Contents

List of Figures

List of Tables

1 Introduction

The explosion of on-line textual material and advances in text processing technology have created a renewed interest in text summarization. In May 1998, the U.S. Government conducted an evaluation of automatic text summarization systems, under the TIPSTER Text Program (Phase III). The TIPSTER Text Summarization Evaluation (SUMMAC) represents a significant step in the field of text summarization, as it is the first large-scale, developer-independent evaluation of text summarization systems. This paper is the final report on this evaluation.

The main aim of this report is to present the goals, methods, and results of the evaluation as clearly and concisely as possible. This has meant sacrificing opportunities for much longer discussion and more detailed comparisons with the related literature, which we defer to future papers.

1.1 The TIPSTER Text Summarization Evaluation (SUMMAC)

TIPSTER is a Defense Advanced Research Projects Agency (DARPA) initiative with participation from multiple U.S. Government agencies and research and commercial institutions to push the state of the art in the text processing technologies of information retrieval (IR), information extraction, and text summarization. The TIPSTER program has been in existence since 1992, with the focus being on information retrieval and information extraction. In 1997, DARPA expanded the program to include the development and evaluation of full text summarization systems. This expansion was a natural extension of the existing research as there was a previous history of successful evaluations for both IR (the Text Retrieval Conference [Harman and Voorhees 1996]) and information extraction (the Message Understanding Conference [MUC-6 1995]).

As part of this expanded effort, in October 1997, a "dry-run" of the evaluation was conducted, limited to the six TIPSTER Phase III summarization contractors. These contractors were Carnegie Group Inc. and Carnegie-Mellon University (CGI/CMU), Cornell University and SabIR Research, Inc. (Cornell/SabIR), GE Research and Development (GE), New Mexico State University (NMSU), the University of Pennsylvania (Penn) and the University of Southern California-Information Sciences Institute (ISI). The purpose of this dry-run was to flesh out some of the issues prior to the "formal" evaluation in May 1998.

The formal evaluation was open to everyone, and sixteen systems participated. These systems included the six dry-run participants, as well as ten others: Lexis-Nexis (LN), the University of Surrey (Surrey), IBM Thomas J. Watson Research (IBM), TextWise LLC, SRA International, British Telecommunications (BT), Intelligent Algorithms (IA), the Center for Intelligent Information Retrieval at the University of Massachussetts (UMass), the Center for Information Research (CIR), and the National Taiwan University (NTU). The participants came from four countries: one from Taiwan (NTU), one from Russia (CIR), two from Britain (BT and Surrey), and twelve from the USA.

The goals of the formal evaluation (referred to as "the TIPSTER evaluation") were to judge individual summarization systems in terms of their usefulness in specific summarization tasks, to gain a better understanding of the issues involved in building and evaluating summarization systems, and to guide the direction of the research to requirements of real world tasks. The evaluation is a "black-box" evaluation, where the systems as a whole are evaluated, as opposed to a "glass-box" evaluation, where components internal to each system are evaluated.

1.2 Text Summarization

Text summarization is the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user (or users) and task (or tasks). Since abridgment of content is essential, an important parameter to summarization is the level of compression (ratio of summary length to source length) desired. This form of the compression will be be used throughout this paper, rather than the complement of this measure used sometimes in the literature.

The uses of text summarization vary with different applications. Summaries can be used to indicate what topics are addressed in the source text, and thus can be used to alert the user as to the source content (the *indicative* function). Summaries can also be used to cover the concepts in the source text to the extent possible given the compression requirements for the summary (the *informative* function). Informative summaries are viewed as standing in place of the source. They can even offer a critique of the source (the *evaluative* function) [Sparck-Jones 1998]. Present technical capabilities are limited to indicative and informative summaries. The information in the summary can differ widely across these cases. In some applications, such as abstracting services for technical texts [Pollock and Zamora 1975], it is useful to have summaries which consist of fluent, well-connected text, whereas in other applications where a mere gist is desired, the summaries may be more fragmentary in nature, consisting of words and phrases. Summaries can be tailored to a reader's interests and expertise, yielding *user-focused* summaries, or they can be aimed at a particular - usually broad - readership community, as in the case of *generic* summaries. (It should be borne in mind that the notion of a truly generic summary is problematic, since some background assumptions of an audience are involved in every case.) Traditionally, generic summaries written by authors or professional abstractors served as surrogates for full-text. However, as our computing environments have continued to accommodate full-text searching, browsing, and personalized information filtering, *user-focused* summaries have assumed increasing importance.

Summaries do not, of course, have to consist of *extracts* of source material, although many systems provide that; summaries can include *abstracts*, where the text in the summary isn't necessarily the same as in the source. The summary can stand in place of the source as a surrogate, or it could be linked to the source, or even be presented in the context of the source (e.g., by highlighting source text). Furthermore, a summary could be in a different language from the source, although at present such a capability is little developed. A summary may stand for a single source, or it may stand for a collection of documents [McKeown and Radev 1995], [Mani and Bloedorn 1999]. The source, or sources, need not in fact be text; while text summarization deals with cases where the input is text, this text input in turn may represent information represented in some other media such as imagery, audio, or video. Finally, the actual form of the summary can range from relatively polished, connected text, or text which is more fragmentary (e.g., a list of phrases).

For reasons of space, we will forego an introduction to the framework and methods for text summarization. For a more detailed introduction to the field, see [Mani and Maybury 1999], [Hahn and Mani 1998], [Hovy and Marcu 1998].

1.3 Participant Technology Summaries

Here we very briefly summarize the approaches used by the different participants.

- BT's ProSum used statistical techniques based on the co-occurrences of word stems, the length of sentences and their position in the original text to calculate the importance of a sentence in the context of the overall text in which it occurs. The most important sentences are then used to construct the summary.

- CIR created a thematic representation of a text that included nodes of thematically related terms simulating topics of the text. Related terms were identified using a thesaurus specially constructed for this task.

