LAMP-TR-115
CAR-TR-999
CS-TR-4610
UMIACS-TR-2004

July 20, 2004

Extrinsic Evaluation of Automatic Metrics for Summarization[*]

Bonnie Dorr, Christof Monz, Douglas Oard, Stacy President, David Zajic
University of Maryland Institute for Advanced Computer Studies

Richard Schwartz
BBN Technologies

Abstract

This paper describes extrinsic-task evaluation of summarization. We show 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. In addition, we demonstrate that the extrinsic task we have selected exhibits a high degree of interannotator agreement, i.e., consistent relevance decisions across subjects. We also conducted a composite experiment that better reflects the actual document selection process and found that using a surrogate improves the processing speed over reading the entire document. Finally, we have found a small yet statistically significant correlation between some of the intrinsic measures and a user's performance in an extrinsic task. The overall conclusion we can draw 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.

[*] This work has been supported, in part, by DARPA TIDES Cooperative Agreement N66001-00-2-8910, BBNT Contract 9500004957, and NSF CISE Research Infrastructure Award EIA0130422.

Introduction

People often prefer to read a summary of a text document, e.g., news headlines, scientific abstracts, movie previews and reviews, meeting minutes. With the explosion of online textual material, automatic summarization is of great interest to those with this need. Automatic summaries may span one or more documents (Radev and McKeown, 1998; Mani and Bloedorn, 1999) and may be obtained monolingually or crosslingually (Dorr et al., 2003a). Our focus in this report is on single-document English summaries of English texts.

In order to demonstrate the usefulness of automatic summaries, the summarization community has turned to the development and evaluation of both intrinsic and extrinsic measures (Sparck-Jones and Galliers, 1996) for use in large-scale evaluations, e.g., SUMMAC (Mani et al., 2002) and the Document Understanding Conference (DUC; Harman and Over, 2001;2002;2003;2004).

Recently, the ROUGE metric (Lin, 2003) has been adopted as NIST's standard for automatic intrinsic evaluation of summarization systems. ROUGE is said to correlate highly with the results of human judgments of content and quality (Lin, 2004). However, no results have been reported on the correlation of ROUGE to human performance on an extrinsic task. (For example, the DUC task involved human comparisons of system output to a set of 'ideal' summaries.) The goal of our work is to determine the degree of correlation of ROUGE—and also a contrastive automatic metric (BLEU; Papaneni, 2002)—to measure human performance on a real-world task.

In selecting the extrinsic task it is important that the task be unambiguous enough that subjects can perform it with a high level of agreement. If the task is so difficult that subjects 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.

In an earlier experiment (Zajic et al., 2004a) we attempted to use document selection in the context of information retrieval as an extrinsic task. Subjects were asked to decide if a document was highly relevant, somewhat relevant or not relevant to a given query. However we found that subjects who had been shown the entire document were only able to agree with each other 75% of the time and agreed with the allegedly correct answers only 70% of the time. We were unable to draw any conclusions about the relative performance of the summarization systems, and thus were not able to make any correlations between human performance and scores on automatic summarization evaluation tools.

This document describes a more constrained type of document relevance judgment as an appropriate extrinsic task for evaluating human performance using automatic summarizations. The task, *event tracking*, has been reported in NIST TDT evaluations to provide the basis for more reliable results. 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.

Subjects are asked to decide if a document contains information related to a particular event in a specific domain. The subject 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 relevant.

We test the hypothesis 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) Our 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/DUC guidelines.

A second hypothesis is that the task we have selected supports a very high degree of interannotator agreement, i.e., consistent relevance decisions across subjects. This is similar to the "consistency test" applied in SUMMAC, except that we apply it not just to the full-text versions of the documents, but also to all types of summaries. (In addition, to validate our hypothesis, we require a much higher degree of agreement—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 (Di 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 accuracy (the extrinsic measure)—in order to establish an automatic (inexpensive) predictor of human performance.

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 a user study. 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 subjects in order to establish statistical significance.

The results of our experiments provide some support for the first two hypotheses. However, the first hypothesis is only weakly supported in an additional "composite-user" experiment that better reflects the actual document selection process. We found that using a surrogate improves the processing speed minimally over reading the entire document—with a significant degradation in recall. Regarding the third hypothesis, we have found a small yet statistically significant correlation between some of the intrinsic measures and a user's performance in an extrinsic task. The overall conclusion we can draw 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.

Summarization Evaluation Methods

Extrinsic evaluation tasks are used to determine how well summarization systems aid humans in performing a specific task. In our experiment, the extrinsic task was that of judging relevance of a document to a specific event. The documents were displayed as output from one of four types of systems: (1) the full document itself (Text)—the upper baseline; (2) a human-generated 75-character summary (Human); (3) a human-generated headline associated with the original document (Headline); or (4) a 75-character summary from one of six automatic summarization systems (described in more detail in the Experimental Setup section below). We also included a lower baseline summary—the first 75 characters of the document.

We used two categories of evaluation methods for evaluating the text summaries: intrinsic and extrinsic. Intrinsic evaluations test the summarization systems themselves. In contrast to the earlier study by (Mani et al., 2002), our intrinsic evaluation methods are automatic: we use BLEU and ROUGE. Recently, Passonneau and Nenkova (2003, 2004) proposed a new intrinsic evaluation method called Pyramid that takes into account how often content units from the peer summary occur in the reference summaries. The method looks very promising and the authors show high correlation with human judgments, but the identification of content units requires some degree of human involvement—although it might be possible to automate this process to some extent—and the purpose of our study is to focus on fully automated measures for intrinsic evaluation. Thus, we restrict ourselves to the aforementioned BLEU and ROUGE measures. Our goal is to test the degree to which these measures correlate with the performance of the systems in an extrinsic task.

