LAMP-TR-133 MAY 2006
CS-TR-4808
UMIACS-TR-2006-28

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

Stacy F. President, Bonnie J. Dorr

Language and Media Processing Laboratory
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742-3275
stacypre@umiacs.umd.edu, bonnie@umiacs.umd.edu

Abstract

This research describes two types of summarization evaluation methods, intrinsic and extrinsic, and concentrates on determining the level of correlation between automatic intrinsic methods and human task-based extrinsic evaluation performance. Suggested experiments and preliminary findings related to exploring correlations and factors affecting correlation (method of summarization, quality of summary, type of intrinsic method used, and genre of source documents) are detailed. A new measurement technique for task-based evaluations, Relevance Prediction, is introduced and contrasted with the current gold-standard based measurements of the summarization evaluation community. Preliminary experimental findings suggest that the Relevance Prediction method yields better performance measurements with human summaries than that of the LDC-Agreement method and that small correlations are seen with one of the automatic intrinsic evaluation metrics and human task-based performance results.

The support of this research by the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-2-0001 is gratefully acknowledged.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
MAY 2006		**00-05-2006 to 00-05-2006**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Text Summarization Evaluation: Correlating Human Performance on an Extrinsic Task with Automatic Intrinsic Metrics	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Maryland,Language and Media Processing Laboratory,Institute for Advanced Computer Studies,College Park,MD,20742-3275**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**109**	

TABLE OF CONTENTS

List of Tables iv

List of Figures viii

1 Introduction 1
1.1 Motivation . . . 2
1.2 Proposed Experiments . . . 3
1.3 Contributions . . . 4
1.4 Preliminary Findings . . . 4
1.5 Outline . . . 5

2 Background 7
2.1 Text Summarization . . . 7
2.1.1 Types of Text Summarization . . . 7
2.1.2 Usefulness of Summarization . . . 10
2.2 Summarization Evaluation . . . 10
2.2.1 Human Intrinsic Measures . . . 11
2.2.2 Automatic Intrinsic Measures . . . 11
2.2.3 Human Extrinsic Evaluations . . . 24
2.2.4 Automatic Extrinsic Evaluations . . . 25
2.3 Summary . . . 25

3 Toward a New Agreement Measure: Relevance Prediction 26
3.1 LDC Agreement . . . 26
3.2 Relevance Prediction . . . 27
3.3 Agreement Measure Validation . . . 29

4 Initial Studies: Correlation of Intrinsic and Extrinsic Measures 33
4.1 LDC General: Correlation of BLEU and ROUGE and Extrinsic Task Performance . . . 33
4.1.1 Hypotheses . . . 33
4.1.2 Experiment Details . . . 34
4.1.3 Experiment Design . . . 35

4.1.4 Preliminary Results and Analysis . . . 36
4.1.5 Discussion . . . 42
4.1.6 Alternate Interpretation of Results . . . 44
4.1.7 Automatic Intrinsic Evaluation . . . 46
4.1.8 Correlation of Intrinsic and Extrinsic Measures . . . 48
4.1.9 Experimental Findings . . . 49
4.2 LDC Event Tracking: Correlation with an Extrinsic Event Tracking Relevance Assessment . . . 49
4.2.1 Hypotheses . . . 50
4.2.2 Experiment Details . . . 50
4.2.3 Experiment Design . . . 52
4.2.4 Preliminary Results and Analysis . . . 54
4.2.5 Discussion . . . 57
4.2.6 Alternate Interpretation of Results . . . 58
4.2.7 Automatic Intrinsic Evaluation . . . 59
4.2.8 Correlation of Intrinsic and Extrinsic Measures . . . 62
4.2.9 Experimental Findings . . . 69
4.3 Memory and Priming Study . . . 72
4.3.1 Experiment Details . . . 72
4.3.2 Experiment Design . . . 72
4.3.3 Preliminary Results and Analysis . . . 74
4.3.4 Discussion . . . 75
4.3.5 Findings . . . 76

5 Proposed Work: Relevance Prediction 77
5.1 Hypotheses . . . 77
5.2 Experiment Details . . . 78
5.3 Experimental Design . . . 79
5.4 Preliminary Results and Analysis . . . 80
5.5 Automatic Intrinsic Evaluation . . . 82
5.6 Correlation of Intrinsic and Extrinsic Measures . . . 83
5.7 Experimental Findings . . . 85

6 Additional Experiments 87
6.1 Experiments 5 & 6: The Relevance Prediction Method with Automatic Summaries . . . 87
6.2 Experiment 7: Multi-Document Summarization and Correlation with the Pyramid Method and Basic Elements . . . 88

7 Topics (Rules of Interpretation) 90

8 Experimental Questionnaire 92

9 Instructions for Document Relevance Experiment 94

LIST OF TABLES

1.1 Overview of Proposed Experiments 5

4.1 Experiment 1: Average Word and Character Counts for Each Surrogate 35

4.2 LDC General Latin Square Experiment Design 36

4.3 Contingency Table for Extrinsic Task 37

4.4 Results of Extrinsic Task Measures on Seven Systems with Strict Relevance, sorted by **A**ccuracy 39

4.5 Results of Extrinsic Task Measures on Seven Systems with Non-Strict Relevance, sorted by **A**ccuracy 39

4.6 Equivalence Classes of Automatic Summarization Systems with respect to Recall for Strict Relevance 40

4.7 Equivalence Classes of Automatic Summarization Systems with respect to F-Score for Strict Relevance 40

4.8 Equivalence Classes of Automatic Summarization Systems with respect to Recall for Non-Strict Relevance 41

4.9 Equivalence Classes of Automatic Summarization Systems with respect to F-Score for Non-Strict Relevance 41

4.10 Results Using Strict Relevance, sorted by LDC-Agreement (Accuracy) 42

4.11 Results Using Non-Strict Relevance, sorted by LDC-Agreement (Accuracy) 43

4.12 Non-Strict Relevance for Surrogate, Strict relevance for Full Text, Sorted by F-score 45

4.13 Non-Strict Relevance for Surrogate, Non-Strict relevance for Full Text, Sorted by F-score 45

4.14 BLEU and ROUGE Scores on Seven Systems, sorted by BLEU-1 . . 47

4.15 Pearson r Correlation between Intrinsic and Extrinsic Scores Grouped by System (including Full Text) for Strict Relevance 48

4.16 Pearson r Correlation between Intrinsic and Extrinsic Scores Grouped by System (excluding Full Text) for Strict Relevance 49

4.17 LDC Event Tracking Experiment: Average Word and Character Counts for Each Surrogate 52

4.18 Example Output From Each Experimental System 53

4.19 LDC Event Tracking Latin Square Experiment Design 54

4.20 Preliminary Results of Extrinsic Task Measures on Ten Systems, sorted by **A**ccuracy 55

4.21 Equivalence Classes of Automatic Summarization Systems with respect to Precision 56

4.22 User Agreement and Kappa Score 57

4.23 Composite Simulation User Results, sorted by Accuracy 58

4.24 ROUGE and BLEU Scores on Ten Systems, sorted by ROUGE-1 . 60

4.25 Honestly Significant Differences for Automatic Summarization Methods Using ROUGE and BLEU 61

4.26 Equivalence Classes of Automatic Summarization Systems with respect to ROUGE-1 62

4.27 Equivalence Classes of Automatic Summarization Systems with respect to BLEU-1 62

4.28 Pearson r Correlation between Intrinsic and Extrinsic Scores Grouped by System (including Full Text) 63

4.29 Pearson r Correlation between Intrinsic and Extrinsic Scores Grouped by System (excluding Full Text) 64

4.30 Spearman ρ Correlation between Intrinsic and Extrinsic Scores Grouped by System (excluding Full Text) 64

4.31 Pearson r Correlation between Intrinsic and Extrinsic Scores Grouped by System-Topic Pair-200 Data Points (including Full Text) 65

4.32 Pearson r Correlation between Intrinsic and Extrinsic Scores Grouped by System-Topic Pair (excluding Full Text) 67

4.33 Adjusted Pearson r Correlation between Intrinsic and Extrinsic Scores Grouped by System-Topic Pair (excluding Full Text) 69

4.34 Spearman ρ Correlation between Intrinsic and Extrinsic Scores Grouped by System-Topic Pair (excluding Full Text) 71

4.35 Spearman ρ Correlation between Intrinsic and Extrinsic Scores Grouped by System-Topic Pair (excluding Full Text) 71

4.36 Comparison of Summary/Document Judgments 74

4.37 Additional Comparison of Summary/Document Judgments 75

4.38 Average Timing for Judgments on Summaries and Full Text Documents (in seconds) 76

5.1 Results of Extrinsic Task Measures on Three Presentation Types, sorted by **A**ccuracy (using LDC Agreement) 80

5.2 Results of Extrinsic Task Measures on Three Presentation Types, sorted by **A**ccuracy (using Relevance Prediction) 80

5.3 Relevance Prediction Rates for Headline and Human Surrogates (Representative Partition of Size 4) 81

5.4 LDC-Agreement Rates for Headline and Human Surrogates (Representative Partition of Size 4) 82

5.5 Average Rouge-1 Scores for Headline and Human Surrogates (Representative Partition of Size 4) 82

5.6 Pearson Correlations with ROUGE-1 for Relevance Prediction (RP) and LDC-Agreement (LDC), where Partition size (P) = 1, 2, and 4 85

5.7 Users' Judgments and Corresponding Average ROUGE-1 Scores . . 85

LIST OF FIGURES

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. . . 16

2.2 Bitext Grid Example from Melamed et al. (2003) 18

2.3 Bitext Grid with Multiple Reference Text and Co-occuring Word Matches from Melamed et al. (2003) 19

2.4 Official definition of F-measure from Lin and Demner-Fushman (2005) 23

4.1 Non-Strict Relevance Average Results 37

4.2 BLEU Scores 46

4.3 ROUGE Scores 47

4.4 ROUGE Results for Ten Systems, (X axis ordered by ROUGE-1) . 60

4.5 BLEU Results for Ten Systems, (X axis ordered by BLEU-1) 61

4.6 Scatter plot of Pearson r Correlation between ROUGE-1 and Accuracy with 200 Data Points (including Full Text) 66

4.7 Scatter plot of Pearson r Correlation between ROUGE-1 and Accuracy with 180 Data Points (excluding Full Text) 68

4.8 Adjusted Pearson r Correlation between Intrinsic and Extrinsic Scores Grouped by System-Topic Pair (excluding Full Text) 70

5.1 Distribution of the Correlation Variation for Relevance Prediction on Headline and Human 84

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.