- CGI/CMU used a technique called "Maximal Marginal Relevance" (MMR), which produces summaries of very long documents by identifying key relevant, non-redundant information found within the document. (This technique is also used for eliminating or clustering redundant information in multi-document summarization applications.)

- Cornell/SabIR used the document ranking and passage retrieval capabilities of the SMART IR engine to effectively identify relevant related passages in a document.

- GE identified the discourse macro structure for each document and selected the passages from each component that scored well using both content and contextual clues.
- IA's infoGIST is a commercial product that uses proprietary algorithms to analyze a document and produce a summary suitable for quick document scanning tasks.
- The IBM approach scored eligible sentences based on word scores, sentence position, paragraph position, and the first mention of salient terms. They used a modified tf.idf calculation to compute the word score.
- ISI used a multi-faceted approach, including the optimal position of a sentence within a text, which varies based on text type, and building thematic representations of texts based on external ontologies.
- LN used a word and phrase-based technique to do statistical sentence weighting and keyword and phrase extraction.
- NMSU used information about the document structure combined with part of speech and proper name recognition to weight and select sentences to be included in their summaries.
- NTU used a multi-step approach to create user-focused summaries. They assigned a part of speech to each word, calculated the extraction strength (ES) for each sentence, partitioned the text into meaningful sentences based on ES, and filtered out the irrelevant sentences to generate the best summary.
- Penn used co-reference resolution as the basis for their summaries, finding information within a document that is naturally linked together by referring to the same individual, organization, or event and extracting that related information to generate a summary.
- SRA extracted summarization features using morphological analysis, name tagging and co-reference resolution. They used a machine learning technique to determine the optimal combination of these features in combination with statistical information from the corpus to identify the best sentences to include in a summary.
- The Surrey system used lexical cohesion analysis to identify relationships within a text that lead to optimal summary creation.
- TextWise primarily concentrated on summaries of multiple documents. They assigned subject field coding of documents as an initial indicator of document content and identified the most relevant paragraphs, combining statistical information about term frequency with linguistic information .
- UMass used a query expansion technique (from the INQUERY IR system) which given a topic and a collection, selects top-ranked documents retrieved from the collection and then adds to the query terms from the context surrounding the topic terms in those documents. They then used a passage retrieval technique (also from INQUERY) to extract one passage per document.

1.4 Evaluation Methods

Text summarization is still an emerging field, and serious questions remain concerning the appropriate methods and types of evaluation. There are a variety of possible bases for comparison of summarization system performance, e.g., summary to source, machine to human-generated, system to system. Methods for evaluating text summarization approaches can be broadly classified into two categories. The first, an *intrinsic* evaluation, judges the quality of the summarization directly based on analysis of the summary. This can involve user judgments of fluency of the summary [Minel et al. 1997], [Brandow et al. 1994], coverage of stipulated "key/essential ideas" [Paice 1990], [Brandow et al. 1994], or similarity to an "ideal" summary, e.g., [Edmundson 1969], [Kupiec et al. 1995]. Measures of fluency can address language complexity (e.g., word and sentence length), presence of dangling anaphors, preservation of structured environments like lists or tables, grammatical and stylistic features, etc. Many of these fluency measures can be detected by automatic means. Judging coverage of key ideas is more subjective, even if humans agree in advance as to what constitutes the set of key ideas.

The problem with matching a system summary against an ideal summary is that the ideal summary is hard to establish. The human summary may be supplied by the author of the article, by a judge asked to construct an abstract, or by a judge asked to extract sentences. There can be a large number of generic and user-focused abstracts that could summarize a given document, just as there can be many ways of describing something. The lack of uniqueness of the ideal abstract extends to extracts as well, as is borne out by reports of low inter-annotator agreement, e.g., [Rath et al. 1961], [Salton et al. 1997] on texts from encyclopedias. There have been several recent studies of sentence extraction, however, that indicate that while judges may not agree on all of the sentences to be included in a summary, they may agree on the most important sentences to include ([Marcu 1997] on expository text; [Jing et al. 1998] on news text).

In contrast to an intrinsic evaluation, the other method of evaluation is an *extrinsic* evaluation, in which the usefulness of a summary is judged based on how it affects the completion of some other task. There have also been a number of extrinsic evaluations, including question-answering and comprehension tasks [Morris et al. 1992], as well as tasks which measure the impact of summarization on determining the relevance of a document to a topic [Mani and Bloedorn 1999], [Jing et al. 1998], [Tombros et al. 1998], [Brandow et al. 1994]. This will be discussed in more detail below.

2 TIPSTER Summarization Tasks

Extrinsic evaluations are especially attractive to a program like TIPSTER, because they are based on objective evaluation measures. Accordingly, the TIPSTER evaluation is mainly an extrinsic evaluation. However, in order to explore relationships between the two kinds of evaluation, it also includes two highly preliminary intrinsic evaluations.

In order to address the goals of the evaluation, two main tasks were defined, based on activities typically carried out by information analysts in the U.S. Government. In the *adhoc task*, the focus is on indicative, user-focused summaries, in particular, where the summaries are tailored to a particular topic. The real-world activity represented by this task is that of an analyst conducting full-text searches using an IR system who must determine quickly and accurately the relevance of a retrieved document. Here the topic is provided as an input to the summarization system, and the evaluation seeks to determine whether the user-focused summary is effective in determining the relevance of the full-text source to a topic. Given a document (which could be a summary or a full-text source - the subject isn't told which), and a topic description, the human subject determines whether the document seen is relevant to the topic. The accuracy of the subject's relevance assessment decision is measured in terms of separately obtained "ground-truth" judgments of the relevance of the full-text source to the topic, which were separately obtained from the Text Retrieval (TREC) [Harman and Voorhees 1996] conferences (which are large-scale evaluations of full-text IR systems).