The second category of evaluation, an extrinsic evaluation, tests the summarization based on how it affects the completion of the tasks of judging document relevance. We expect that the three human systems, Text, Human and Headline allow the users to make the most accurate judgments—although we show that humans are not significantly more accurate than the top-performing summarization systems. Of the three, the performance of the Text system is the upper baseline, since none of the information is eliminated in the document display. The work of Marcus et al. (1978) also supports this idea by suggesting that the quality of a *surrogate* (i.e., the description used by the human subject to judge relevance of the associated document) is proportional to its length. That is, the greater the length (where the full text is the maximal length), the higher the quality of the surrogate and the better the performance of the human in making the associated judgment.

Data Set

In our experiment, we selected 20 topics from the TDT-3 corpus (Allan et al., 1999). For each topic, a 20-document subset was created from the top 100 ranked documents retrieved by the FlexIR information retrieval system (Monz and de Rijke, 2001). Crucially, each subset was constructed such that exactly 50% of the documents were relevant to the topic. The full-text documents ranged in length from 42 to 3083 words. The documents are long enough to be worth summarizing, but short enough to be read within a reasonably short amount of time. The documents consisted of a combination of news stories stemming from the Associated Press newswire and the New York Times. The topics includes 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 I for details.)

Each topic included an event description and a set of 20 documents. An example of an event description is shown in Table 1. The *Rule of Interpretation* 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.

PanAm Lockerbie Bombing Trial
WHAT: Kofi Annan visits Libya to appeal for surrender of PanAm bombing suspects *WHERE:* Tripoli, Libya *WHO:* U.N. Secretary-General Kofi Annan; Libyan leader Moammar Gadhafi *WHEN:* December, 1998
Kofi Annan went to Libya in December to appeal to the Libyan government to hand over suspects in the 1988 bombing of Pan Am Flight 103, which killed 270 people over the Scottish town of Lockerbie. Libya has resisted handing over the suspects for fear that they would not receive a fair trial. The legal disputes surrounding this case have been going on for years; there was a flurry of activity in the case in late 1998 surrounding the time of Annan's trip. On topic: Stories covering any aspect of this trial: legal and political negotiations, criminal investigations, reactions from around the world to the case. *NOTE:* Although the seminal event for this topic focuses on the trial rather than the crime itself, our rule of interpretation for legal/criminal cases extends the topic to include everything from the crime to the trial and sentencing. Therefore, stories about the Lockerbie bombing itself are on topic.
Rule of Interpretation Rule 3: Legal/Criminal Cases

Table 1: Example of Event Description

The TDT-3 data also provided ‘gold-standard’ judgments – each document was marked ‘relevant’ or ‘not relevant’ with respect to the associated event. These ‘gold-standard’ judgments were used in our analysis to produce accuracy, precision, recall, and f measure results.

Systems Evaluated

Ten systems were used to produce surrogates in our experiments—six of which were automatic summarization systems. As in the SUMMAC evaluations, we designed our experiment as a black-box evaluation, i.e., we examined the summarization systems as a whole rather than their internal components. No attempt was made to classify the different technologies to study their effect on performance. The 10 systems are described in Table 2.

An example of each system’s output associated with the event in Table 1 is displayed in Table 3. The output of each system, excluding the full document text and the original document headline, was constrained to be no more than 75 characters. With the exception of UTD, the automatic summaries were extractive—i.e., consisting only of words from the full text document. Note that the output of each system contains bold-faced *descriptor words*, i.e., words that occur in the event description, so that the user is presented with a readable format that most closely simulates that of a real-world task.

System	Description	Mean Length Chars	Mean Length Words
TEXT	Full text of the document.	3696	594
Headline	The original headline associated with the document	54	9
Human	A 75-character summary written by a human. The summaries were commissioned for this experiment.	71	11
First75	The first 75 characters of the document. This is intended to be a lower-bound baseline	75	12
KWIC	Keywords in Context (Monz, 2003, 2004)	71	11
GOSP	Global word selection with localized phrase clusters (Zhou and Hovy, 2003)	75	11
ISIKWD	Topic-independent keyword summary (Hovy and Lin, 1997)	75	11
UTD	Unsupervised Topic Discovery. (Schwartz et al., 2001)	71	9
Trimmer	Fluent headline based on a linguistically-motivated parse-and-trim approach (Dorr et al., 2003b)	56	8
Topiary	Hybrid topic list and fluent headline based on integration of UTD and Trimmer (Zajic et al., 2004b)	73	10

Table 2: Ten Summarization Systems

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
KWIC	**Gadhafi** to surrender two Libyans wanted in the **1988 bombing** of a **PanAm**
Trimmer	Ugandan President Yoweri Museveni flew apparently violating **U.N.** sanctions
First75	Ugandan President Yoweri Museveni flew to **Libya**, apparently violating **U.N.**
Topiary	NEWS **LIBYA** Ugandan President Yoweri Museveni flew violating **U.N.** sanctions
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

Table 3: Example Output From Each Experimental System

Experiment Design

Fourteen undergraduate and six graduate students at the University of Maryland at College Park were recruited through posted experiment advertisements to participate in the experiment. They were asked to provide information about their educational background and experience (Appendix II). 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 III.

Each of the 20 topics, T_1 through T_{20}, consists of 20 documents corresponding to one event. The twenty human subjects 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, we were able to collect data for an analysis of within-group judgment agreement.

	T_1,T_2	T_3,T_4	T_5,T_6	T_7,T_8	T_9,T_{10}	T_{11},T_{12}	T_{13},T_{14}	T_{15},T_{16}	T_{17},T_{18}	T_{19},T_{20}
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
KWIC	D	E	F	G	H	I	J	A	B	C
First75	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
TOPIARY	H	I	J	A	B	C	D	E	F	G
TRIMMER	I	J	A	B	C	D	E	F	G	H
UTD	J	A	B	C	D	E	F	G	H	I

Table 4: Latin Square Experiment Design

Each human subject was asked to evaluate 22 topics (including two practice event topics not included in our 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 $20\times2=40$ 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, is 8000.

A Latin square design (as shown above in Table 4) 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 subjects performed the experiment on a Windows or Unix workstation, using a web-based interface we developed to display the event, document descriptions and to record the judgments. The subjects 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.