It becomes necessary 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, two automatic evaluation metrics and one semi-automatic method have been developed and are becoming more widely used in the text summarization evaluation community. These methods claim to correlate *highly* (Papineni et al., 2002) or *surprisingly well* (Lin and Hovy, 2003) with human evaluation, but this had not been validated in a broad-scale independent real-world task based evaluation.

To investigate these claims, this research explores several relevance assessment tasks for comparing automatic evaluation metrics against human judgment performance. In addition, this research aims to test the validity of a new measurement technique, Relevance Prediction (RP), for evaluating the effectiveness of a summarization system. In studies prior to the development of this new measure (Zajic et al., 2004b; Dorr et al., 2004), users[1] 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) (LDC, 2006). These gold standards were considered to be the "correct" judgments, yet the interannotator agreement

[1]The term "user" is used interchangeably throughout this text to mean a real-world user, i.e. information retrieval data analyst, or a participant in the particular experimental study being discussed.

was very low and inconsistencies were found in the user's judgments. Thus, it was difficult to make strong statistical statements from the outcome of these earlier experiments.

Relevance Prediction eliminates the need for an external "gold standard" by making use of the same user's relevance judgment on both the summary and the corresponding full text. Preliminary studies have been conducted using Relevance Prediction (Dorr et al., 2005). This measure has been shown to be more reliable and has produced better agreement scores than the LDC-agreement method.

1.1 Motivation

Text summarization evaluation is an area wrought with many challenges. Human evaluations 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. At least four total participants are usually needed to produce representative results, although using more participants is likely to increase the reliability of the results.

Another challenge with human evaluations is that human judgments vary greatly and pose an issue for evaluation measurements based on gold standard judgments. Two previous studies (Mani, 2001; Tombros and Sanderson, 1998) have reported very low agreement rates in studies that use gold standards.

Because of these challenges, automatic summarization evaluation methods have been proposed. These are known to be fast, inexpensive, easy to use, and reusable. Moreover, automatic methods allow developers to continuously check for improvements based on small changes to their summarizing system. However, previous studies have shown only minimal (if any) correlations between automatic summarization measures of human task performance (Zajic et al., 2004b; Dorr et al., 2004, 2005). Also, automatic summarization evaluations still rely on humans for the provision of reference summaries or for annotation of the full text.

This research investigates various methods for text summarization evaluation and attempts to correlate automatic and semi-automatic methods with human performance on relevance assessment tasks. To address some of the challenges above, numerous human evaluations are proposed with feedback improvements made from one experiment to the next. These evaluations utilize human and automatic summaries, and compare the human judgment results with measurements made by the current evaluation methods. To address the issue of low agreement rates, a new measurement technique is proposed to remove reliance on external gold standard judgments.

1.2 Proposed Experiments

The 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. Seven experimental studies are proposed and briefly outlined below. Detailed information about the first three experiments including the experimental hypotheses, design, and findings are presented in Chapter 4. The last four experiments are proposed work and are discussed in Chapters 5 and 6.

- 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 correlations.

- 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 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 LDC-Agreement is used as the basis of the analysis.

- Experiment 3: *Memory and Priming Study.* This study explores effects of ordering of documents and summaries on user performance. Results 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 a new Relevance Prediction (RP) method with scores produced by LDC-Agreement. The correlation of BLEU and ROUGE 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 to investigate cor-

relations of BLEU and ROUGE. Two human summary types, one automatic summarizer and the first-75 character baseline are tested.

- Experiment 6: *Constrained RP Dual Summary.* This study is similar to the RP Dual Summary experiment but uses fewer event and document sets to for a shorter experiment (approximately 1 - 1.5 hours long) to minimize fatigue and boredom among participants. The Relevance Prediction and LDC-Agreement methods are compared and correlations of BLEU and ROUGE are investigated. Two human summary types, one automatic summarizer and the first-75 character baseline are tested.

- Experiment 7: *Multi-Document Summarization.* This study investigates differences in correlations with summaries spanning more than one source document. Correlations of ROUGE, Basic Elements (BE) and the Pyramid method with human performance are investigated.

An overview and brief description of each of the experiments is shown in Table 1.1. The current status of each experiment (whether it is complete, preliminarily complete, or if it is suggested for future work) is also noted.

1.3 Contributions

The specific contributions of this work are:

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

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

- A method for measuring human performance that is more reliable than current "gold standard" methods.

- 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).

- Use of the results of the human evaluations to compare summarization techniques.

1.4 Preliminary Findings

Preliminary experimental findings using human summaries suggest that the Relevance Prediction method yields a better performance metric than that of the LDC-Agreement method and that the elimination of gold standards produces more

Experiment	Objective	Status		
		Done	Prelim	Future
1. LDC General	To determine whether BLEU and ROUGE correlate with human performance on an extrinsic task. Uses NIST topic and document sets.	X		
2. LDC Event Tracking	Investigates correlation of BLEU and ROUGE, but uses an event tracking relevance assessment task.	X		
3. Memory and Priming	Explores effects of ordering of documents and summaries on user performance. Compares results of ten different orderings in a two part experiment.	X		
4. RP with Human Summaries	Introduces the Relevance Prediction method and compares RP scores with LDC-Agreement scores. Investigates correlation of BLEU and ROUGE and tests only the human summary types.		X	
5. RP Dual Summary	Continues to test the Relevance Prediction method and tests both human and automatic summaries.		X	
6. Constrained RP Dual Summary	Similar to RP Dual Summary experiment, but uses fewer event and document sets to for a shorter experiment to minimize fatigue among participants.			X
7. Multi-Document Summaries	Investigates differences in correlations with summaries spanning more than one source document. Investigates correlations of ROUGE, BE and the Pyramid method with human performance.			X

Table 1.1: Overview of Proposed Experiments

stable results. The findings also show small, positive correlations with some automatic extrinsic evaluations metrics and human task-based Relevance Prediction measurements. These findings are based on human-generated summaries from a single source text. Therefore, continued experimentation is necessary to investigate the performance of the Relevance Prediction method with automatic summaries and to determine if using multi-document summaries affect scoring and correlations.

1.5 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 presents a previous relevance-assessment method and compares this to the new Relevance-Prediction method introduced in this paper. Chapter 4 compares and

contrasts the evaluation methods of previous sections and discusses the preliminary findings of their levels of correlation. Chapter 5 reports some preliminary findings of the proposed Relevance-Prediction method in an experimental study. Finally, the details of continued research and future experiments are provided in Chapter 6.

Chapter 2

Background

This chapter motivates work in evaluation by introducing detailed background information on text summarization. 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 qual-

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

Abstractive summaries 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 not provide an accurate reflection of the contents of the document, are good examples of abstractive summaries.

---

[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.

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

Extractive summaries use information directly from the source document(s). Automatic summarizers are more likely to produce extractive summaries. Many of them 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.

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 summarizing 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 is shown that the evaluation results of the same 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. Two hypotheses in this research are: (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 that 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 paper 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. Preliminary experiments have been conducted that 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. These experiments are discussed more in Chapter 3 and the associated findings in Chapter 4.

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 paper 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. Six examples of fully automatic intrinsic measures (Bleu, Rouge, BE, GTM, Meteor, and Pourpre) and one semi-automatic measure (Pyramid) are described in the sections below. Each of the measures takes a slightly different approach to summarization evaluation, some built upon the shortcomings of the previous methods. The measures also all claim to correlate highly with human extrinsic evaluations (described below in Section 2.2.3). Proposed experiments to test these correlation claims are described in Chapter 4.

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 authors 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-grams* co-occuring in a candidate (peer) summary and a reference (model) summary, and $Count(n\text{-}gram)$ is the number of *n-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.

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

- 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) 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

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

(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 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.

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.

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

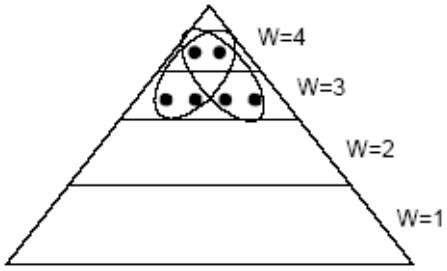

Figure 2.1: Example of a Pyramid with SCUs Identified and Marked for the Top Two Tiers from Nenkova and Passonneau (2004). W indicates the number of references associated with the SCUs at each level.

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 the Pyramid Method has for evaluation 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, the Pyramid Method does have drawbacks in that it is not automatic and 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).

2.2.2.4 Basic Elements

Although the approach of the Pyramid Method seems promising, the method is still primarily manual and relies heavily on human participation. In an effort to

explore the evaluation of summaries with smaller meaningful units of information (rather than previous reliance on sentence-level comparisons) 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). 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)").

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.

[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).