TREC data was used because of lack of resources needed to create another pool of relevant documents. This of course restricted us to the TREC documents, TREC topics, and their associated boolean relevance judgments. The documents for the TIPSTER evaluation were English documents drawn from TREC CDs 4 and 5, and the topics were all drawn from topics used in TREC. Both documents and topics have SGML annotations, with documents having an identifier as well as a TEXT tag indicating the body of the text.

In the *categorization task*, the focus is on generic summaries. The real-world activity represented by the categorization task is that of a person manually routing or filtering information who must quickly decide whether a document contains information about any of several related topic areas. Here the topic isn't known to the summarization system. Given a document, which could be a generic summary or a full-text source (the subject isn't told which), the human subject chooses a single category (out of five categories, each of which has an

	Mean	Standard Deviation	Range
Adhoc	5032	2231	2018 - 18324
Categorization	5291	2231	2065 - 19012

Table 1: Document Lengths (bytes)

204 - Nuclear power plants	257 - Cigarette consumption
207 - Quebec independence	258 - Computer security
210 - Medical waste dumping	266 - Professional scuba diving
211 - DWI regulations	268 - Cost of national defense
215 - Infant mortality rates	271 - Solar power
219 - Japanese auto imports	273 - Volcanic activity levels
222 - Capital punishment	274 - Electric automobiles
226 - Lotteries	282 - Violent juvenile crimes
254 - Procedures for heart ailments	289 - For-profit hospitals
255 - Environmental protection	297 - Right to die

Table 2: Adhoc Topics

associated topic description) to which the document is relevant, or else chooses "none of the above". Here the evaluation seeks to determine whether the generic summary is effective in capturing whatever information in the document is needed to correctly categorize the document. As in the adhoc task, the accuracy of the subject's relevance assessment decision is measured in terms of TREC "ground-truth" judgments of the relevance of the full-text source to (each of) the topics (documents are chosen so they are relevant to no more than one topic).

A third *question-answering task* involves an intrinsic evaluation which measures the extent to which each summary provides an answer to a set of questions that must be satisfied for a document to be judged relevant to a topic. The correct answers to these questions represent the "obligatory" aspects of a topic. This task is intended to support an information analyst responsible for writing a report on a topic. The challenge for the systems is to understand the topic in relation to each document and to produce an "informative" summary that covers all obligatory aspects of the topic in as short a summary as possible.

Finally, we also conducted an *acceptability task* to measure the overall "acceptability" of the summaries in relation to the full-text, in an intrinsic evaluation similar to [Brandow et al. 1994]. Subjects read the full-text document and the variable-length summaries (the term is introduced below in Section 3.2) of all the systems participating in the categorization task. They were asked to determine if the summary was "good" or not. No criteria for "goodness" were supplied to the subjects (unlike, say, [Brandow et al. 1994]'s study).

The latter two evaluations were highly preliminary in nature. In contrast to the first two evaluations, due to resource limitations, they were not carried out on a large enough scale to permit statistical analyses of some aspects of the results. They were exploratory in nature, and intended to provide some additional intuitions as to what the relationships are between performance in intrinsic and extrinsic evaluations. They are discussed in Sections 8 and 9.

3 Data Selection

3.1 Topics and Full-text Sources

A variety of criteria were used to select topics and documents in each task. In the adhoc task, 20 topics were selected (from TREC topics 200-300). For each topic, a 50-document subset was created from the top 200 most relevant documents retrieved by a standard IR system were chosen. This subset was constructed such that all its documents had relevance judgments provided by TREC, and such that 25%-75% of the documents were relevant

Environment	Global Economy
208 - Bioconversion	251 - Exportation of industry
228 - Pollution recovery	269 - Foreign trade
237 - Alternative Fuels	284 - International drug enforcement
243 - Fossil fuels	290 - Foreign car makers
249 - Rain forests	291 - Worldwide tax sources

Table 3: Categorization Topics

```
<top> <num> Number:  208
<desc> Description:
What are the latest developments in bioconversion --
the conversion of biological waste, garbage, and
plant material into energy, fertilizer, and other
useful products?
</top>

<top> <num> Number: 254
<title> Topic: Non-invasive procedures for persons with heart ailments
<desc> Description:
The document will discuss all those cases in which medications and
procedures were used instead of or prior to heart surgery.
<narr> Narrative:
A relevant document will report/discuss those cases in which
persons diagnosed with heart ailments were treated with
medications and/or techniques such as angioplasty, stents, lasers,
arthrectomy (roto router) etc., in place of surgery.  Also advantages
of non-invasive procedures over surgery and comparative studies
which show any disparity in longevity when either procedure is used.
</top>
```

Figure 1: Topic Examples

to the topic. The resulting 20 subsets (i.e., one subset per topic) had no documents in common. Documents shorter than 2000 bytes were excluded under the assumption that summarization might not be useful for such short documents. Documents longer than 20,000 bytes were excluded because reading texts of that length might be infeasible. The distribution of document lengths is shown in Table 1[1].

For the categorization task, the main difference from the adhoc task in terms of data preparation was in the selection of topics: only 10 topics were selected, with 100 documents used per topic. These topics were selected such that they could be grouped into two mutually exclusive classes: *environment* and *global economy*. In the dry-run evaluation, it turned out that the topics used in the categorization task were so distinct that very few instances were assigned to an incorrect category other than "none of the above". Therefore, in the formal evaluation, topics within a group were inspected to ensure that they were "similar", based on subjective judgments of similarity, as well as the existence of relevant documents in common across pairs of topics within a group (the overlapping documents were then excluded from the test set).

The topics chosen are shown in Tables 2 and 3. Each topic includes a description and (for topics 251-300) narrative fields. Two sample topics are shown in Figure 1.

The breakdown of different text sources in the two main tasks is shown in Table 4. The documents are all newspaper sources, the vast majority of which are news stories, but also containing sundry material such as letters to the editor. While we recognize that a more diverse set of text genres would be more interesting for

[1] The standard deviation for adhoc and categorization lengths are in fact the same.

