while maintaining a similar level of task judgment accuracy as seen with the full text documents.

- Creation and implementation of a new evaluation approach incorporating a 5-point likert scale for evaluation of multi-document summaries.

- Exploration of the factors that affect performance scoring including Single versus Multi-document summarization, and summary type (abstractive versus extractive).

- Use of the results of the human evaluations to compare summarization techniques.

8.3 Future Work

This section identifies possible directions for future work motivated by the findings of this thesis.

Decreased variation of relevance levels for evaluation with multi-document summaries: One difference with the evaluations of single- and multi-

document summaries is that the external relevance level of the documents in the single-document experiments were kept constant at 50% relevant, 50% not relevant. The multi-document evaluation used external relevance levels varying from 0% relevant (with 100% not relevant documents), to 100% relevant (without not relevant documents). An investigation of multi-document summaries with a more constrained set of external relevance levels could help in making direct comparisons with the results of the single-document evaluations.

Increased variation of relevance levels for evaluation with single-document summaries: As stated above, the single-document summaries in the experiments reflected a 50/50 relevant/not-relevant split across topics. Evaluations with single-document summaries using more varied sets of per-topic external relevance levels may enable comparisons with the results of the multi-document evaluations

Creation of a method for evaluation of multi-document summaries that eliminates the need for both strict and non-strict relevance: A finding of the first two single-document summary experiments was that using strict and non-strict scoring made it difficult to make definitive statements about the evaluation and correlation results. The creation of a scoring method for multi-document summaries that maps more directly to a binary system (possibly offering an even number of judgment options) may encourage more stable results and definitive findings between Relevance Prediction and LDC Agreement for multi-document summaries.

Use of Signal Detection Theory methods to analyze multi-document summary results: As stated above, the use of a method that can analyze the resulting data without the need for multiple formats (strict versus non-strict, basic versus bonus scoring) may lead to better results. Using additional Signal Detection Theory methods such as the Receiver Operating Characteristic (ROC) curve would eliminate these varying formats while continuing to test for differences between the systems and compare the LDC-Agreement and Relevance Prediction methods.

Extension of these methods to evaluation of question-answering (QA) tasks: The methods and investigations of the question-answering field are somewhat similar to that of summarization evaluation. The extension and investigation of the Relevance-Prediction method for human question-answering tasks as an alternative to comparisons with external gold-standards (answer keys created by external annotators) may produce interesting results.

In summary, this thesis introduced a new measure for human task-based summarization evaluation, Relevance Prediction, that is shown to be a more intuitive and more reliable measure of human task performance than an external gold-standard based agreement measure, LDC Agreement. This thesis also investigated correlations with current automatic intrinsic summarization evaluation metrics and extrinsic evaluation results produced with the Relevance-Prediction and LDC-Agreement methods. Moderate, yet consistent correlations were found with the ROUGE measure and Relevance Prediction.

Appendix A

Topics (Rules of Interpretation)

1. *Elections*: Examples - New people in office, new public officials, change in governments or parliaments (in other countries), voter scandals. The event might be the confirmation of a new person into office, the activity around voting in a particular place and time, the opposing parties' or peoples' campaigns, or the election results. The topic would be the entire process, nominations, campaigns, elections, voting, ceremonies of inauguration.

2. *Scandals/Hearings*: Examples - Monica Lewinsky, Kenneth Starr's investigations. The event could be the investigation, independent counsels assigned to a new case, the discovery of a potential scandal, the subpoena of political figures. The topic would include all pieces of the scandal or the hearing including the allegations or the crime, the hearings, the negotiations with lawyers, the trial (if there is one), and even media coverage.

3. *Legal/Criminal Cases*: Examples - crimes, arrests, cases. The event might

be the crime, the arrest, the sentencing, the arraignment, the search for a suspect. The topic is the whole package; crime, investigation, searches, victims, witnesses, trial, counsel, sentencing, punishment and other similarly related things.