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 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

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

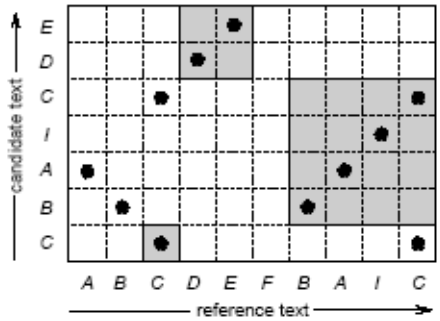

Figure 2.2: Bitext Grid Example from Melamed et al. (2003)

An important concept of the GTM method is the notion of co-ordination 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 co-ordination of some word in the reference text with some word in the candidate 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

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

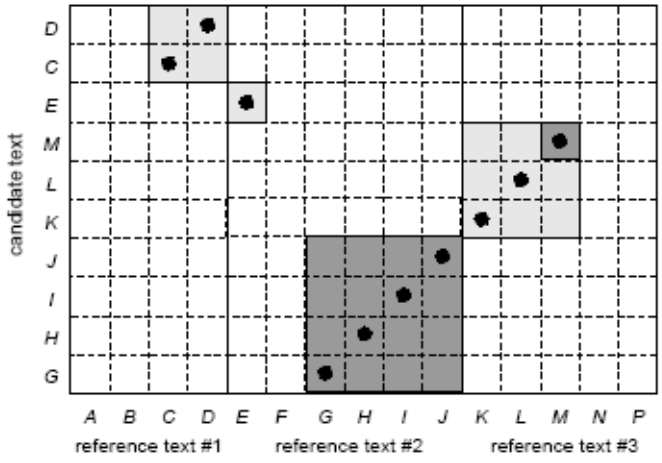

Figure 2.3: Bitext Grid with Multiple Reference Text and Co-occuring Word Matches from Melamed et al. (2003)

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.

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.

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:

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

[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).

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 Translation 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 are used, who are fluent speakers of the target language, 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.

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 metrics for question-answering 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.

[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.

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). The official F-measure is defined in Figure 2.4.

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.

Figure 2.4: 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] *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 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

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 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. To date, human judgments on these tasks have been 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 human judgments on a summary with his or her own judgment on the full text document instead of relying on external *gold standard* judgments.

The approach proposed here addresses the shortcomings of the SUMMAC studies (Mani et al., 2002). Specifically, this research introduces Relevance Prediction as an alternative to a gold standard. This measure is a more realistic agreement measure for assessing usefulness in a relevance assessment task. For

---

probability of occurring in relevant documents (Jones, 1980).

[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.

example, users performing browsing tasks examine document surrogates but open the full-text version only if they expect the document to be interesting to them. They are not trying to decide if the document will be interesting to someone else.

2.2.4 Automatic Extrinsic Evaluations

Currently, there are no automatic extrinsic evaluators, 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, 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 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 and The Pyramid Method and Basic Elements are suggested as intrinsic evaluation methods for the proposed studies in 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 assessed in Chapter 5 using both an existing gold-standard measure (LDC Agreement) and a new measure called Relevance Prediction. These measures are described in the next chapter.

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 information or that provided the user with too little information and encouraged an *incorrect* judgment.

One variant of gold-standard judgment, LDC-Agreement, uses LDC-commisioned judgments for relevance assessment (see Section 3.1). However, I maintain 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 the user's own judgments. In an experiment described in Chapter 5, LDC-Agreement and Relevance Prediction are compared for their correlation with human judgments.

We will now examine LDC-Agreement and Relevance Prediction in more detail and will then discuss the issue of Agreement Measure Validation.

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" to which other judgments should be compared against. 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). A new method that eliminates external gold-standard judgments and is thought to be more reliable than the LDC-Agreement method is proposed in the next section. In Section 3.3, factors that affect human judgment and illuminate problems associated with external gold-standard measurements are described.

3.2 Relevance Prediction

I propose a measure called *Relevance Prediction*, where each user builds their own "gold standard" based on the full-text documents. Agreement is measured by comparing users' surrogate-based judgments against their own judgments on the corresponding texts. If a user makes a judgment on a summary consistent with the judgment made on corresponding full text document, this signifies that the summary 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 judgement, and 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:

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

In the proposed experiments (as discussed in Chapter 4), the users make relevance judgments on a subset of all the summaries with a given system (for example human summaries) and then judgments for a subset of the documents. This ensures that the user does not make a judgment on an individual summary immediately before seeing the corresponding document. It usually takes users a few minutes to complete judgments on summaries and move on to the full text. In cases where more than one summary system is used, the users make judgments on a subset of documents within a given system prior to judging the corresponding subset for the next summary system until all summary systems are exhausted. Then, much later judgments are made for the full text documents. In all cases, the users have time delays of about fifteen to twenty minutes between judgments on summaries and judgments on the corresponding full text.[1]

It is believed that this approach is a more reliable comparison mechanism than LDC-Agreement because it does not rely on gold-standard judgments provided by other individuals. Specifically, 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 preliminary experimental results using this method are discussed in Chapter 5.

A concern that people may have with this method is that, once a user sees a summary, their judgment could be biased on the full-text document (or vice-versa). In the proposed experiments all source documents are similar (on the same topic or have enough related information to seem similar), and therefore, this may not be a problem. It would be hard for a user to immediately associate a given summary with a specific document, and in all cases of the regular experiments, the summaries are always shown before the full-text documents. To explore whether such a bias would be present, I conducted a Memory and Priming study (in Section 4.3), that confirmed that the order in which summaries and documents were presented did not affect the users judgments.

In light of the new measure, an interesting question to consider is why do external gold-standard based measures such as LDC-Agreement produce low per-

[1]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 Section 4.3.

formance scores? As is described in Chapter 5, the elimination of the external gold-standard (in the Relevance Prediction measure) produces higher performance scores than LDC-Agreement. This leads to another question: What factors influence differences in the decisions that humans make in the relevance judgment task? Some background on human judgments and factors that influence the variance of judgments are described next.

3.3 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 the specified topic or event is central to each individual's beliefs. Gold-standard based measurements try to impose a 'correct' answer and judge other individual's performance by those criteria. A key factor in the creation of the Relevance Prediction method is the accommodation of the variance of human judgments. As will be discussed below, individuals make relevance judgments on documents based on their background knowledge of the event or topic, their personal views on what information is salient in a text (which differs from user to user), and their cognitive biases.

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 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 are 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 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 manipulation 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%.[2] 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 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 informa-

[2]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).

tion 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 7 and 9).

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 (see Table 1.1 in Chapter 1) relevant to determining the level of correlation of the various intrinsic measures (i.e. ROUGE, Pyramid, BE, BLEU) to human performance on a document relevance assessment task. The findings from these first three studies have encouraged modifications to subsequent experiment designs, hypotheses and methods in the proposed work (to be described in Chapter 6).

The three experiments were referred to as LDC General, LDC Event Tracking and Memory and Priming in Chapter 1. These experiments aim to examine the use of the LDC-Agreement method to test the effectiveness of summaries for the task of relevance assessment by examining correlations between automatic metrics and human task performance. 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 keywords 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 (>80%) 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 expeirment, 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;

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

- **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 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

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

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. 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 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 Preliminary 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"

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

Table 4.3: Contingency Table for Extrinsic Task

LDC judgment.

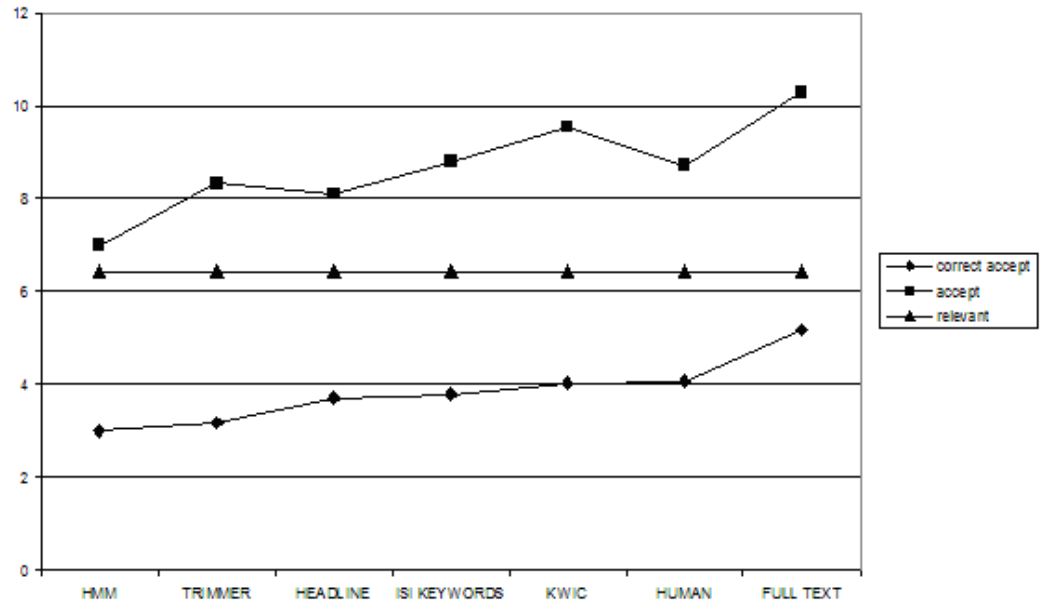

Figure 4.1: Non-Strict Relevance Average Results

Figure 4.1 shows the average counts out of 20 of accepted documents, correctly accepted documents, and actually relevant documents, using non-strict relevance. The number of actually relevant documents is constant across systems. The distance between the relevant line and the correct accept line indicates the average number of missed relevant documents in each group of 20. The distance between the correct accept line and the accept line indicates the average number of incorrectly accepted documents in each group of 20.

Although accuracy is the primary metric used to investigate the correlations between intrinsic and extrinsic measures, other metrics commonly used in the IR literature are imported (following the lead of the SUMMAC experimenters). The contingency table for the extrinsic task is shown in Table 4.3, 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 full set of extrinsic measures is given here:

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

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

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

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

Table 4.4 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 4.5. 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).