Source	Adhoc	Categorization
Associated Press	47	49
Wall Street Journal	27	22
Financial Times	19	21
San Jose Mercury	7	8

Table 4: Percentage Documents By Source

summarization, we had to limit ourselves to TREC data that met our requirements.

3.2 Summaries

In each task, participants submitted two summaries: a fixed-length summary limited to 10% of the character length of the source, and a "best" summary which was not limited in length (also called a variable-length summary). As with the source documents, summaries were marked up in SGML and had to conform to SGML specifications (DTDs). The 10% limit was based on number of characters within the TEXT tag in the summary and source, excluding whitespace. Fixed-length summaries longer than the limit were automatically truncated, with completion allowed to the end of a word.

Although sophisticated presentation and interaction strategies can make a substantial difference in the effectiveness of summarization, we decided not to focus on user interface issues in this evaluation. This strategy has the advantages of avoiding significant complications to the evaluation, as well as leveling the playing field across participants. Each summary was required to be an ASCII submission constituting a summary of just the TEXT portion of the source. The summary could be presented in a format of the participant's choosing, within the range of readable ASCII text. No additional formatting was allowed (e.g. highlighting, bold-facing, underlining). Also, any "headline" information (including tags like "HL", "1ST_LINE", "TITLE", etc., from the original document) found within the summaries was stripped out, along with any text within those tags.

To help make this more concrete, the Appendix shows an example document, topic, and some sample summaries.

3.3 Training

Several months prior to the evaluation, sample topics and documents were constructed along the lines above, and supplied as training data to the participants. These topics and documents did not overlap with those used in the evaluation. The training data was intended for participants to become familiar with the tasks and data, for them to submit sample summaries which were then validated, and to allow their systems to improve performance by using corpus-based methods. To help ensure that everyone was counting summary length for the fixed-length summaries the same way, a character-length counting program was supplied to the participants.

4 Experimental Hypotheses and Design

In meeting our evaluation goals, the main question to be answered was whether summarization saved time in relevance assessment, without impairing accuracy.

The first test was a *summarization condition test*: we wanted to determine whether subjects' relevance assessment performance in terms of time taken and accuracy was affected by different summarization conditions: full-text (F), fixed-length summaries (S1), variable-length summaries (S2), and baseline summaries (B). The latter were generated by picking the first 10% of the full-source text (i.e., just the material within the TEXT portion). The use of leading text as a baseline is motivated by an intrinsic evaluation by [Brandow et al. 1994], who found leading text summaries of newswire texts were as acceptable as summaries generated using sophisticated techniques.

The second test was a *participant technology test*: we were also interested in comparing the performance of different participants' systems.

The third test was a *consistency test*: to determine how much agreement there was between subjects on their relevance decisions on full-text sources, and to compare that agreement with the TREC ground-truth. In order to address the consistency test, we had, in addition to the main adhoc and categorization tasks, "consistency" versions of these tasks, where the subjects were just shown full-text sources. The subjects' relevance decisions from the consistency task could also be used as an alternative yardstick of ground-truth. The differences obtained from applying this yardstick rather than TREC relevance judgments is described in Section 7.1.5.

To arrive at definitive conclusions regarding these comparisons, we assumed a statistical methodology based on significance testing. This methodology is as follows: In each of the two tests above, we assume a null hypothesis. In the first test, the null hypothesis is that there is no difference in performance of subjects between different summarization conditions. In the second test, the null hypothesis is that there is no difference in performance of subjects for different participant summaries. We apply a statistical test to assess the difference in performance between groups (e.g., summarization conditions or participants), with the result from the test being subject to two parameters, α and β. The significance level α is the probability of rejecting the null hypothesis (i.e., in our case the hypothesis that there's no difference) when it's true, in other words, the probability of inferring a difference when there isn't one. Conventionally, $\alpha \leq 0.05$; this value is used in the significance results reported in this paper. The parameter β is the probability of failing to reject a false null hypothesis, in other words, the probability of missing a significant difference. β $(= (1 - Power))$ can be calculated from power tables (e.g., [Cohen 1969], [Kirk 1968]); given a statistical test, a sample size, and a significance level α, these tables compute the power to detect a small, medium and large effect size (i.e., the size of the difference between summarization conditions). Conventions for "small", "medium", and "large" are arrived at from the experimental literature [Cohen 1969].

In designing our experiment, we assumed we would have in the main tasks at least 16 subjects, each assessing 1000 documents. Each of these 1000 documents would be divided between full-text (F), fixed-length summaries (S1), variable-length summaries (S2), and baseline summaries (B). The dry-run evaluation conducted on the adhoc and categorization tasks used fewer subjects and participants. The dry-run indicated that we could expect very large effect sizes for the summarization condition and participant technology tests. For the formal evaluation, given effect sizes as large as those in the dry-run, and given 1000 documents to be analyzed per subject, a power analysis indicated that α and β could be kept acceptably small for the summarization condition test with as few as 20 full-text versions and 20 baseline summaries, leaving 480 documents to be allocated to S1 and 480 to S2. These 960 S1 and S2 documents not only provided for acceptably small α and β for the participant technology test between participants' systems, but also allowed us to sample as many of the participants' summaries as possible, which enabled us to offer detailed feedback to each participant. (In the formal evaluation, there were 11 participants for the adhoc task and 14 for the categorization task.)

In the adhoc and categorization tasks, the 1000 documents assigned to a subject were allocated among F, B, S1, and S2 conditions through random selection without replacement: 20 were randomly selected from the 1000 and allocated to F, then 20 were randomly selected from the remaining 980 and allocated to B, 480 were randomly selected from the remaining 960 and allocated to S1 and the remaining 480 were allocated to S2. The presentation order of the documents was varied between subjects.

For the consistency tasks, each subject was assigned full-text versions of the same 1000 documents, and the presentation order was varied between subjects.