Extrinsic Evaluation of Human Performance

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

To compute accuracy, we rely on the TDT-3 'gold-standard' relevance judgments associated with each event. Based on these judgments, we take accuracy is 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, we have enforced a 50% relevant/irrelevant split across each document set. This balanced split justifies the inclusion of true negatives in our 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.)

Although accuracy is the primary metric we used to investigate the correlations between intrinsic and extrinsic measures, we imported other metrics commonly used in the IR literature (following the lead of the SUMMAC experimenters). The contingency table for the extrinsic task is shown in Table 5, 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:

	Judged Relevant	Judged Non-Relevant
Relevant is True	TP	FN
Relevant is False	FP	TN

Table 5: Contingency Table for Extrinsic Task

Using this contingency table, the full set of extrinsic measures is given here:

Accuracy = (TP + TN)/(TP + TN + FP + FN)
Precision = TP/(TP + FP)
Recall = TP/(TP + FN)
F-score = (2 × Precision × Recall)/(Precision + Recall)

In addition to computing these four metrics, we will use Tukey's Studentized Range criterion, called the Honestly Significant Difference (HSD) (cf., Hinton, 1995, for a cogent description), to test whether differences between groups of systems are statistically significant.

Table 6 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 on Accuracy, which is the focus of our attention for the remainder of this report.

	TP	FP	FN	TN	A	P	R	F	T
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 6: Results of Extrinsic Tasks Measures on Ten Systems, sorted by A(ccuracy)

We computed one-factor repeated-measures ANOVA to determine if the differences among the systems were statistically significant for five measures: precision, recall, f-score, accuracy, and time. Each subject saw each system twice during the experiment, so each sample consisted of a subject's judgments on the 40 documents that comprised the two times the subject 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 the differences are significant for all five measures, with p<0.01. However, this test only guarantees that one pair of systems is significantly different. In order to determine which pairs of system are significantly different, we applied the post hoc Tukey test to compute the Honestly Significant Difference (HSD).

The HSD is shown for each measure in the bottom row of Table 6 with p<0.05. If the difference in measures between two systems is greater then the HSD, then we can claim a significant difference between the systems. 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 we cannot claim a significant difference 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 we can claim a significant difference between UTD and HUMAN on accuracy.

Unfortunately, we are unable to claim any significant differences with p<0.05 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. Thus, we reanalyzed the automatic summarization systems without the human-generated summaries or the full text.

Using the mean scores from Table 6, we tested whether the results of ANOVA indicate significant differences among the extrinsic measures for 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 us to group the automatic systems into two overlapping sets, the members of which are not significantly distinct according to the Tukey test. This is shown in Table 7.

First 75	A	
GOSP	A	
ISIKWD	A	
TOPIARY	A	B
TRIMMER	A	B
UTD		B
KWIC		B

Table 7: Equivalence Classes of Automatic Summarization Systems with respect to Precision

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

Note that the human-generated systems—Text, Human and Headline—performed best with respect to Accuracy, with the Text system as the upper baseline, consistent with our initial expectations. However, our 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 our 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.

Agreement on Relevant and Irrelevant Documents

Recall that our second hypothesis is that the task we have selected supports a very high degree of interannotator agreement—beyond the low rate of agreement (16-69%) achieved in the SUMMAC experiments. Table 8 shows “Subject

Agreement," i.e., agreement of both relevant and irrelevant judgments of subjects within a group, and the kappa score based on subject agreement.

	Subject Agreement	Kappa Score
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 8: Subject Agreement and Kappa Score

The Kappa score is calculated as:

$$(P_A - P_E) / (1 - P_E)$$

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

Intrinsic Evaluation of Summarization Systems

In contrast to SUMMAC—which focused on an extrinsic-task evaluation—we have also examined the problem of intrinsic evaluation using automatic metrics. We will use BLEU (Papineni et al. 2002) and ROUGE (Lin and Hovy, 2003) 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. We commissioned three 75-characters summaries (in addition to the summaries in the HUMAN system) to use as references. BLEU and ROUGE were run with 1-grams through 4-grams, and two additional variants of ROUGE (-L and –W-1.2) were run. The results are shown in Table 9.

	R1	R2	R3	R4	RL	RW	B1	B2	B3	B4
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 9: ROUGE and BLEU scores on Ten Systems, sorted by ROUGE-1

Analogously to the extrinsic evaluation measures discussed above, we computed the ANOVA values 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 Figure 1 and Figure 2, respectively. In both graphic representations, the 95% confidence interval is shown by the error bars on each line.