One-factor repeated-measures ANOVA 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 guarantees that one pair of systems is significantly different. 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 4.4 and 4.5 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

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 4.4: 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 4.5: Results of Extrinsic Task Measures on Seven Systems with Non-Strict Relevance, sorted by **A**ccuracy

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

Table 4.6: 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 4.7: Equivalence Classes of Automatic Summarization Systems with respect to F-Score for Strict Relevance

systems can only be claimed for the Recall measure with Non-Strict Relevance (Table 4.5), 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 overlapping sets, the members of which are not significantly distinct according to the Tukey test. This is shown in Tables 4.6 and 4.7.

For Non-Strict Relevance (as seen in Table 4.5), the HSD at $p<0.05$ is 0.142 for recall and 0.143 for the f-score. The equivalence classes for Recall and F-score are shown in Tables 4.8 and 4.9.

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

Table 4.8: 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 4.9: Equivalence Classes of Automatic Summarization Systems with respect to F-Score for Non-Strict Relevance

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 4.10: Results Using Strict Relevance, sorted by LDC-Agreement (Accuracy)

Additional measures were also calculated:

- 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.

4.1.5 Discussion

The experiments 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. These low scores 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 4.4, 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

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

an upper bound for all performance measures. The Full Text did perform the best of all the systems with Non-Strict Relevance (Table 4.5 supporting the main hypothesis). Also, for both Strict and Non-Strict Relevance (Tables 4.4 and 4.5), 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 for Strict Relevance. The HMM system tied with the Human system for the highest accuracy result, and achieved a higher result than both the headline and human systems for precision. The results with Non-Strict Relevance did rank the Headline and Human systems as the highest for the accuracy precision, and f-score measures, supporting the second hypothesis.

The KWIC system had the lowest performance results for Accuracy and Precision with Strict relevance (Table 4.4), but ranked higher than one automatic system for Non-Strict relevance (Table 4.5). It generated mid-range scores for recall and the f-measure, which does not support the 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 4.10 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 4.4 and 4.5 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 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 Alternate Interpretation of Results

When using a search engine, users often make a tentative decision that a document might be relevant by looking at a summary and then they finalize their decision by looking at the full document. This behavior can be simulated by constructing composite users from the actual experiment data. The implementation of a composite-user experiment using the results of these experiments will now be described.

Imagine that each relevance judgment represents one part of a two-stage process. First a Topic$_i$-System$_j$ user makes a judgment about a document. If the first-stage judgment is positive, the judgment of the Topic$_i$-FullText user is considered as the second stage. If the second-stage judgment is also positive, the composite user is thought to have made a positive judgment. If either the first-stage or second-stage judgment is negative, the composite user is thought to have made a negative judgment. The time for one session of 20 documents is computed as the time taken by the first-stage user for that session, plus $\frac{n}{20}$ of the time taken by the second-stage user, where n is number of documents accepted by the first-stage user. The results of this composite-user experiment are shown in Tables 4.12 and 4.13 for Strict Document Relevance and Non-Strict Document Relevance, respectively. For comparison, the results using the document alone are reproduced in the first row of each table.

From these results it is concluded that using a surrogate improves the processing speed over reading the entire document. If the performance of the surrogate systems is compared to the performance of the composite system, an improvement in precision and degradation in recall are observed. This corresponds to the intuition that reading the entire document allows the composite participant to reject documents incorrectly accepted by viewing the surrogate, and that some documents are incorrectly rejected after viewing the surrogate.

System	Precision	Recall	F-Score	Time (S)
Full Text	0.54	0.52	0.53	33.15
Human	0.70	0.35	0.47	4.33
Headline	0.64	0.34	0.45	3.94
ISIKWD	0.63	0.33	0.43	4.09
KWIC	0.60	0.32	0.41	3.57
HMM	0.58	0.29	0.38	4.01
Trimmer	0.56	0.29	0.38	3.87

Table 4.12: Non-Strict Relevance for Surrogate, Strict relevance for Full Text, Sorted by F-score

System	Precision	Recall	F-Score	Time (S)
Full Text	0.50	0.81	0.62	34.33
Human	0.64	0.52	0.57	3.43
Headline	0.60	0.48	0.53	3.31
KWIC	0.58	0.49	0.53	2.93
ISIKWD	0.57	0.49	0.53	3.25
Trimmer	0.54	0.42	0.47	3.26
HMM	0.54	0.38	0.45	3.47

Table 4.13: Non-Strict Relevance for Surrogate, Non-Strict relevance for Full Text, Sorted by F-score

4.1.7 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. Three additional short human summaries were commissioned for use as references in the automatic testing. BLEU was used used with 1-grams through 4-grams and the results are shown in Figure 4.2.

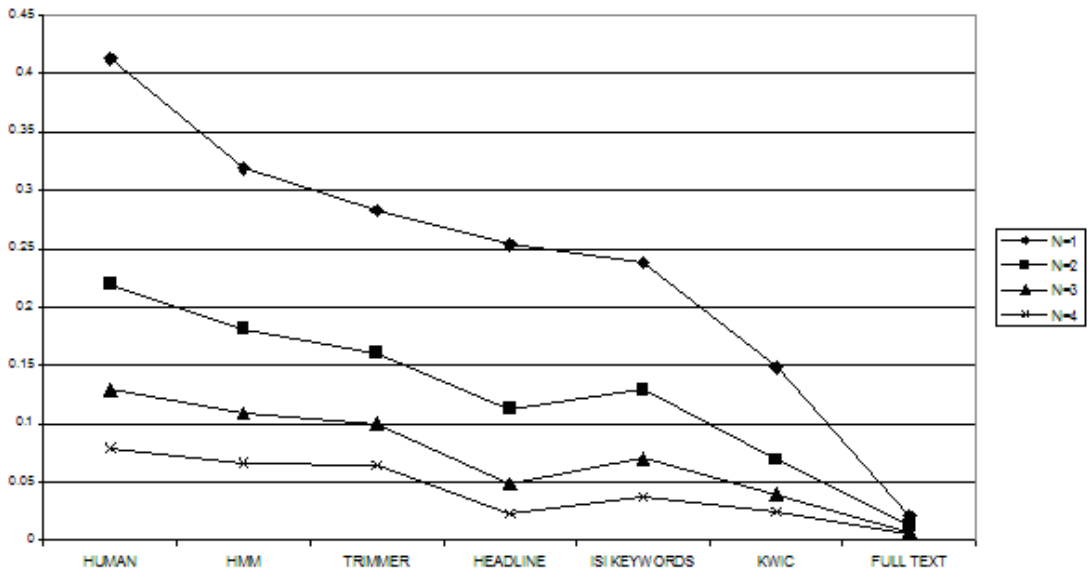

Figure 4.2: BLEU Scores

Unsurprisingly, the human summary 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 4.3, 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 4.14. 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 human-generated outputs (Full Text, Human and Headline).

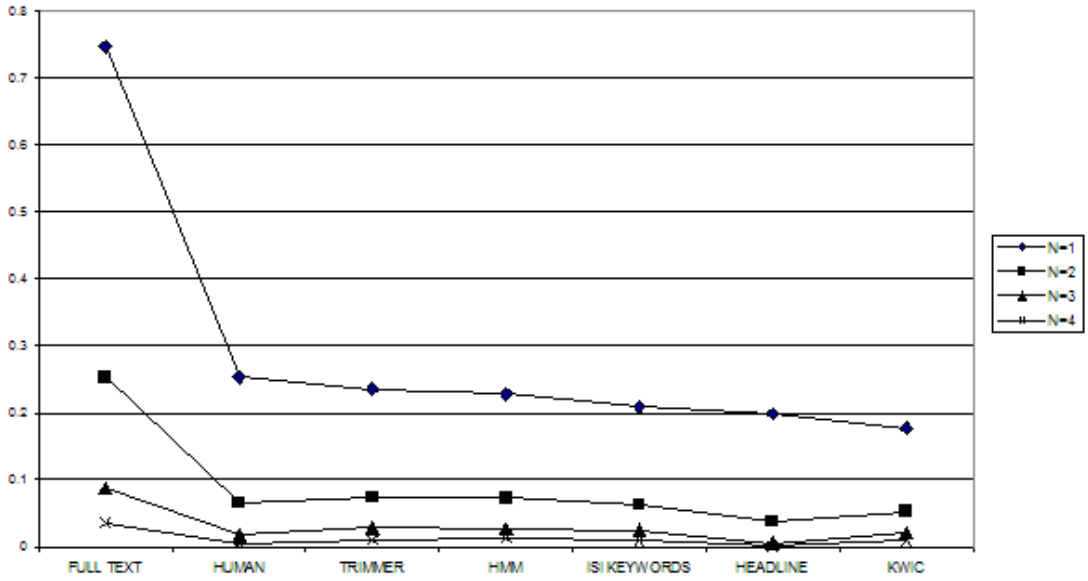

Figure 4.3: ROUGE Scores

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

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

4.1.8 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 4.15 and Table 4.16 show the rank correlations—using Pearson r (Siegel and Castellan, 1988)—between the average system scores assigned by the task-based metrics from Table 4.4 and the automatic metrics from Table 4.14. 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 4.15 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, this correlation decreases dramatically, and the BLEU-1 measure 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 (>80%) correlation with human performance. It must be noted that without the inclusion of the Full Text system, this hypothesis is not supported.

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

4.1.9 Experimental Findings

A concern with this experiment was the low individual performance, low interannotator agreement 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 cases, some systems were 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, users should be constrained to making only a "Relevant" or "Not Relevant" judgment. The elimination of "Somewhat Relevant" will help to 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 preliminary 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 0.6 Kappa score as opposed to the .38 Kappa score achieved in the SUMMAC experiments. The reader is 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 4.15, 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, 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.

And finally, as before, the "Full Text" document was included as a system and used as the 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 4.17. In this experiment, the outputs of each of the experimental systems 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

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

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 4.18. The *Rules of Interpretation* (Appendix A) is 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—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-measure 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, 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.