5 Experimental Method

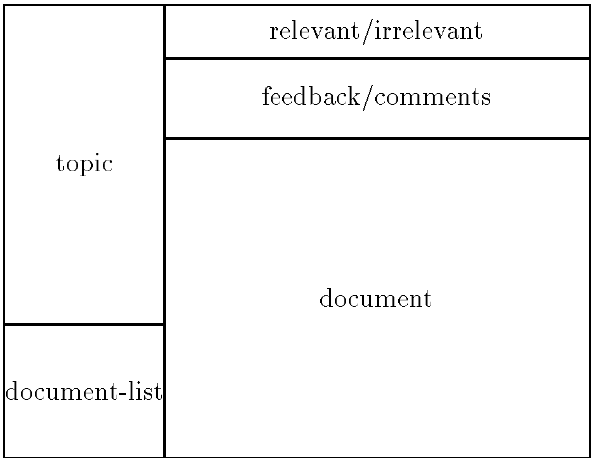


Figure 2: Schematic for ASSESS User Interface (adhoc version)

Question	Answer Type
Preferred length	1=Shorter, 2=Just Right, 3=Longer
Intelligibility	1=Low, 2=OK, 3=Fine
Confidence in decision	1=Low, 2=Medium, 3=High
Comments	Free text
Was the category you chose the main theme of the text? (categorization task only)	Yes/No

Table 5: Questions in Feedback Form

The experimental subjects were 51 professional information analysts. The evaluation was conducted over a one month period, approximately 16-20 hours per subject. The main adhoc task had 21 subjects, the main categorization 24 subjects; the consistency adhoc task had 14 subjects, the consistency categorization 7 subjects. The 14 subjects in the consistency adhoc task included 10 subjects who had completed the main categorization task, and the 7 subjects in the consistency categorization task included 5 subjects who had completed the main adhoc task. The subjects were told they were working with documents that included summaries, and that their goal was to examine each document to determine if it was relevant to the topic.

The subjects were presented with the topics and documents using the ASSESS user interface from the National Institute for Standards and Technology (NIST) used in the TREC relevance assessments, but modified by MITRE for the needs of this evaluation. The modified ASSESS interface for the adhoc tasks is shown schematically in Figure 2. The left top pane of the interface contains a topic description, the left bottom pane shows a list of documents by title, the right bottom pane shows the document, and the right top pane has radio buttons to select relevant versus non-relevant, along with a feedback form that pops up when relevance is assessed. The system automatically cycles the topic-document pairs to the subject in a pre-determined order, with the next document for a topic being brought up once the feedback form is filled out, and with the next topic being brought up when the previous topic's documents have all been assessed. Each decision the subject makes is logged along with a time-stamp; the subject also has a 'take-a-break' button, which can be used at any time. The interface for the categorization tasks is similar, except that six radio buttons are used with short labels for categories, instead of two radio buttons. The feedback form shown to the subjects after each decision contained a number of questions with multiple-choice and yes/no answers, along with a comment field for free text. These are shown in Table 5.

Before the evaluation began, the subjects were walked through the instructions, and were provided with

Ground Truth	Subjects' Judgment	
	Relevant	**Irrelevant**
Relevant is True	TP	FN
Irrelevant is True	FP	TN

Table 6: Adhoc Task Contingency Table. TP=true positive, FP = false positive, TN= true negative, FN=false negative.

Ground Truth	Subjects' Judgment		
	X	**Y**	**None**
X is True	TP	FN	FN
None is True	FP	FP	TN

Table 7: Categorization Task Contingency Table. X and Y are distinct categories other than None-of-the-above, represented as None.

warm-up examples of topics and documents, different from those used in the actual evaluation. Subjects were told that their decisions would be timed (although they should not attempt to 'beat the clock'), and that they were to take breaks only when in feedback mode.

6 Performance Metrics

There are two main measures of performance, time and accuracy. The time of each individual's decision can be measured straightforwardly (from the log files), and is reported in seconds. Measuring accuracy is more complicated, and is explained below.

6.1 Basic Measures of Accuracy

The accuracy decisions made by subjects in the adhoc and categorization tasks are categorial decisions, for which a contingency table analysis is appropriate. Table 6 shows such a table for the adhoc task. Given a topic and a document, a subject's decision can be placed in exactly one cell in this table. The four cells form a partitioning of the set of decision categories, in that they are exhaustive and mutually exclusive. As mentioned earlier, the ground truth judgments are obtained from the TREC data. In these relevance assessment tasks, an indicative summary is "accurate" if it accurately reflects the relevance or irrelevance of the corresponding source. Therefore, when the subject is looking at a summary and makes a judgment of relevant/irrelevant, the correctness of this judgment is with respect to the corresponding full-text source.

The above contingency table analysis is ideally suited to problems of binary classification as in the adhoc task. For the categorization task, where one is assigning a document to exactly one of six categories, there are a number of different ways of measuring accuracy. Among them are treating each of the categories as a separate binary classification task and then averaging across classes, or collapsing all the outcomes to two possibilities, correct or incorrect. We preferred the latter; although it loses some information, it codes the results the same way as the adhoc task, allowing for better comparisons, and does not hide what is going on. The semantics of TP, TN, etc. for this six-way classification is indicated in Table 7.

In keeping with the analysis of accuracy in terms of categorial decisions, differences between groups can be compared for significance in terms of a non-parametric test such as chi-squared:

$$\chi^2 = \sum \frac{(f - e)^2}{e} \tag{1}$$

where the summarization is across all decision categories and all conditions being tested. In equation (1), f is the observed frequency within a particular decision category for a particular condition (e.g., F, B, S1 or S2). When testing for a null hypothesis that the distributions are the same across decision categories for each summarization condition (F, B, S1 and S2) and/or for each technology (across participants), the particular expected frequency e contrasted with each observed f is based on the overall average frequency for that decision category. Thus, in comparing conditions (e.g., rows F, B, S1 or S2), we sum up the scores in each category (e.g., columns TP, TN, etc.), and for each row, we make the expected value for each column cell to be that fraction of the row total given by the percentage that the column's total is of the grand total of the columns.