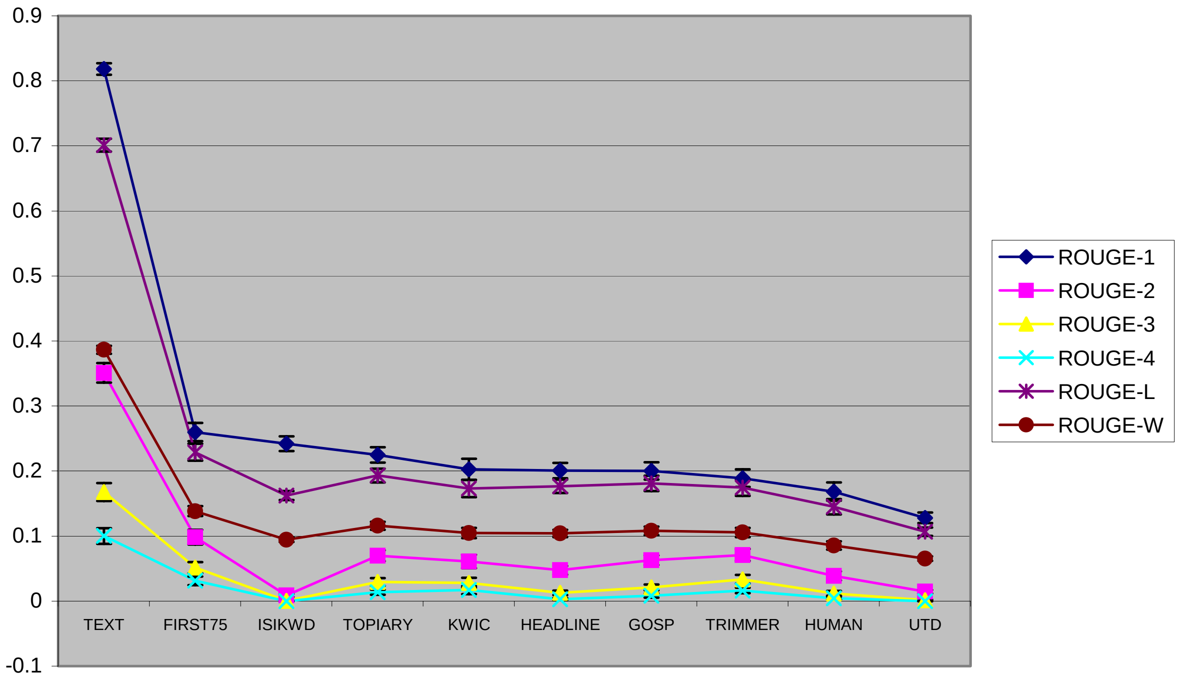

Figure 1: ROUGE Results for Ten Systems (X axis ordered by ROUGE-1)

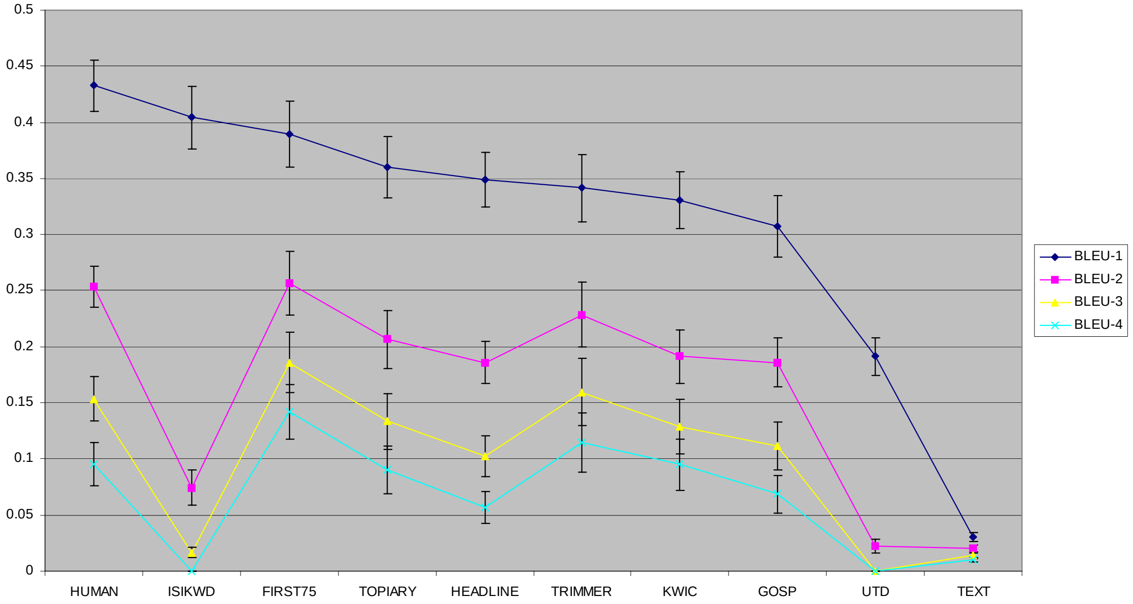

Figure 2: BLEU Results for Ten Systems (X axis ordered by BLEU-1)

In Figure 1, we see that 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 2, we see that the full document representation performs 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.

We also performed the ANOVA test 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 us to compute the honestly significant differences (HSD) for each measure, which is shown in Table 10.

	ROUGE-1	**ROUGE-2**	**ROUGE-3**	**ROUGE-4**	**ROUGE-L**	**ROUGE-W**	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**
HSD,p<.05	**0.04**	**0.03**	**0.02**	**0.01**	**0.04**	**0.02**	**0.09**	**0.09**	**0.06**	**0.05**

Table 10: Honestly Significant Differences for Automatic Summarization Methods Using Intrinsic Measures

As we did for the extrinsic measures above, we can group the different summarization systems, based on the honestly significant difference. For illustration purposes we show the groupings for ROUGE-1 and BLEU-1 in Table 11 and Table 12.

FIRST75	**A**			
ISIKWD	A	B		
TOPIARY	A	B	C	
KWIC		B	C	
GOSP		B	C	
TRIMMER			C	
UTD				D

Table 11: 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 12: Equivalence Classes of Automatic Summarization Systems with respect to BLEU-1

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

Alternate Interpretation: Composite Subjects

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. We will now describe our implementation of a composite-user experiment using the results of our experiments.

Two subjects made judgments about 20 documents for each combination of topic and summarization system. In particular two subjects 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 subjects who made judgments about Topic$_i$-System$_j$ and two different subjects who made judgments about Topic$_i$-FullText. There are four possible combinations of Topic$_i$-System$_j$ subjects and Topic$_i$-FullText subjects. Each combination is considered to be a composite subject.

We imagine that each relevance judgment represents one part of a two-stage process. First a Topic$_i$-System$_j$ subject makes a judgment about a document. If the first-stage judgment is positive, we consider the judgment of the Topic$_i$-FullText subject as the second stage. If the second-stage judgment is also positive, we consider the composite subject to have made a positive judgment. If either the first-stage or second-stage judgment is negative, we consider the composite subject to have made a negative judgment. The time for one session of 20 documents is computed as the time taken by the first-stage subject for that session, plus $\frac{n}{20}$ of the time taken by the second-stage subject, where n is number of documents accepted by the first-stage subject. The results of this composite-subject experiment are shown in Table 13.

System	**TP**	**FP**	**FN**	**TN**	P	R	F	A	T
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<0.5					0.091	0.120	0.107	0.059	

Table 13: Composite User simulation results, sorted by Accuracy

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

This simulation has some predictable differences from the activity it models. If the first-stage subjects 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, we expect that the first-stage subjects will pass about half of the documents on to the second stage. Therefore we would expect the time to be approximately the time of the summary systems plus half the time of the full text. Using values from Table 6:

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

The times calculated in the composite-subject 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 subjects 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 subjects 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 subjects 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.)