4. *Natural Disasters*: Examples - tornado, snow and ice storms, floods, droughts, mud-slide, volcanic eruptions. The event would include causal activity (El Nino, in many cases this year) and direct consequences. The topic would also include; the declaration of a Federal Disaster Area, victims and losses, rebuilding, any predictions that were made, evacuation and relief efforts.

5. *Accidents*: Examples - plane- car- train crash, bridge collapse, accidental shootings, boats sinking. The event would be causal activities and unavoidable consequences like death tolls, injuries, loss of property. The topic includes mourners pursuit of legal action, investigations, issues with responsible parties (like drug and alcohol tests for drivers etc.)

6. *Ongoing violence or war*: Examples - terrorism in Algeria, crisis in Iraq, the Israeli/Palestinian conflict. In these cases the event might be a single act of violence, a series of attacks based on a single issue or a retaliatory act. The topic would expand to include all violence related to the same people, place, issue and time frame. These are the hardest to define, since war is often so complex and multi-layered. Consequences or causes often include (and would therefore be topic relevant) preparations for fighting, technology,

weapons, negotiations, casualties, politics, underlying issues.

7. *Science and Discovery News*: Examples - John Glenn being sent back into space, archaeological discoveries. The event is the discovery or the decision or the breakthrough. The topic, then, would include the technology developed to make this event happen, the researchers/scientists involved in the process, the impact on every day life, all history and research that was involved in the discovery.

8. *Finances*: Examples - Asian economy, major corporate mergers. The topic here could include information about job losses, impacts on businesses in other countries, IMF involvement and sometimes bail out, NYSE reactions (heavy trading BECAUSE Tokyo closed incredibly low). Again, anything that can be defined as a CAUSE of the event or a direct consequence of the event are topic-relevant.

9. *New Laws*: Examples - Proposed Amendments, new legislation passed. While the event may be the vote to pass a proposed amendment, or the proposal for new legislation, the topic includes the proposal, the lobbying or campaigning, the votes (either public voting or House or Senate voting etc.), consequences of the new legislation like protesting or court cases testing it's constitutionality.

10. *Sports News*: Examples - Olympics, Super Bowl, Figure Skating Championships, Tournaments. The event is probably a particular competition or

game, and the topic includes the training for the game or competition, announcements of (medal) winners or losers, injuries during the game or competition, stories about athletes or teams involved and their preparations and stories about victory celebrations.

11. *MISC. News*: Examples - Dr. Spock's Death, Madeleine Albright's trip to Canada, David Satcher's confirmation. These events are not easily categorized but might trigger many stories about the event. In these cases, keep in mind that we are defining topic as the seminal event and all directly related events and activities. (include here causes and consequences) If the event is the death of someone, the causes (illness) and the consequences (memorial services) will all be on topic. A diplomatic trip topic would include plans made for the trip, results of the trip (a GREAT relationship with Canada) would be on topic.

Appendix B

Experimental Questionnaire

Userid # _____

1. What's the highest degree/diploma you received or are pursuing?

   degree: _____

   major: _____

   year: _____

2. What is your occupation? _____

3. What is your gender? (Please circle one)

   male

   female

4. What is your age? _____

5. How often do you use the internet for document searching? (Please circle one)

every day

a few times per week

a few times per month

not very often

never

6. If you do use the internet for document searching what is your preferred method? (Please circle one)

   Google

   Ask Jeeves

   Yahoo

   Other - Please specify _____

7. How long have you been doing online searches? _____

8. Please circle the number closest to your experience:

How much experience have you had in:	none		some		lots
Using a point and click interface	1	2	3	4	5
Searching on computerized library catalogs	1	2	3	4	5
Searching on commercial on line systems (e.g. BRS Afterdark, Dialog, Lexis-Nexis)	1	2	3	4	5
Searching on world wide web search services (e.g. Alta Vista, Google, Excite, Yahoo, HotBot, WebCrawler)	1	2	3	4	5