Based on the chi-squared measure, we can determine the size of the differences between groups using the effect size measure introduced earlier. The effect size e [Cohen 1969] is defined as:

$$e = \frac{\chi^2}{n} \tag{2}$$

where n is the sample size.

Conventions for "small", "medium", and "large" are arrived at from the experimental literature. [Cohen 1969] suggests the criterion for small is < 0.05, for large > 0.10.

6.2 Aggregate Measures of Accuracy

In addition to the basic measures above, it is possible to import into the summarization evaluation other accuracy metrics commonly used in the IR literature. These are aggregate measures, computed for a set of decisions, unlike TP, etc., which are assigned to each individual decision. Further, these measures involve division, and so are undefined when the denominator is zero for the set of decisions.

$$Precision = TP/(TP + FP) \tag{3}$$

$$Recall = TP/(TP + FN) \tag{4}$$

$$Fallout = FP/(FP + TN) \tag{5}$$

$$Fscore = 2 * Precision * Recall/(Precision + Recall) \tag{6}$$

In contrast to the basic measures of accuracy above, differences between groups on these aggregate measures are compared by means of an analysis of variance. An F-ratio value (not to be confused with F-scores) provides an overall test of differences based on a conventional ratio between the variance among the group means (i.e., the between-group mean square[2]) and the general population variance (estimated from the aggregated variances among scores within groups, i.e., the within-group means square).

When the F-ratio indicates that overall significant differences exist at an α-level criterion, then specific differences between groups are tested using Tukey's Studentized Range criterion, called the Honestly Significant Difference (HSD). When exploring data for effects, as in this study, where many comparisons are being made, HSD ensures that inferring erroneously that even one difference is true (when it truely isn't) has a probability of only α.

We also mention in passing a statistic we will use often in the paper, to compare aggregate measures of accuracy. This is Pearson's r, the product-moment correlation coefficient (e.g., see [Press et al. 1988]). It is

[2] The mean square is the sum of squares of deviations from the mean divided by the degrees of freedom.

used to compare two variables X and Y whose values $x_1 \ldots x_n$ and $y_1 \ldots y_n$ are ordinal, and is defined as follows (here $\overline{x}$ is the mean of the X values):

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{7}$$

r ranges from -1 (perfect negative correlation) to +1 (perfect positive correlation), with a value near zero indicating that the variables are uncorrelated. It should be noted that this compares scores rather than comparing ranks. r gets a higher absolute value when each positive (negative) deviation score on X gets multiplied by a similarly large positive (negative) deviation score on Y.

We now discuss some issues associated with using the above IR measures to measure summarization accuracy. In an IR evaluation, precision measures the percentage of relevant documents in a set retrieved as a result of a query. Clearly, this measures the preciseness of the system. In the summarization evaluation, however, precision measures the system's ability to summarize relevant information from relevant documents (thus increasing the likelihood of a TP) while not summarizing information that appears to be relevant from nonrelevant documents (which would otherwise increase the likelihood of a FP).

Also, in an IR evaluation, recall measures how many relevant documents of the total relevant document set a system is able to retrieve. This is clearly an important measure for IR systems which, in combination with precision, gives a workable measure of a system's effectiveness. For summarization, recall measures how well the system is able to capture the relevant content of a document. It measures a system's ability to extract relevant information from documents known to contain relevant information. Consider, however, the case of a long document that contains only a small text portion relevant to a given topic. Should the document be considered relevant to the topic? If the document should be considered topic relevant, should the generic summary (as produced for the categorization task) include the relevant section? If it does not include the relevant section, then the produced summary will be judged as nonrelevant, resulting in a FN, adversely affecting recall. If the document should be considered not relevant to the topic, should the user-focused summary (as produced for the adhoc task) include the relevant section? If it does include it, then the produced summary will be judged as relevant, resulting in a FP based on TREC criteria of ground truth, adversely affecting precision.

The generation of a relevant summary from a TREC-determined nonrelevant document, which produces a false positive result, may be considered good behavior in a user-focused summarizer. There are several reasons why a false positive might occur: 1) The TREC assessor might have made an initial error in judgment, 2) The summarization assessor might disagree with the TREC assessor, or 3) The summary might have captured the relevant part of a document that causes the IR system to retrieve it in the first place. It is this last possibility that makes the false positives interesting. If the TREC topic is not the main subject of the document, this relevant section might have been obscured so that the article appeared to be nonrelevant. If a user-focused summarizer is able to recover these relevant sections, the use of such systems can be very advantageous, particularly with longer documents.

In general, we would prefer summaries which are accurate and which take less time to process than full-text. There are various ways of looking at the trade-off between accuracy and time; we have resorted here to plotting accuracy against time, where the best solution is found in the high-accuracy, low-time region of the plot.

7 Results: adhoc and categorization tasks

7.1 Adhoc tasks

7.1.1 Performance by Condition

The performance in terms of F-score and Time for the different conditions (averaged across subjects) is shown in Tables 8 and 9. The striking result is that variable-length summaries are not significantly different in accuracy

Condition	F-score
F	.67
S2	.64
S1	.53
B	.42

Table 8: Adhoc F-score by Condition. All F-score differences are significant, except between F (Full-Text) and S2 (HSD=0.10, F-ratio = 18.5).

Condition	Mean Time	Time Standard Deviation
F	58.89	56.86
S2	33.12	36.19
B	23.15	21.82
S1	19.75	26.96

Table 9: Adhoc Time by Condition (seconds per decision). All time differences are significant except between B and S1 (HSD=9.8, F-ratio=420.81).

from full-text. As might be expected, the summaries as a whole save time compared to full-text. The fixed-length summaries are not significantly quicker to process than the baselines. All other differences in accuracy and time are significant at the $\alpha = 0.05$ level.