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 4.18: Example Output From Each Experimental System

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 4.19: LDC Event Tracking Latin Square Experiment Design

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 4.19) 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, 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 Preliminary 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

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 4.20: Preliminary Results of Extrinsic Task Measures on Ten Systems, sorted by **A**ccuracy

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 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 4.3, the extrinsic measures used for this experiment are: Accuracy, Precision, Recall, and F-score.

The Tukey Honestly Significant Difference (HSD) is also computed to determine whether differences found between groups of systems are statistically significant.

Table 4.20 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 was computed 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 4.20 with

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

Table 4.21: Equivalence Classes of Automatic Summarization Systems with respect to Precision

$p<0.05$. 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 systems with the highest accuracy were First75 and ISIKWD (0.748) and the lowest was UTD (0.675). The difference between them is 0.073, which is less than the HSD for accuracy (0.099), so a significant difference cannot be claimed between UTD and ISIKWD. On the other hand, the difference between UTD and HUMAN accuracy (0.815) is 0.140, greater than the HSD, so a significant difference between UTD and HUMAN can be claimed on accuracy.

Unfortunately, significant differences with $p<0.05$ cannot be claimed among any of automatic systems for precision, recall, f-score or accuracy using the Tukey Test. The issue is that Tukey is a more conservative test than ANOVA, which means there might have been real differences it did not detect. The automatic summarization systems were also 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 the Tukey (HSD) Test.

Using the mean scores from Table 4.20, the results of 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, the members of which are not significantly distinct according to the Tukey test. This is shown in Table 4.21.

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 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,

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

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.

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:

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

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

System	TP	FP	TN	FN	P	R	F	A	T (s)
Human	532	46	260	762	0.920	0.672	0.777	0.809	17.0
Headline	490	37	302	771	0.930	0.619	0.743	0.788	16.1
First75	448	32	344	776	0.933	0.566	0.704	0.765	14.8
Topiary	475	59	317	749	0.890	0.600	0.716	0.765	17.0
KWIC	498	86	294	722	0.853	0.629	0.724	0.762	18.9
ISIKWD	438	38	354	770	0.920	0.553	0.691	0.755	16.5
GOSP	435	37	357	771	0.922	0.549	0.688	0.754	15.1
UTD	453	54	339	754	0.893	0.572	0.697	0.754	16.8
Trimmer	415	44	377	764	0.904	0.524	0.663	0.737	15.5
HSD, $p < 0.05$					0.091	0.120	0.107	0.059	

Table 4.23: Composite Simulation User Results, sorted by Accuracy

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 Alternate Interpretation of Results

Again, a composite experiment is implemented using these results. Two users made judgments about 20 documents for each combination of topic and summarization system. In particular two users viewed the full text for each topic. Let Topic_i denote a particular topic and System_j denote a particular system other than FullText. There are two users who made judgments about Topic_i-System_j and two different users who made judgments about Topic_i-FullText. There are four possible combinations of Topic_i-System_j users and Topic_i-FullText users. Each combination is considered to be a composite user.

Recall that in Section 4.1.6, time for one session of 20 documents in the composite experiment was computed as the time taken by the first-stage user for that session, plus $\frac{n}{20}$ of the time taken by the second-stage user, where n is number of documents accepted by the first-stage user. Time is computed similarly in this composite experiment, and the results are shown in Table 4.23.

A one-factor independent-measures ANOVA is performed on the results of the simulation. Independent-measures is used because the composite users did not occur multiple times and thus were not a source of variance. Significant differences were found across systems at $p<0.05$ for Accuracy (A), Recall (R) and F-score (F), and at $p<0.01$ for Precision (P). A significant difference across systems was not found for Time (T).

This simulation has some predictable differences from the activity it models. If the first-stage users had been told that their job was not to decide whether a document was relevant, but only to eliminate the obviously irrelevant documents, they would probably have performed the task more quickly. Also, because the

numbers of relevant and non-relevant documents were equal in this experiment, it is expected that the first-stage users will pass about half of the documents on to the second stage. Therefore, the time is expected to be approximately the time of the summary systems plus half the time of the full text. Using values from Table 4.20:

$$7 + \frac{23}{2} = 18.5$$

The times calculated in the composite-user simulation were generally slightly lower than 18.5 seconds per document, however still within the same range, due to the high precision and low recall in this simulated experiment. In particular, the first-stage users frequently failed to pass relevant documents to the second-stage, so the number of documents judged in the second stage was low, and thus the time taken to make the judgments was low. However, even if the users had always made correct judgments, the expected time of 18.5 seconds per document is 80% of the observed time using full documents, which is not enough of a time improvement to justify the summarization.

In order to meaningfully test this approach, it will be necessary to create a scenario in which there is a high ratio of non-relevant to relevant documents and instruct first-stage users to favor recall over precision. (The practicality of such a scenario is currently a subject of debate: Given that 50% to 80% of the highest scoring documents returned by a typical IR engine are relevant, it is not clear that creating a result set with a low density of relevant documents is a realistic scenario.)

4.2.7 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 4.24.

Analogously to the extrinsic evaluation measures discussed above, the ANOVA values were computed to see whether there are differences between the systems for 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 4.4 and 4.5, respectively. In both graphic representations, the 95% confidence interval is shown by the error bars on each line.

In Figure 4.4, 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 4.5, the full document representation performs

System	R1	R2	R3	R4	RL	RW	B1	B2	B3	B4
Full Text	0.81808	0.35100	0.16782	0.10014	0.70117	0.38659	0.0301	0.0202	0.0139	0.0101
First75	0.25998	0.09824	0.05134	0.03119	0.22888	0.13837	0.3893	0.2564	0.1859	0.1420
ISIKWD	0.24188	0.00866	0.00027	0.00000	0.16230	0.09463	0.4043	0.0743	0.0166	0.0000
Topiary	0.22476	0.06992	0.02962	0.01369	0.19310	0.11582	0.3604	0.2067	0.1334	0.0903
KWIC	0.20265	0.06093	0.02813	0.01689	0.17310	0.10478	0.3306	0.1912	0.1289	0.0949
Headline	0.20084	0.04744	0.01282	0.00297	0.17669	0.10404	0.3491	0.1857	0.1020	0.0571
GOSP	0.20035	0.06285	0.02114	0.00844	0.18101	0.10798	0.3074	0.1858	0.1115	0.0686
Trimmer	0.18901	0.07095	0.03351	0.01633	0.17453	0.10548	0.3414	0.2282	0.1597	0.1148
Human	0.16838	0.03872	0.01180	0.00457	0.14508	0.08565	0.4326	0.2537	0.1536	0.0955
UTD	0.12802	0.01444	0.00128	0.00000	0.10684	0.06541	0.1913	0.0228	0.0000	0.0000
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 4.24: ROUGE and BLEU Scores on Ten Systems, sorted by ROUGE-1

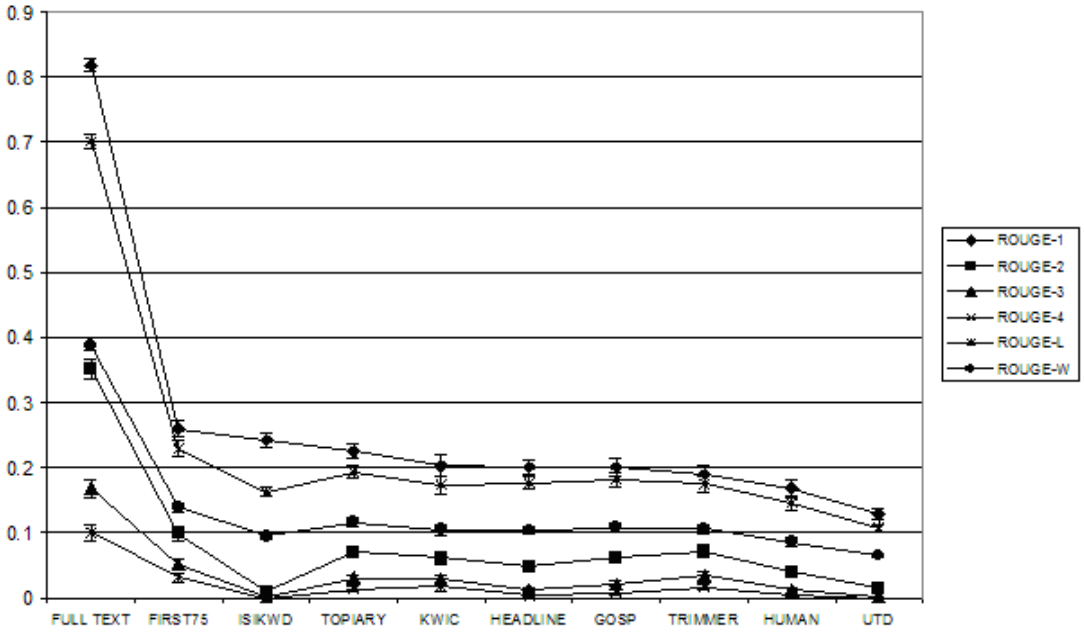

Figure 4.4: ROUGE Results for Ten Systems, (X axis ordered by ROUGE-1)