9. Please circle the number closest to your searching behavior:

	never	once or twice a year	once or twice a month	once or twice a week	once or twice a day
How often do you conduct a search on any kind of system?	1	2	3	4	5

10. Please circle the number that indicates to what extend you agree with the following statement:

	strongly disagree	disagree	neutral	agree	strongly agree
I enjoy carrying out information searches	1	2	3	4	5

Appendix C

Instructions for Document Relevance Experiment

General Instructions

Your task is to review a topic description, and to mark subsequent displayed news stories (documents) as **relevant** or **not relevant** to that topic. The listing for each topic includes the title of an event and helpful, but possibly incomplete, information about that event. There will be a total of 20 documents displayed with each topic, and the document can be displayed as the entire news story text, or the news story headline. Some of the documents texts or headlines may contain information that is relevant to the topic, some may contain information that is not relevant. Mark a document **RELEVANT** if it discusses the topic in a substantial way (at least 10% of the document is devoted to that topic or the headline describes a document focusing on that topic). Mark a document **NOT RELEVANT** if less than 10% or none of the document is devoted to that topic or the headline describes

a document that does not focus on that topic. It is okay if you have some difficulty in deciding if a document is relevant or not. When deciding the relevance of a document, you are also asked to mark your confidence in that judgment. If you are sure that your relevant/not-relevant judgment is probably correct, please mark **high confidence**. If you are somewhat unsure, but believe it may be correct, please mark **medium confidence**. If you are totally unsure if your judgment for that document is correct, please mark **low confidence**. Finally, each topic will list a "Rule of Interpretation." Use the attached sheet to find specific details on how to determine whether documents are related to a particular topic.

General Definitions

TOPIC- A topic is an event or activity, along with all directly related events and activities. A set of 60 topics will be defined for the TDT3 corpus.

EVENT- An event is something that happens at some specific time and place, and the unavoidable consequences. Specific elections, accidents, crimes and natural disasters are examples of events.

ACTIVITY- An activity is a connected set of actions that have a common focus or purpose. Specific campaigns, investigations, and disaster relief efforts are examples of activities.

Appendix D

Instructions for the Multi-Doc Relevance Experiment

General Instructions

For this task, you will be given summaries and the purpose is to rate the quality of summaries in reference to a specific topic. You will be given a topic title and a topic description, and a list of summaries.

You can imagine that you are searching for information on the internet about the displayed topic. Your job is to rate the summaries you are given on how well they provide you with the information requested in the topic description.

Use the five choices below to rate the summaries when they are presented by themselves, or in reference to source documents that are related to the topic.

1. **Very Well** - the summary provided you with quality information about the topic and the summary contained no or almost no unrelated information.

2. **Well** - the summary provided you with quality information about the topic,

but also contained unrelated information.

3. **Somewhat** - the summary contained some information about the topic, but it also contained almost an equal amount of unrelated information; OR the summary contained some information about the topic, but not the information that was outlined in the topic description.

4. **Very Poorly** - the summary contained very little information about the topic and contained mostly unrelated information.

5. **Not at all** - the summary did not contain any information about the topic, and contained all unrelated information.

After you rate the summaries, you will be shown the summaries again, and the twenty text documents that the summaries were created from. All, some or none of the text documents may be related to the topic. If all or some of the text documents contain information that is related to the topic, your task then is to look at the information in the related documents only, and then rate how well the summaries pulled out the important information (as specified by the topic and description) in those documents. Please rate them again based on the five choices described above.

If none of the documents contain information related to the topic, your job then is to determine how good was the summary in letting you know that none of the documents were related, and rate them on the five choices described below.

Use the five choices below to rate the summaries in reference to all unrelated source documents:

1. **Very Well** - the summary provided you with quality information about the topics of the source texts, and by reading the summary only you would be sure that the source texts did not contain information related to the topic. (If you were doing an internet search on the topic, after viewing the summary, you would not want to open the source documents because you were sure they were not related to the topic).