Correlation of Intrinsic and Extrinsic measures

To test our third hypothesis, we compared the results of the automatic metrics with 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 we will see shortly, this correlation is low when we consider the summaries alone (i.e., if we exclude the full text).

We start by computing correlation on the basis of the average performance of a system for all topics. As we saw above, there are significant differences between human performance measures and the scoring by the automatic evaluation systems. Table 14 through Table 16 below show the rank correlations between the average system scores assigned by the task-based metrics from Table 6 and the automatic metrics from Table 9. We used two methods for computing this correlation: Pearson r and Spearman ρ (Siegel and Castellan, 1988).

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.

First, we compute the intrinsic and extrinsic scores for each summarization method, averaging over the individual topics. 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 14 shows the results for Pearson *r* correlation (Siegel and Castellan, 1988), where s_i are the values of e.g. precision score of system *i*, and $\bar{s}$ are the average scores e.g., precision, over all systems, including full text. 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 our primary focus for the remainder of this report.

	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 14: Pearson *r* Correlation between Intrinsic and Extrinsic Scores Grouped by System (including full text)

Looking back at Figure 1 and Figure 2, we see that full text has much higher ROUGE scores than any of the other systems, and also that 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 us to draw valid statistical inferences. If we treat the full text as an outlier, removing it from the set of systems, the correlations are significantly weaker. (We will return to this point again shortly.) Table 15 shows the results for Pearson *r* over all systems, excluding full text.

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

Table 15: Pearson *r* Correlation between Intrinsic and Extrinsic Scores Grouped by System (excluding full text)

Spearman *ρ* is computed exactly like the Pearson *r* correlation, but instead of comparing actual scores, one compares the system ranking based on an intrinsic measure with the system ranking based on an extrinsic measure. We show the Spearman *ρ* correlation between intrinsic and extrinsic scores—excluding the full text—in Table 16 below.

	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 16: Spearman *ρ* Correlation between Intrinsic and Extrinsic Scores Grouped by System (excluding full text)

Table 15 and Table 16 show that there is a positive correlation in some cases, but it also shows that all positive correlations are rather low. Tests of statistical significance indicate that none of the Pearson *r* and Spearman *ρ* correlations is statistically significant at level p<0.05.

Computing correlation on the basis of the average performance of a system for all topics has the disadvantage that there are only 10 data points which leads to rather unstable statistical conclusions. In order to increase the number of data points we redefine a data point 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 (we compare ten summarization systems) and *n* = 1...20 (we are using 20 topics). 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 17.

	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 17: Pearson *r* Correlation between Extrinsic and Intrinsic Scores Grouped by System-Topic Pair—200 Data Points (including full text)

Having a sufficiently large number of data points allows us also to inspect a scatter plot showing the correlation between intrinsic and extrinsic measures. Figure 3 shows the scatter plot corresponding to the ROUGE-1/Accuracy Pearson correlation given in Table 17.

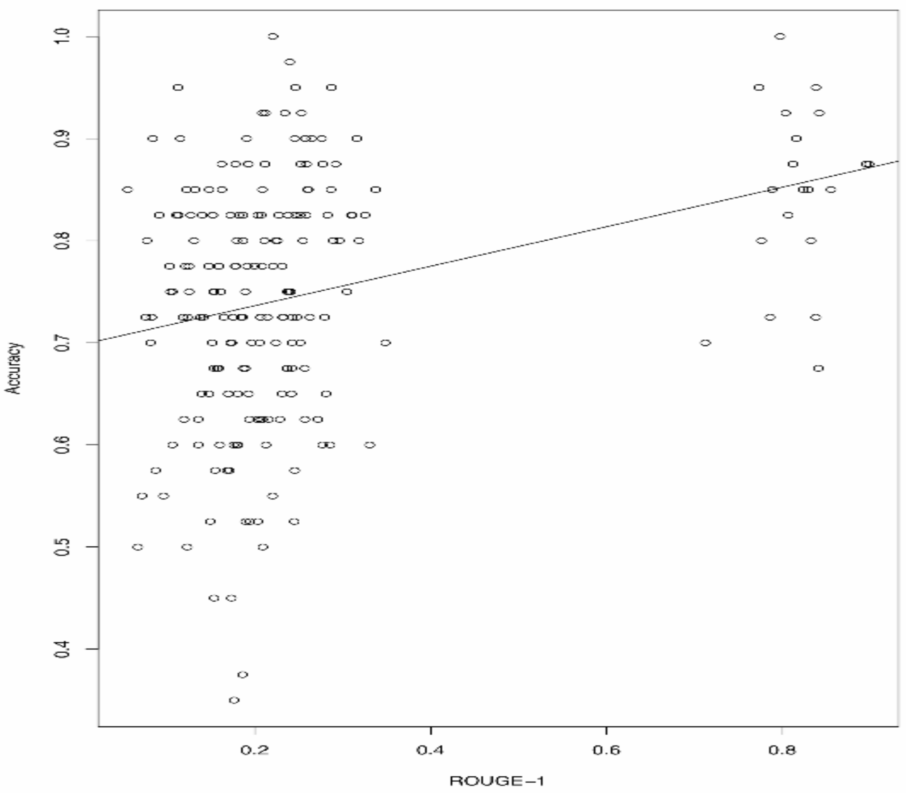

Figure 3: Scatter plot of Pearson r Correlation between ROUGE-1 and Accuracy—200 Data Points (including full text)

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, we see that this is not the case in Figure 3. 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 our main interest.

Because we are primarily interested in the performance with respect to summaries only, we will remove the 20 data points that use full text from the data set and the following discussion is based on the remaining 180 data points only. The Pearson *r* correlation for all pairs of intrinsic and extrinsic measures on all systems, excluding the full text, is shown in Table 18.