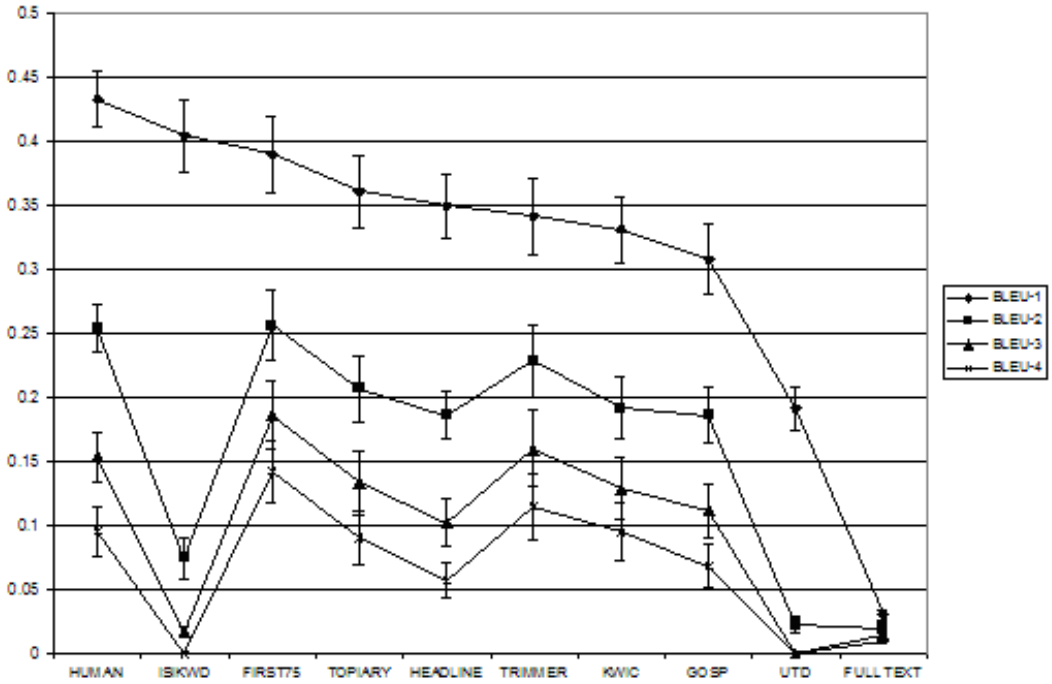

Figure 4.5: BLEU Results for Ten Systems, (X axis ordered by BLEU-1)

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

rather poorly. This is 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, which allows the honestly significant differences (HSD) to be computed for each measure, which is shown in Table 4.25.

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 4.26 and 4.27.

Whereas evaluation with ROUGE-1 allows for a rather differentiated grouping of the summarization methods, evaluating with BLEU-1 only resulted in two

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

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

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

Table 4.27: Equivalence Classes of Automatic Summarization Systems with respect to BLEU-1

groups.

4.2.8 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 showed that there is a statistically significant correlation between different intrinsic evaluation measures and common measures used in 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 4.28 through Table 4.30 below show the rank correlations between the average system scores assigned by the task-based metrics from Table 4.20 and the automatic metrics from Table 4.24. Two methods were used for computing this correlation: Pearson r as used for comparison with the previous experiment,

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

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 computed, 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 4.28 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 (*). Note that the strongest correlation is between ROUGE-1 and Accuracy. Thus, the ROUGE-1/Accuracy correlation will be the primary focus for the remainder of this section.

Looking back at Figures 4.4 and 4.5, 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. (I will return to this point again shortly.) Table 4.29 shows the results for Pearson r over all systems, 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 4.30 below.

Tables 4.29 and 4.30 show that there is a positive correlation in some cases, but it also shows that all positive correlations are rather low. Tests of statistical

	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 4.29: 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	0.3	0.216	-0.25	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	-0.2

Table 4.30: Spearman ρ Correlation between Intrinsic and Extrinsic Scores Grouped by System (excluding Full Text)

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

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 4.31.

Having a sufficiently large number of data points allows the creation of a scatter plot showing the correlation between intrinsic and extrinsic measures. Figure 4.6 shows the scatter plot corresponding to the ROUGE-1/Accuracy Pearson correlation given in Table 4.31.

In the case of a strong positive correlation, one would expect the data points to gather along the straight line that characterizes the least sum of squared differences between all points in the plot. However, Figure 4.6 shows that this is not the case. Rather, the plot shows two separate formations, where the data points in the upper right corner are the data points using full text. Clearly these are outliers. Including these data points results in an artificially high correlation that is largely dominated by the fact that both Rouge-1 and Accuracy can distinguish between summaries and full text, which is not the main interest here.

Because the primary interest is in the performance with respect to summaries only, the 20 data points that use full text will be removed from the data set and the

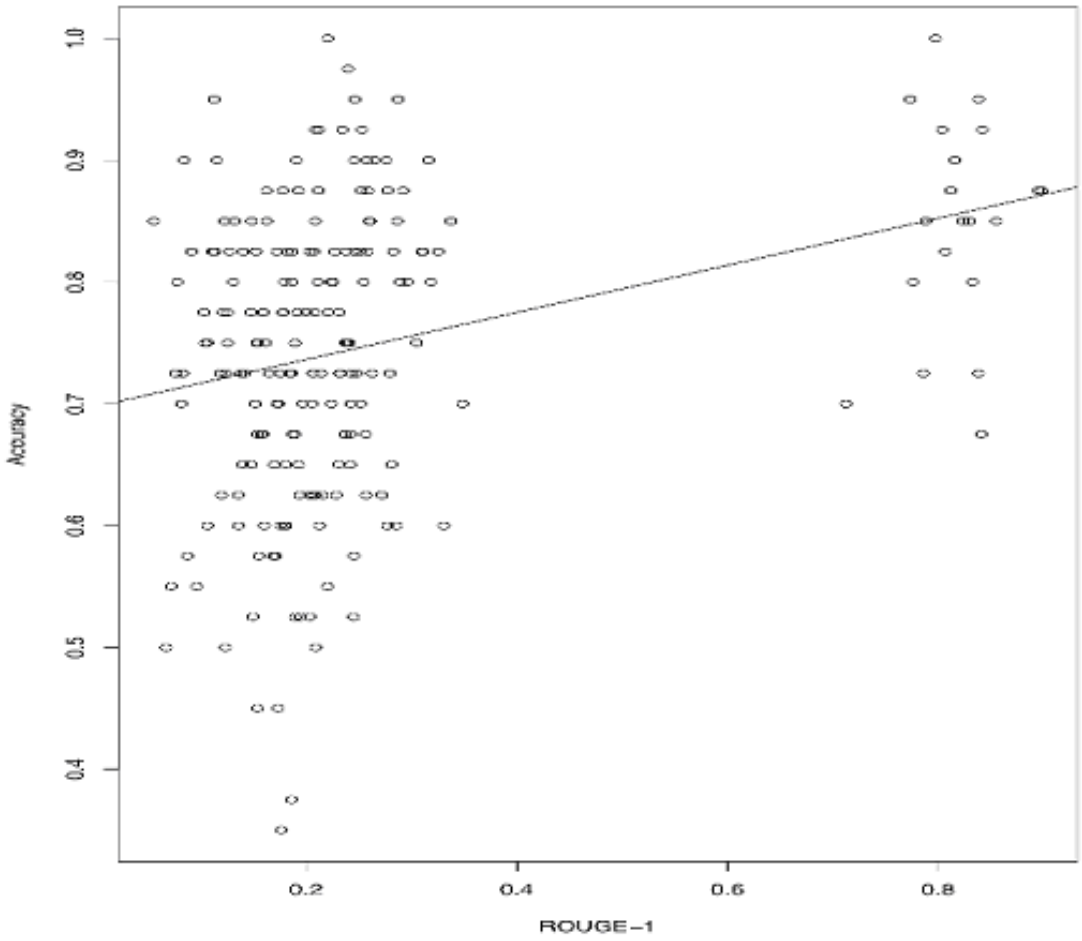

Figure 4.6: Scatter plot of Pearson r Correlation between ROUGE-1 and Accuracy with 200 Data Points (including Full Text)

	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	0.1374	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 4.32: Pearson r Correlation between Intrinsic and Extrinsic Scores Grouped by System-Topic Pair (excluding Full Text)

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 4.32.

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 (*).

Figure 4.7 shows the scatter plot corresponding to the ROUGE-1/Accuracy Pearson r correlation given in Table 4.32. As one can see, the data points form a rather evenly distributed cloud. The straight line that characterizes the least sum of squared differences between all points in the plot has a slope of 0.3569 and an intercept of 0.6665. The 0.3569 slope suggests there is some positive correlation between the accuracy and ROUGE-1, but the cloud-like appearance of the data points indicates that this correlation is weak.