2. **Well** - the summary provided you with quality information about the topics of the source texts, but you would be a little unsure if the summary missed information contained in the source texts that were related to the displayed topic. (If you were doing an internet search on the topic, after viewing the summary, you would consider opening the source documents).

3. **Somewhat** - the summary contained some information about the source texts but left you unsure about the topics the source texts covered. (If you were doing an internet search on the topic, after viewing the summary, you would know about some of the topics covered in the source text but would open them to determine the additional topics).

4. **Very Poorly** - the summary poorly reflected the source texts, and provided very little information about the topics covered. (If you were doing an in-

ternet search on the topic, after viewing the summary, you would open the source texts to determine the topics that were covered).

5. **Not at all** - the summary did not reflect the source texts at all. (You would need to read the source texts to find any information about the source topics).

BIBLIOGRAPHY

Khurshid Ahmad, Bogdan Vrusias, and Paulo C F de Oliveira. Summary Evaluation and Text Categorization. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, Canada, July 2003.

James Allan, Hubert Jin, Martin Rajman, Charles Wayne, Daniel Gildea, Victor Lavrenko, Rose Hoberman, and David Caputo. Topic-based Novelty Detection. Technical Report 1999 Summer Workshop at CLSP Final Report, Johns Hopkins, Maryland, 1999.

Satanjeev Banerjee and Alon Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the Association for Computational Linguistics (ACL) Workshop on Intrinsic and Extrinsic Measures for Machine Translation and/or Summarization*, Ann Arbor, MI, June 2005.

Ronald Brandow, Karl Mitze, and Lisa F. Rau. Automatic Condensation of Electronic Publications by Sentence Selection. *Information Processing and Management*, 31(5):675–685, 1995.

Jean Carletta. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22(2):249–254, June 1996.

Hoa Trang Dang. Overview of DUC 2005. In *Proceedings of the Document Understanding Conferences (DUC)*, Vancouver, Canada, Oct 2005.

Hoa Trang Dang. Overview of DUC 2006. In *Proceedings of the Document Understanding Conferences (DUC)*, Brooklyn, NY, June 2006.

René Descartes. Principles of philosophy. In John Cottingham, Robert Stoothoff, and Dugald Murdoch, editors, *The Philosophical Writings of Descartes*, volume 1, pages 193–291. Cambridge University Press, Cambridge, England, 1984. Original work published 1644.

Bonnie J. Dorr, David Zajic, and Richard Schwartz. Hedge Trimmer: A Parse-and-Trim Approach to Headline Generation. In *Proceedings of the Human Language Technology - North American Chapter of the Association for Computational Linguistics (HLT-NAACL) Text Summarization Workshop*, Alberta, Canada, May 2003.

Harold P. Edmundson. New Methods in Automatic Abstracting. *Journal of the Association for Computing Machinery*, 16(2):264–285, 1969.

Gunes Erkan and Dragomir Radev. The University of Michigan at DUC 2004. In *Proceedings of the Document Understanding Conferences (DUC)*, Boston, MA, May 2004.

Barbara Di Eugenio and Michael Glass. Squibs and Discussions - The Kappa Statistic: A Second Look. *Computational Linguistics*, pages 95–101, 2004.

Atefeh Farzindar, Frédérik Rozon, and Guy Lapalme. CATS a topic-oriented multi-document summarization system at DUC 2005. In *Proceedings of the Document Understanding Conferences (DUC)*, Vancouver, Canada, Oct 2005.

William Goldstein and Robin Hogarth, editors. *Research on Judgment and Decision Making : Currents, Connections, and Controversies*. Cambridge University Press, Cambridge, United Kingdom, 1997.

Cleotilde Gonzalez. Task Workload and Cognitive Abilities in Dynamic Decision Making. *Human Factors*, 47(1):92–101, 2005.