	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	0.134	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 18: Pearson *r* Correlation between Extrinsic and Intrinsic Scores Grouped by System-Topic Pair (excluding full text)

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

Figure 4 shows the scatter plot corresponding to the ROUGE-1/Accuracy Pearson *r* correlation given in Table 18. 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 we are mainly interested in the effect of system performance, we want to eliminate the effect of topic difficulty while maintaining a reasonable sample size of data points.

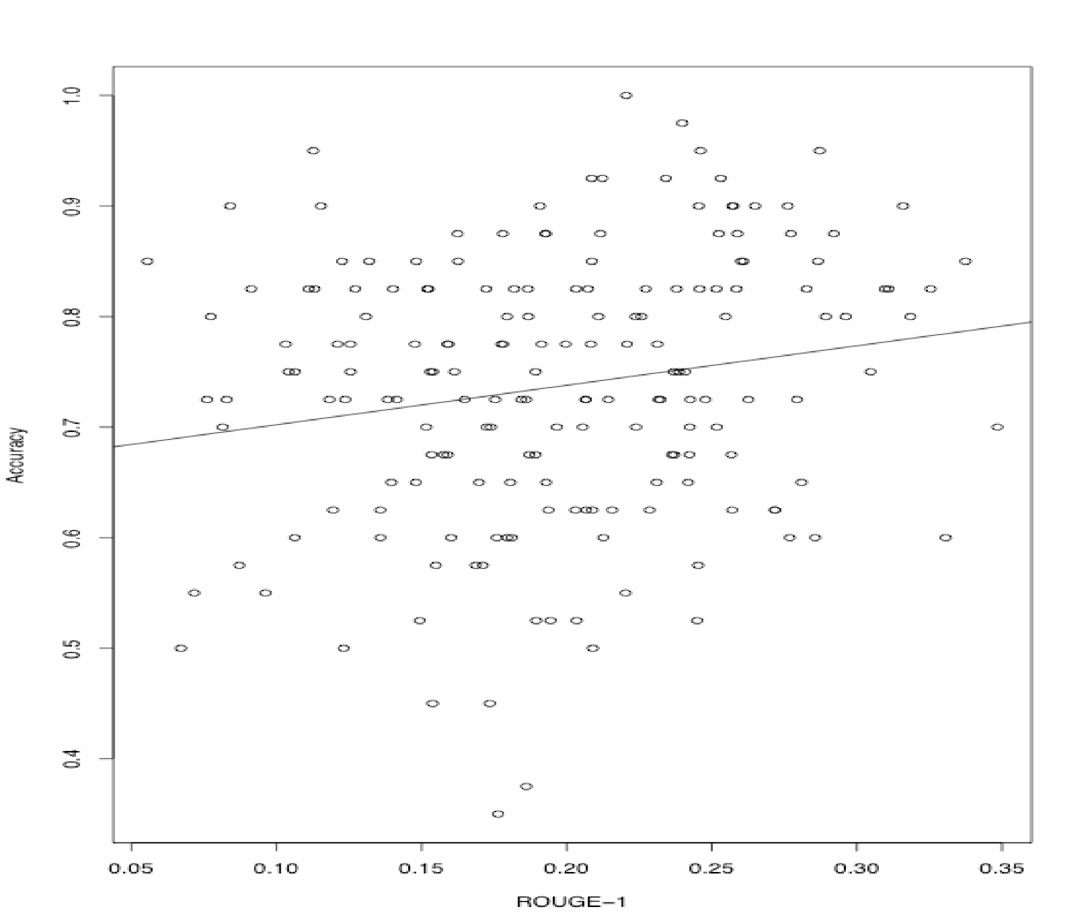

Figure 4: Scatter plot of Pearson *r* Correlation between ROUGE-1 and Accuracy—180 Data Points (excluding full text)

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

Table 19 shows the adjusted correlation—using Pearson r—for all pairs of intrinsic and extrinsic measures on all systems (excluding the full text). Figure 5 shows the scatter plot corresponding to the ROUGE-1/Accuracy Pearson *r* correlation given in Table 19.

	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 19: Adjusted Pearson *r* Correlation between Intrinsic and Extrinsic Scores Grouped by System-Topic Pair (excluding full text)