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

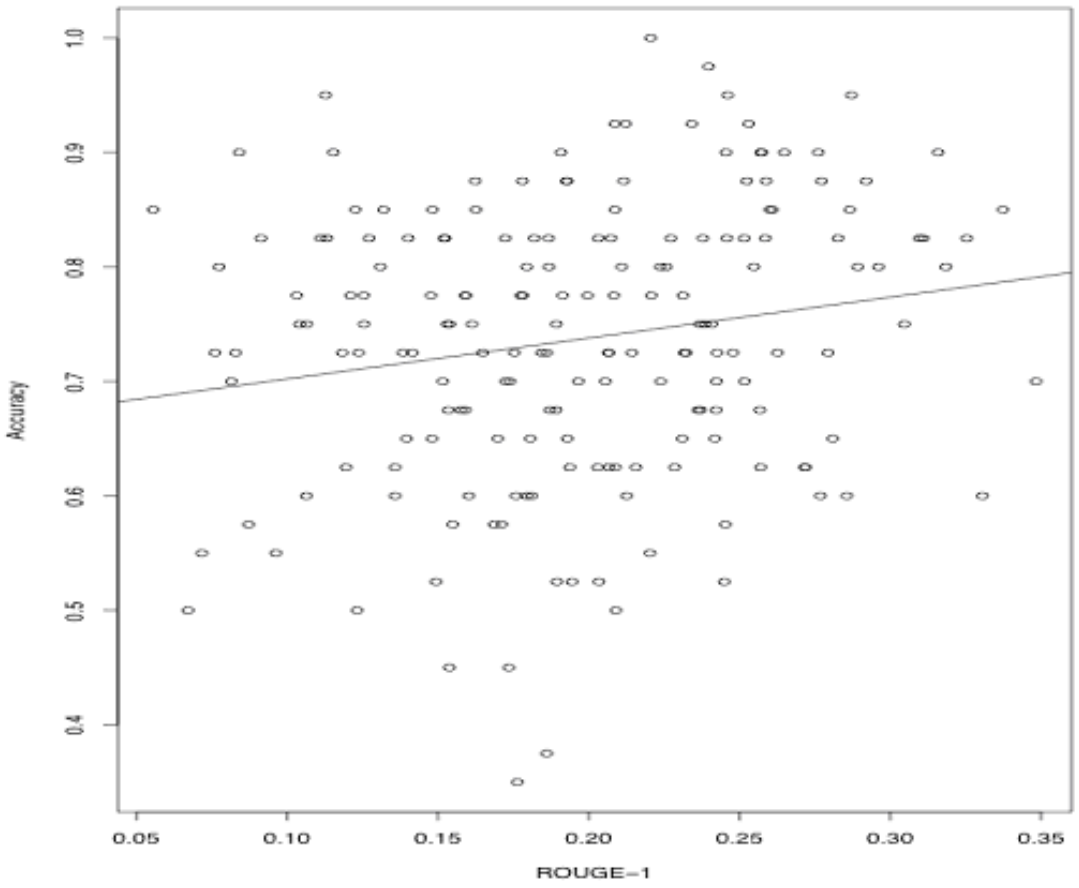

Figure 4.7: Scatter plot of Pearson r Correlation between ROUGE-1 and Accuracy with 180 Data Points (excluding Full Text)

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

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 4.33 shows the adjusted correlation—using Pearson r—for all pairs of intrinsic and extrinsic measures on all systems (excluding the full text). Figure 4.8 shows the scatter plot corresponding to the ROUGE-1/Accuracy Pearson r correlation given in Table 4.33.

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 4.34 and 4.35). 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 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.

4.2.9 Experimental Findings

These experiments 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

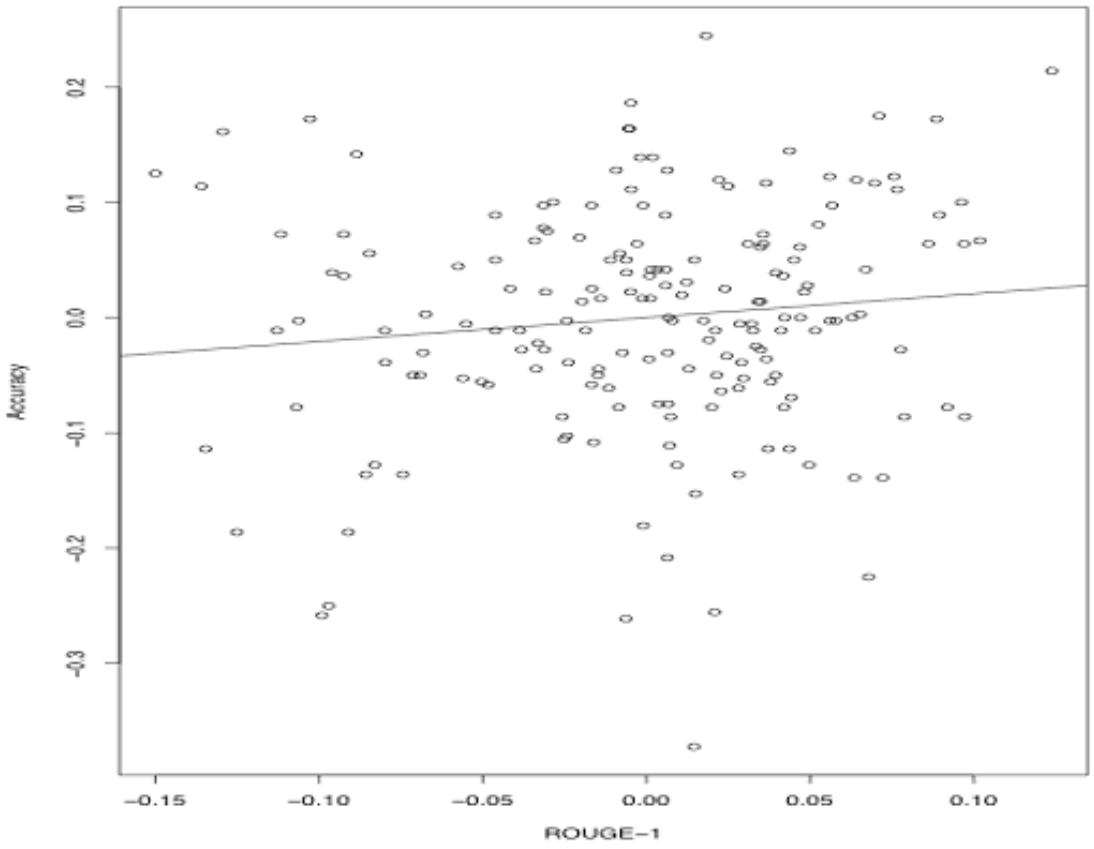

Figure 4.8: Adjusted Pearson r Correlation between Intrinsic and Extrinsic Scores Grouped by System-Topic Pair (excluding Full Text)

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

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

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 will explore 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 one 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

	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 4.36: Comparison of Summary/Document Judgments

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. Experiment part 2 was completed exactly a week after experiment part 1. This was designed to aid in decreasing or factoring out possible memory effects on making a summary judgment then its full text judgment or vice versa. In Methods SD1 and SD2, memory effects becomes a concern in that the judgments for the full text are made immediately after 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 addressed 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.

4.3.3 Preliminary Results and Analysis

Tables 4.36 and 4.37 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 4.37 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

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

Table 4.37: Additional Comparison of Summary/Document Judgments

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 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.

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

Table 4.38: Average Timing for Judgments on Summaries and Full Text Documents (in seconds)

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 4.38.

4.3.5 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 concerns with these experiments have been the low agreement results shown in Tables 4.10, 4.11, 4.22 and 4.23. It was then hypothesized 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 and 6 detail four 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

Proposed Work: 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. We have seen (in Chapter 4) that a measure that uses low-agreement human-produced annotations does not yield stable results. We have also 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 research is the introduction and use of a new measurement technique, the Relevance Prediction method, that yields more stable results.

This chapter reports some preliminary findings and lays the groundwork for further experiments using this new measurement technique. The experiment presented here (listed as RP with Human Summaries in Table 1.1) 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 method.

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

[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.

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).

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 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

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

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

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 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. 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

[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).

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)

Appendices A and C) were given to aid participants in determining requimements 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 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 Preliminary 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 yield 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), thus confirming the first hypothesis. The Relevance Prediction Precision and F-score results were also significantly higher than the LDC Agreement results with $p < 0.01$.

However, the second hypothesis was not confirmed. The Headline Relevance Prediction 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

Surrogate	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	Avg
Headline	0.80	0.80	0.85	0.70	0.73	0.60	0.80	0.75	0.60	0.75	0.88	0.68	0.80	0.93	0.83	0.77
Human	0.83	0.88	0.85	0.68	0.75	0.75	0.93	0.75	0.98	0.90	0.75	0.70	0.80	0.90	0.78	0.82

Table 5.3: Relevance Prediction Rates for Headline and Human Surrogates (Representative Partition of Size 4)

able to make consistent relevance decisions from the non-extractive Headline surrogates, even though these were shorter and less informative than the Human surrogates.

A closer look reveals that the Headline summaries sometimes contained enough information to judge relevance, yielding almost the same number of true positives (and true negatives) as the Human summaries. For example, a document about the formation of a coalition government to avoid violence in Cambodia has the Headline surrogate *Cambodians hope new government can avoid past mistakes*. By contrast, the Human surrogate for this same event was *Rival parties to form a coalition government to avoid violence in Cambodia*. Although the Headline surrogate uses words that do not appear in the original document (*hope* and *mistakes*), the task participant may infer the relevance of this surrogate by relating *hope* to the notion of forming a coalition government and *mistakes* to violence.

On the other hand, the lower degree of informativeness of Headline surrogates gave rise to over 50% more false negatives than the Human summaries (117 vs. 77). This statistically significant difference will be discussed further in Section 5.7.

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.

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 single-factor ANOVA. The higher Relevance Prediction rate supports our 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 our correlation studies (to be presented in Section 5.6) we did a further analysis where we took steps to reduce the effect of outliers. Specifically, we computed an average over all judgments for each user (20 judgments $\times$

Surrogate	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	Avg
Headline	0.70	0.73	0.85	0.70	0.63	0.60	0.60	0.85	0.50	0.73	0.70	0.78	0.65	0.63	0.73	0.69
Human	0.68	0.75	0.58	0.68	0.75	0.70	0.68	0.80	0.88	0.58	0.63	0.55	0.55	0.60	0.78	0.68

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

Surrogate	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 5.5: Average Rouge-1 Scores for Headline and Human Surrogates (Representative Partition of Size 4)

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 our 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 5.3 and 5.4. 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 additional provides further support for our hypothesis that Relevance Prediction is more reliable than that LDC Agreement 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 5.5 shows the ROUGE scores, based on 3 reference summaries per document, for partitions P1–P15 used in the previous tables.[5] The ROUGE 1-gram measurement (R1) is included here.[6] The ROUGE scores for Headline surrogates were slightly lower than those for Human surrogates. This is consistent with the earlier statements about the difference between 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 4.1.8).

As one might expect, there is some variability in the correlation between ROUGE and human judgments for the different partitions. However, the boxplots for both Headline and Human indicate that the first and third quartile were relatively close to the median (see Figure 5.1).

Table 5.6 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. The highest correlation was observed for Relevance Prediction on Headline.

The conclusion is that ROUGE correlates more highly with the Relevance Prediction measurement than the LDC-Agreement measurement, although it must be noted that none of the correlations in Table 5.6 were statistically significant at

[5] 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).