Therese Firmin Hand. A Proposal for Task-Based Evaluation of Text Summarization Systems. In *Proceedings of the ACL/EACL-97 Summarization Workshop*, Madrid Spain, July 1997.

Donna Harman and Daniel Marcu. *Proceedings of the Document Understanding Conference (DUC) 2001*. New Orleans, LA, 2001.

Donna Harman and Paul Over. *Proceedings of the Document Understanding Conference (DUC) 2004*. Boston, MA, 2004.

Perry R. Hinton. *Statistics Explained: A Guide for Social Science Students*. Routledge, New York, NY, 1995.

Eduard Hovy and Chin-Yew Lin. Automated Text Summarization in SUMMARIST. In *Proceedings of the Association for Computational Linguistics (ACL) Workshop on Intelligent Scalable Text Summarization*, Madrid, Spain, August 1997.

Eduard Hovy, Chin-Yew Lin, and Liang Zhou. Evaluating DUC 2005 Using Basic Elements. In *Proceedings of the Document Understanding Conferences (DUC)*, Vancouver, Canada, Oct 2005.

Eduard Hovy, Chin-Yew Lin, Liang Zhou, and Junichi Fukumoto. Automated Summarization Evaluation with Basic Elements. In *Proceedings of the Fifth Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy, April 2006.

Hongyan Jing, Regina Barzilay, Kathleen McKeown, and Michael Elhadad. Summarization evaluation methods: Experiments and analysis. In *Proceedings of the AAAI Symposium on Intelligent Summarization*, Stanford University, CA, March 23-25 1998.

Karen Spärck Jones. A Statistical Interpretation of Term Specificity and its Application to Retrieval. *Journal of Documentation*, 28:11–21, 1980.

Dharmendra Kanejiya, Arun Kumar, and Surendra Prasad. Automatic Evaluation of Students' Answers Using Syntactically Enhanced LSA. In *Proceedings of the HLT-NAACL 2003 Workshop on Building educational Applications Using Natural Language Processing*, pages 53–60, Morristown, NJ, MAY 2003.

Klaus Krippendorf, editor. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, 1980.

Thomas Landauer, Peter Foltz, and Darrell Laham. Introduction to Latent Semantic Analysis. *Discourse Processes*, 25:259–284, 1998.

Alon Lavie, Kenji Sagae, and Shyamsundar Jayaraman. The Significance of Recall in Automatic Metrics for MT Evaluation. In *Proceedings of the Association for Machine Translation in the Americas (AMTA) - 2004*, Washington, DC, September 2004.

LDC. *Data Annotation*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, 2006. http://www.ldc.upenn.edu.

Chin-Yew Lin. ROUGE: a Package for Automatic Evaluation of Summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, Barcelona, Spain, July 25–26 2004.

Chin-Yew Lin and Eduard Hovy. Manual and Automatic Evaluation of Summaries. In *Proceedings of the Association for Computational Linguistics (ACL) Workshop on Automatic Summarization*, Philadelphia, PA, July 2002.

Chin-Yew Lin and Eduard Hovy. Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics. In *Proceedings of the Joint Annual Meeting of Human Language Technology (HLT) and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 71–78, Edmonton, Canada, May-June 2003.

Chin-Yew Lin and Franz Joseph Och. ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland, August 23–27 2004.

Jimmy Lin and Dina Demner-Fushman. Automatically Evaluating Answers to Definition Questions. In *Proceedings of the 2005 Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, pages 931–938, Vancouver, Canada, October 2005.

Inderjeet Mani. Summarization Evaluation: An Overview. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL) Workshop on Automatic Summarization*, 2001.

Inderjeet Mani, Gary Klein, David House, and Lynette Hirschman. SUMMAC: a text summarization evaluation. *Natural Language Engineering*, 8(1):43–68, 2002.

Interjeet Mani and Eric Bloedorn. Summarizing Similarities and Differences Among Related Documents. *Information Retrieval*, 1(1):35–67, 1999.