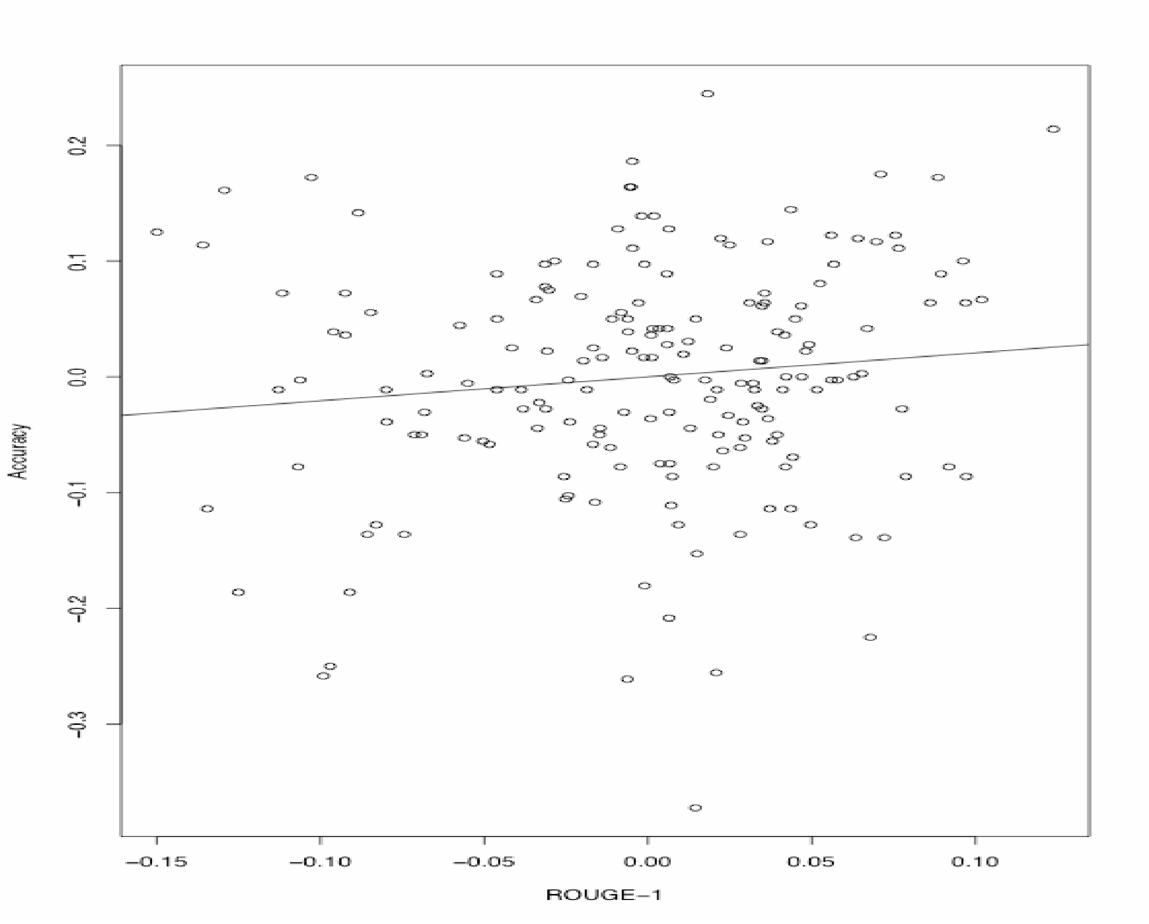

Figure 5: Scatter Plot of Adjusted Pearson *r* Correlation between ROUGE-1 and Accuracy—180 Data Points (excluding full text)

For completeness, as above, we computed the Spearman *ρ* rank correlation between intrinsic and extrinsic evaluation measures, for both the non-adjusted and adjusted cases. (See Table 20 and Table 21.) 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.

	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 20: Spearman *ρ C*orrelation between Intrinsic and Extrinsic Measures 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 21: Adjusted Spearman *ρ* Correlation between Intrinsic and Extrinsic Measures Grouped by System-Topic Pair (excluding full text)

Conclusions

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 of cases, a stricter non-parametric correlation measure such as Spearman ρ only showed a significant correlation in one case.

The overall conclusion we can draw 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.

References

Allan, J., H. Jin, M. Rajman, C. Wayne, D. Gildea, V. Lavrenko, R. Hoberman, and D. Caputo. Topic-based novelty detection: 1999 summer workshop at CLSP, final report, 1999.

Carletta, J. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics* 22(2):249-254, 1996.

Di Eugenio, B. and M. Glass. The Kappa statistic: A second look. Computational Linguistics 30(1). 2004

Dorr, Bonnie J., David Zajic, and Richard Schwartz. 2003a. "Cross-Language Headline Generation for Hindi," *ACM Transactions on Asian Language Information Processing (TALIP)*, 2:2.

Dorr, Bonnie J., David Zajic and Richard Schwartz. 2003b. "Hedge Trimmer: A parse-and-trim approach to headline generation," In *Proceedings of the HLT-NAACL 2003 Text Summarization Workshop,* Edmonton, Alberta, Canada.

Harman, Donna and Paul Over. 2001. *Proceedings of DUC-2001.*

Harman, Donna and Paul Over. 2002. *Proceedings of DUC-2002.*

Harman, Donna and Paul Over. 2003. *Proceedings of DUC-2003.*

Harman, Donna and Paul Over. 2004. *Proceedings of DUC-2004.*

Hinton, Perry, 1995. “Statistics Explained,” Routledge, New York, NY.

Hovy, Eduard and Lin, Chin Yew (1997). “Automated Text Summarization in SUMMARIST,” In *Proceedings of the Workshop on Intelligent Scalable Text Summarization*, pages 18-24, Madrid, Spain, August 1997. Association for Computational Linguistics.

Kendall, M.. A new measure of rank correlation. *Biometrika,* 30(1 2):81-93, 1938.

Lin, Chin-Yew and Eduard Hovy. 2003. “Automatic Evaluation of Summaries Using N-gram Co-Occurrences Statistics,” *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, Edmonton, Alberta.

Marcus, R. S., P. Kugel, and A.R. Benenfeld. 1978. Catalog information and text as indicators of relevance. *Journal of the American Society for Information Science*, 29, 15-30.

Mani, I. and Bloedorn, E. 1999. “Summarizing Similarities and Differences Among Related Documents”. *Information Retrieval* 1(1), 35-67.

Mani, I., G. Klein, D. House, and L. Hirschmann. 2002. SUMMAC: A Text Summarization Evaluation. *Natural Language Engineering*. 8(1):43-68.

Monz, C. and M. de Rijke. The University of Amsterdam at CLEF 2001. In *Proceedings of the Cross Language Evaluation Forum Workshop (CLEF 2001)*, 165-169, 2001.

Monz, C. *From Document Retrieval to Question Answering*. PhD Thesis, University of Amsterdam, 2003.

Monz, C. Minimal Span Weighting Retrieval for Question Answering. In Proceedings of the SIGIR 2004 Workshop on Information Retrieval for Question Answering. To appear.

Nenkova, A. and R. Passonneau. Evaluating Content Selection in Summarization: The Pyramid Method. In Proceedings of NAACL-HLT 2004.

Papineni K., S. Roukos, T. Ward, and W. Zhu. 2002. "Bleu: a Method for Automatic Evaluation of Machine Translation," *Proceedings of Association of Computational Linguistics*, Philadelphia, PA.

Passonneau, R. and A. Nenkova. Evaluating Content Selection in Human- or Machine-Generated Summaries: The Pyramid Scoring Method Columbia University, CS Department Technical Report, CUCS-025-03, 2003.

Radev, Dragomir and Kathleen McKeown. 1998. "*Generating Natural Language Summaries from Multiple On-Line Sources," Computational Linguistics*, 24(3), 469-500.