When we examine other accuracy measures, as in Table 10, several interesting patterns emerge. First, when we drill down to Precison and Recall measures, the ordering of $F \simeq S2 > S1 > B$ on F-score accuracy continues to hold for Recall, while Precision differences aren't significant. As can be seen, performance in terms of recall is much poorer than performance in terms of precision: the main source of error in all conditions is FNs, not FPs. The proportion of FNs is significantly higher than that of the FPs across conditions ($\chi^2 = 69.18$). This means that subjects are not recognizing relevant documents. Further, the FPs are not significantly larger a source of error in summaries than they are in full-text. In other words, the main weakness with these summaries compared to full-text is that they are more often missing relevant information from the source (resulting in higher FNs). This suggests that within the available compression, improved tailoring of the summary to the topic is required.

Figure 3 shows the time taken as a function of summary length. Each point on the graph represents the set of variable-length summaries at that summary length range (sets of size less than 20 aren't plotted). It can be seen that the time taken steadily increases with summary length in a more or less linear manner until about 1000 bytes, after which there is less linearity and fewer data points.

7.1.2 Performance by Participant

Condition	TP	FP	FN	TN	Precision	Recall	Fallout
F	.38	.08	.26	.28	.83	.59	.22
S2	.35	.08	.28	.28	.80	.56	.23
S1	.27	.07	.35	.31	.79	.43	.19
B	.18	.05	.41	.35	.81	.30	.12

Table 10: Adhoc: Other Accuracy Measures by Condition. Here TP, FP, FN, TN are expressed as percentage of totals observed in all four categories. Precision differences aren't significant. All Recall differences between conditions are significant, except between F and S2 (HSD=0.12, F-ratio=16.40). Fallout differences between S2 and B only are significant (HSD=0.10, F-ratio=2.98).

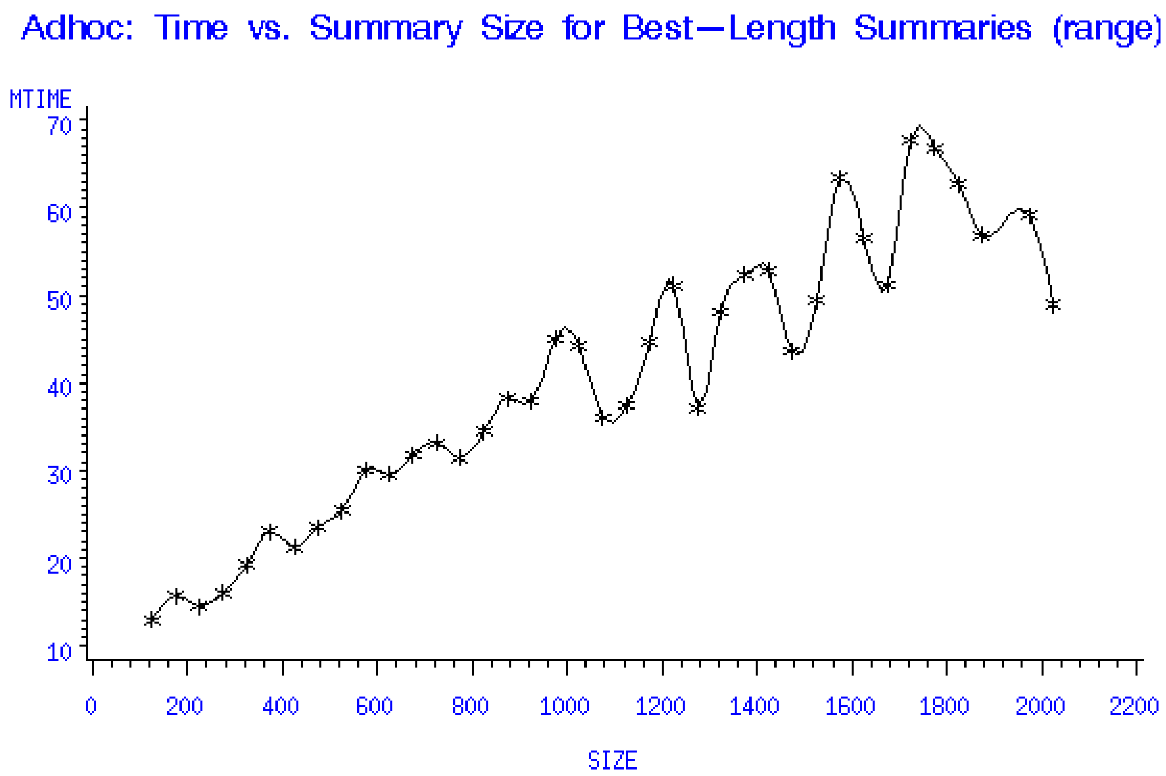


Figure 3: Adhoc Time versus Summary Length in bytes (variable-length summaries).

Party	Precision	Recall	Fallout	F-score
CGI/CMU	.82	.66	.25	.72
Cornell/SabIR	.78	.67	.32	.70
GE	.78	.60	.27	.67
LN	.78	.58	.25	.65
Penn	.81	.57	.21	.65
UMass	.80	.54	.23	.63
NMSU	.80	.54	.22	.63
TextWise	.81	.51	.18	.61
SRA	.82	.49	.19	.60
NTU	.80	.49	.22	.59
ISI	.80	.46	.19	.56

Table 11: Adhoc Accuracy (variable-length) by Participant. Precision and Fallout differences aren't significant. CGI/CMU and Cornell/SabIR are significantly different from SRA, NTU, and ISI in Recall (HSD=0.17, F-ratio=3.41) and from ISI in F-score (HSD=0.13, F-ratio=2.67).