Mark Maybury. Generating Summaries from Event Data. *Information Processing and Management*, 31(5):735–751, 1995.

I. Dan Melamed, Ryan Green, and Joseph P. Turian. Precision and Recall of Machine Translation. In *Proceedings of the Joint Annual Meeting of Human Language Technology (HLT) and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, Edmonton, Canada, May 2003.

Christof Monz. Minimal Span Weighting Retrieval for Question Answering. In *Proceedings of the Special Interest Group on Information Retrieval (SIGIR) Workshop on Information Retrieval for Question Answering*, Pittsburgh, PA, May 2004.

Christof Monz and Maarten de Rijke. The University of Amsterdam at CLEF 2001. In *Proceedings of the Cross Language Evaluation Forum Workshop (CLEF 2001)*, pages 165–169, Darmstadt, Germany, September 2001.

Diana Mutz and Ross Chanin. Comedy or News? Viewer Processing of Political News from Late Night Comedy Shows. In *Proceedings of the Political Communication Pre-Conference: Fun, Faith and Futuramas*, Chicago, Illinois, September 2004.

Ani Nenkova and Rebecca J. Passonneau. Evaluating Content Selection in Summarization: The Pyramid Method. In *Proceedings of the Joint Annual Meeting of Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, Boston, MA, May 2004.

Chris D. Paice. Constructing Literature Abstracts by Computer: Techniques and Prospects. *Information Processing and Management*, 26(1):171–186, 1990.

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 for Computational Linguistics (ACL 2004)*, Philadelphia, PA, July 2002.

Rebecca J. Passonneau and Ani Nenkova. Evaluating Content Selection in Human- or Machine-Generated Summaries: The Pyramid Method. Technical report, Columbia, New York, NY, 2003. CUCS-025-03.

Martin Porter. An Algorithm for Suffix Stripping. *Program*, 14(3):130–137, 1980.

Martin Porter. *Porter Stemming Algorithm*, 2006. http://www.tartarus.org/∼martin/PorterStemmer.

G. J. Rath, A. Resnick, and R. Savage. The Formation of Abstracts by the Selection of Sentences: Part 1: Sentence Selection by Man and Machines. *American Documentation*, 2(12):139–208, 1961.

Gerard Salton, Amit Singhal, Mandar Mitra, and Chris Buckley. Automatic Text Structuring and Summarization. *Information Processing and Management*, 33 (2):193–207, 1997.

Richard Schwartz, Sreenivasa Sista, and Timothy Leek. Unsupervised Topic Discovery. In *Proceedings of the Advanced Research and Development Activity in Information Technology (ARDA) Workshop on Language Modeling and Information Retrieval*, Pittsburgh, PA, May 2001.

Sidney Siegel and N. John Castellan, Jr. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York, second edition, 1988.

Vladimir M. Sloutsky and Anna V. Fisher. Induction and Categorization in Young Children: A Similarity-Based Model. *Journal of Experimental Psychology: General*, 133(2):166–188, 2004.

Anastasios Tombros and Mark Sanderson. Advantages of Query Biased Summaries in Information Retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2–10, 1998.

Cornelis Joost van Rijsbergen. *Information Retrieval*. Butterworths, London, England, 1979. 2nd Edition.

David Zajic, Bonnie J. Dorr, and Richard Schwartz. BBN/UMD at DUC2 2004: Topiary. In *Proceedings of the Document Understanding Conference (DUC)*, Boston, MA, May 2004a.

David Zajic, Bonnie J. Dorr, Richard Schwartz, and Stacy President. Headline Evaluation Experiment Results. Technical report, University of Maryland, College Park, MD, 2004b. UMIACS-TR-2004-18.

Liang Zhou and Eduard Hovy. Web-Trained Extraction Summarization System. In *Proceedings of the Joint Annual Meeting of Human Language Technology (HLT) and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, Alberta, Canada, May 2003.