Schwartz, Richard Sreenivasa Sista and Timothy Leek. 2001. "Unsupervised Topic Discovery," *Proceedings of Workshop on Language Modeling and Information Retrieval,* Pittsburgh, PA. http://la.lti.cs.cmu.edu/callan/Workshops/lmir01/WorkshopProcs/Papers/schwartz.pdf

Siegel, S. and N. Castellan. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, 2nd edition, 1988.

Sparck-Jones K., Gallier, J.R. 1996. "Evaluating Natural Language Processing Systems: An Analysis and Review," Springer, Berlin.

Zajic, David, Bonnie J. Dorr and Richard Schwartz and Stacy President. 2004a. "Headline Evaluation Experiment Results," UMIACS-TR-2004-18, University of Maryland Institute for Advanced Computing Studies, College Park, Maryland.

Zajic, David, Bonnie J. Dorr and Richard Schwartz. 2004b. "BBN/UMD at DUC2 2004: Topiary," *Proceedings of Document Understanding Workshop*, Boston, MA.

Zhou, Liang and Eduard Hovy. 2003. "Web-Trained Extraction Summarization System," In *Proceedings of the HLT-NAACL conference*, May.

Appendix I: Topics (Rules of Interpretation)

1. Elections:
Examples - New people in office, new public officials, change in governments or parliaments (in other countries), voter scandals.
The event might be the confirmation of a new person into office, the activity around voting in a particular place and time, the opposing parties' or peoples' campaigns, or the election results. The topic would be the entire process, nominations, campaigns, elections, voting, ceremonies of inauguration.

2. Scandals/Hearings:
Examples - Monica Lewinsky, Kenneth Starr's investigations.
The event could be the investigation, independent counsels assigned to a new case, the discovery of a potential scandal, the subpoena of political figures. The topic would include all pieces of the scandal or the hearing including the allegations or the crime, the hearings, the negotiations with lawyers, the trial (if there is one), and even media coverage.

3. Legal /Criminal Cases:
Examples - crimes, arrests, cases.
The event might be the crime, the arrest, the sentencing, the arraignment, the search for a suspect. The topic is the whole package; crime, investigation, searches, victims, witnesses, trial, counsel, sentencing, punishment and other similarly related things.

4. Natural Disasters:
Examples - tornado, snow and ice storms, floods, droughts, mud-slide, volcanic eruptions.
The event would include causal activity (El Nino, in many cases this year) and direct consequences. The topic would also include; the declaration of a Federal Disaster Area, victims and losses, rebuilding, any predictions that were made, evacuation and relief efforts.

5. Accidents:
Examples - plane- car- train crash, bridge collapse, accidental shootings, boats sinking.
The event would be causal activities and unavoidable consequences like death tolls, injuries, loss of property. The topic includes mourners pursuit of legal action, investigations, issues with responsible parties (like drug and alcohol tests for drivers etc.)

6. Ongoing violence or war:
Examples - terrorism in Algeria, crisis in Iraq, the Israeli/Palestinian conflict.
In these cases the event might be a single act of violence, a series of attacks based on a single issue or a retaliatory act. The topic would expand to include all violence related to the same people, place, issue and time frame. These are the hardest to define, since war is often so complex and multi-layered. Consequences or causes often include (and would therefore be topic relevant) preparations for fighting, technology, weapons, negotiations, casualties, politics, underlying issues.

7. Science and Discovery News:
Examples - John Glenn being sent back into space, archaeological discoveries.
The event is the discovery or the decision or the breakthrough. The topic, then, would include the technology developed to make this event happen, the researchers/scientists involved in the process, the impact on every day life, all history and research that was involved in the discovery.

8. Finances:
Examples - Asian economy, major corporate mergers.
The topic here could include information about job losses, impacts on businesses in other countries, IMF involvement and sometimes bail out, NYSE reactions (heavy trading BECAUSE Tokyo closed incredibly low). Again, anything that can be defined as a CAUSE of the event or a direct consequence of the event are topic-relevant.

9. New Laws :
Examples - Proposed Amendments, new legislation passed.
While the event may be the vote to pass a proposed amendment, or the proposal for new legislation, the topic includes the proposal, the lobbying or campaigning, the votes (either public voting or House or Senate voting etc.), consequences of the new legislation like protesting or court cases testing it's constitutionality.

10. Sports News :
Examples - Olympics, Super Bowl, Figure Skating Championships, Tournaments.
The event is probably a particular competition or game, and the topic includes the training for the game or competition, announcements of (medal) winners or losers, injuries during the game or competition, stories about athletes or teams involved and their preparations and stories about victory celebrations.

11. MISC. News :
Examples - Dr. Spock's Death, Madeleine Albright's trip to Canada, David Satcher's confirmation.
These events are not easily categorized but might trigger many stories about the event. In these cases, keep in mind that we are defining topic as the seminal event and all directly related events and activities. (include here causes and consequences) If the event is the death of someone, the causes (illness) and the consequences (memorial services) will all be on topic. A diplomatic trip topic would include plans made for the trip, results of the trip (a GREAT relationship with Canada??) would be on topic.

Appendix II: University of Maryland Document Relevance Selection Study: Entry 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?
   □ male
   □ female

4. What is your age?_______________________________

5. How often do you use the internet for document searching?
   □ 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?
   □ Google
   □ Ask Jeeves
   □ Yahoo
   □ Other – Please specify _______________________

7. How long have you been doing online searches?_______

Please circle the number closest to your experience

How much experience have you had in:	none		some		lots
10. Using a point-and-click interface	1	2	3	4	5
11. Searching on computerized library catalogs	1	2	3	4	5
12. Searching on commercial on line systems (e.g. BRS Afterdark, Dialog, Lexis-Nexis)	1	2	3	4	5
13. Searching on world wide web search services (e.g. Alta Vista, Google, Excite, Yahoo, HotBot, WebCrawler)	1	2	3	4	5

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
14.How often do you conduct a search on any kind of system	1	2	3	4	5

Please circle the number that indicates to what extend you agree with the following statement:

	strongly disagree	disagree	neutral	agree	strongly agree
15. I enjoy carrying out information searches	1	2	3	4	5

Appendix III: 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.