[6] ROUGE 2-gram, ROUGE L and ROUGE W were also computed, but the trend for these did not differ from ROUGE-1.

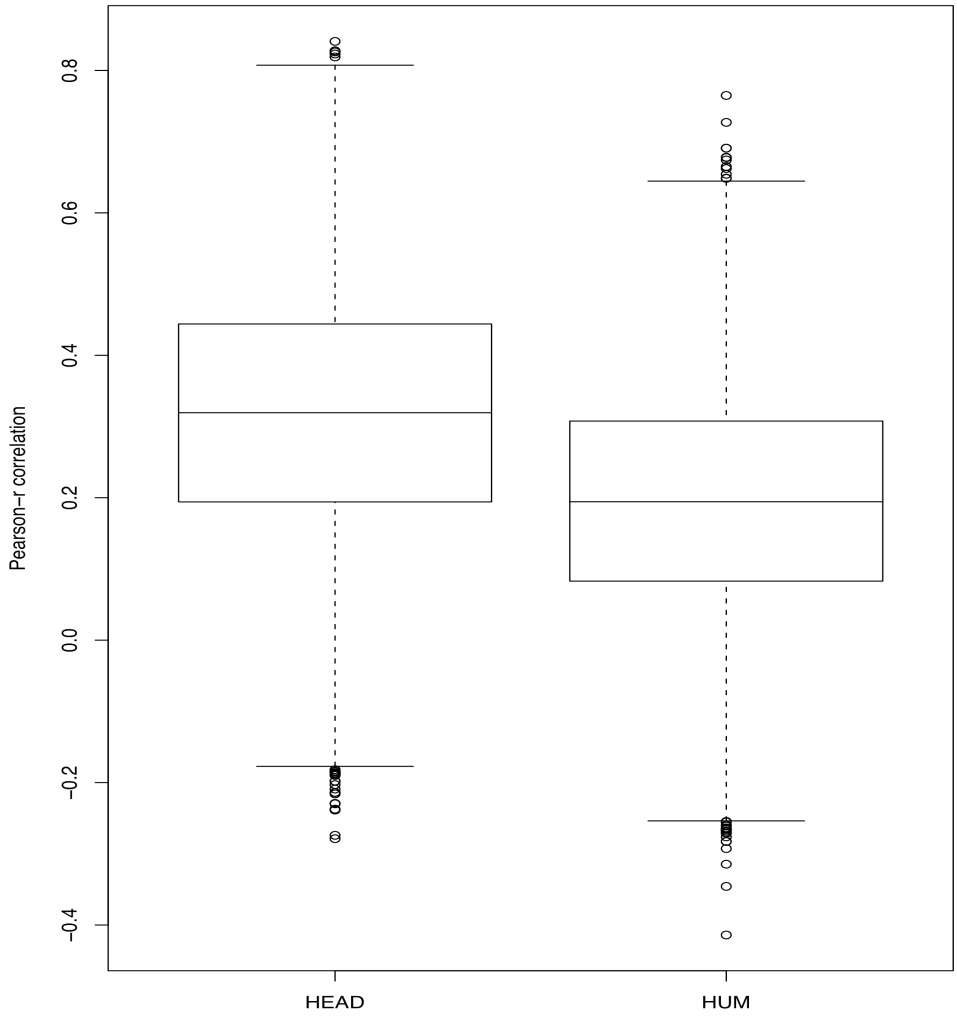

Figure 5.1: Distribution of the Correlation Variation for Relevance Prediction on Headline and Human

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

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 5.7: Users' Judgments and Corresponding Average ROUGE-1 Scores

$p < 0.05$. The low LDC-Agreement scores are consistent with previous studies where poor correlations were attributed to low interannotator agreement rates.

5.7 Experimental Findings

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.

A more detailed analysis of the users' judgments and the corresponding ROUGE-1 scores is given in Table 5.7, 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 5.7 the numbers in parentheses after each ROUGE score refer to the standard deviation for that score. This was computed as follows:

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

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.

Chapter 6

Additional Experiments

The primary contribution of this work is to determine the level of correlation between automatic intrinsic evaluation metrics and human extrinsic task performance. A new method for measuring agreement on extrinsic tasks, Relevance Prediction, was introduced and compared against the previous gold-standard based LDC-Agreement method. These contributions have been explored through preliminary studies, and the findings from those experiments encourage additional research. The Relevance Prediction method has been evaluated on human extrinsic tasks using only human-generated summaries. I propose conducting two additional experiments that incorporate both human and automatic summaries to evaluate the Relevance Prediction method on a relevance assessment task. These proposed experiments are described in more detail in Section 6.1 below.

In addition, all of the preliminary experiments discussed in Chapter 4 used only summaries generated from a single source document (single document summarization). The summarization community has recently become more interested in summaries that span more than one document (multi-document summarization) as can be seen with the Document Understanding Conference's transition from using single to multi-document summaries for their evaluation tasks (Dang, 2005). Therefore, an experiment investigating the correlations of automatic intrinsic measures with human extrinsic task performance and the results of the Relevance Prediction method with multi-document summaries is also proposed as future work and further described in Section 6.2.

6.1 Experiments 5 & 6: The Relevance Prediction Method with Automatic Summaries

In Experiment 4, RP with Dual Summaries (described in Chapter 5), the Relevance Prediction method was tested and compared against the LDC-Agreement method. This preliminary experiment used only human summaries; the original document headline and human-generated summaries that were commissioned for the study. The preliminary results of the Relevance Prediction method show that these measurements were more reliable than the those of the LDC-Agreement method for

human summaries. To further explore these findings, Experiments 5 and 6 were designed to incorporate automatic summaries in the evaluations.

Experiment 5, RP Dual Summary, is proposed to test the Relevance Prediction method using both human and automatic summaries. Participants made judgments on six events with the following systems:

- **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 experiment);
- **First75** – an automatic summary that uses the first 75 characters of the document;
- **Topiary** – an automatic summary that uses a hybrid topic list and fluent headline based on integration of the Unsupervised Topic Discovery and Trimmer systems (Zajic et al., 2004a).

An issue discovered with this experiment was that it was very long (it took approximately 3-4 hours to complete) and users commented that they grew tired and bored during their participation. To minimize the effects of user fatigue on the judgments, a similar and shorter experiment is proposed. This experiment, Constrained RP Dual Summary (Experiment 6), continues to investigate the Relevance Prediction method with the summarization systems listed above, but uses only three topics, and takes users approximately 1.5 hours to complete. Both Experiments 5 and 6 are ongoing work and the results are in the preliminary stages of evaluation.

6.2 Experiment 7: Multi-Document Summarization and Correlation with the Pyramid Method and Basic Elements

The primary focus of the 2005 Document Understanding Conference was to develop and test new intrinsic evaluation methods that take into account the variability of human-generated summaries (Dang, 2005). One of the previous issues with the Bleu and ROUGE metrics was their reliance on model summaries for comparison. It is possible for a good summary to be created that did not match the model summaries and would therefore receive a low Bleu and ROUGE score. The Pyramid Method and Basic Elements were created as intrinsic evaluation metrics that aim to compare the *content* of summaries for scoring rather than relying completely on term matching with a reference summary (see Section 2.2.2 for a description of these and other intrinsic metrics). The preliminary experiments discussed in this

paper were completed before data for the Pyramid Method and Basic Elements became available and therefore, an additional experiment that utilizes these two methods as the automatic intrinsic measures for correlations is suggested as future work. This will help to determine how the Relevance Prediction measure results correlate with these methods and to determine if these intrinsic methods correlate more highly than the Bleu and ROUGE metrics used in prior experiments.

The 2005 Document Understanding Conference also highlights the summarization community's concentration on summaries spanning two or more documents rather than summaries from a single source text. Significant advances in the area of single document summarization has prompted a shift in focus to multi-document summarization. The preliminary experiments discussed in this paper have all relied on single document summaries, therefore, multi-document summaries will be used in Experiment 7 to investigate the differences that summarization type may have on the correlations and user judgment.

The above proposed experiments will further investigate the Relevance Prediction method with automatic summaries and summaries that span more than one document. Some of the newer extrinsic evaluation methods, such as the Pyramid Method and Basic Elements, will also be used to determine their level of correlation with human performance on the extrinsic task.

It is hypothesized that the proposed experiments will show continued high agreement results with the Relevance Prediction method, higher correlations with Relevance Prediction than with LDC-Agreement, statistically significant differences in the performance of the experimental systems, and correlation results that yield strong statistical statements about the predictive power of the automatic and semi-automatic intrinsic measures.

The expected contributions of the additional experiments are:

- The continued exploration of agreement measurements produced by the Relevance Prediction method.
- Further investigation of correlations with current extrinsic evaluation metrics and extrinsic human task performance using the Relevance Prediction method.
- Determination of any differences in human performance and correlations with single document versus multi-document summaries.
- Further motivation of the usefulness of text summarization.

Chapter 7

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.

Chapter 8

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

Chapter 9

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.

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 Lingusitics*, 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.

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, Christof Monz, Douglas Oard, Stacy President, and David Zajic. Extrinsic Evaluation of Automatic Metrics for Summarization. Technical report, University of Maryland, College Park, MD, 2004. LAMP-TR-115, CAR-TR-999, CS-TR-4610, UMIACS-TR-2004-48.

Bonnie J Dorr, Christof Monz, Stacy President, Richard Schwartz, and David Zajic. A Methodology of Extrinsic Evaluation of Text Summarization: Does ROUGE Correlate? In *Proceedings of the Association for Computational Linguistics (ACL) Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, MI, June 2005.

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.

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.

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. Technical report, University of Maryland, College Park, MD, 2005. LAMP-TR-119, CS-TR-4695, UMIACS-TR-2005-14.

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.