Party	Precision	Recall	Fallout	F-score
CGI/CMU	.76	.52	.26	.60
Cornell/SabIR	.79	.47	.20	.56
UMass	.81	.47	.21	.56
GE	.77	.45	.22	.55
LN	.81	.45	.16	.55
Penn	.76	.45	19	.53
TextWise	.79	.41	.19	.52
NMSU	.80	.40	.18	.52
SRA	.79	.37	.14	.48
ISI	.82	.36	.16	.47
NTU	.82	.34	.12	.46

Table 12: Adhoc Accuracy (fixed-length) by Participant. No significant differences on any of the measures.

Group	Members
Group I	CGI/CMU, Cornell/SabIR
Group II	GE, LN, NMSU, NTU, Penn, SRA, TextWise, UMass
Group III	ISI

Table 13: Adhoc Accuracy: Participant Groups for Variable-length summaries. Groups I and III are significantly different in F-score.

The performance of participants on their variable-length summaries are shown in Table 11. The systems (shown in order of decreasing F-score) are all very close in accuracy. A similar result in seen for the fixed-length summaries shown in Table 12. However, three groups of systems are evident in the variable-length adhoc F-score accuracy data, as shown in Table 13. Group I is significantly more accurate in terms of F-score than Group III; accuracy differences within groups and between Group II and the others are not significant. The effect size (e.g., F-ratio=2.67) is very small, however.

Figures 4 and 5 show the adhoc accuracy (F-score) versus time for different participants. For variable-length summaries, the Group I systems are not the fastest in terms of time. As can be seen, for variable length summaries (Figure 4), TextWise, GE and Penn are the closest in terms of Cartesian distance from the ideal performance in terms of both accuracy and time. These three systems are equivalent (Group II) in accuracy. For fixed-length summaries (Figure 5), LN, UMass, and CGI/CMU are the closest in terms of Cartesian distance from the ideal, but the accuracy and time differences between these and the rest aren't significant.

7.1.3 Effect of Compression Rate

Figure 6 shows the number of summaries at different compression rates. (Recall that the compression is defined as the number of characters in the summary divided by number of characters in the source.) It can be seen that the histogram peaks at 17-19% compression. The mean and standard deviation of the compression for each participant are shown in Table 14. More than three-fourths of the (variable-length) summaries have between 17%-30% mean compression. The systems which we identified in Figure 4 as being relatively closer to the ideal in terms of accuracy and time (TextWise, GE, and Penn) have mean compressions at the lower end of the range (less than 20%).

We now turn to the impact of compression on F-score. This is shown in Figure 7, where each point on the graph represents the set of variable-length summaries at that compression. The curve shows a logarithmic relation between F-score and compression. As the compression increases, the summary length becomes a larger fraction of source length. One might expect the F-score to increase as the compression rate increases, as the summary has a greater chance of including topic-related information from the source; however, this must be

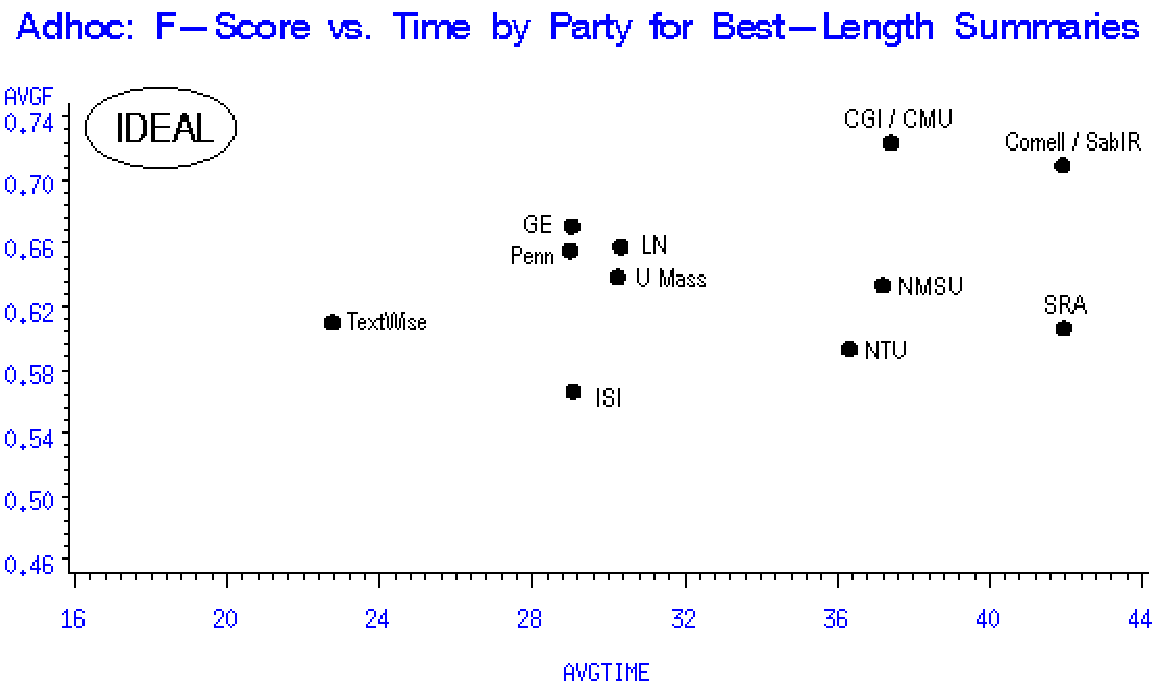


Figure 4: Adhoc F-score versus Time by Participant (variable-length summaries). HSD(F-score) is 0.13. HSD(Time) = 12.88, F-ratio(Time)= 4.78. GE, Penn, and TextWise are significantly faster than SRA and Cornell/SabIR.

Participant	**Mean Compression**	**Compression Standard Deviation**
Cornell/SabIR	29.64	7.75
SRA	28.91	6.47
NMSU	27.41	2.35
CGI/CMU	25.42	4.08
NTU	25.04	5.07
LN	20.42	7.10
Penn	20.18	9.22
UMass	18.76	2.57
GE	17.43	4.77
ISI	17.09	2.46
TextWise	14.39	5.67

Table 14: Adhoc: Compression by Participant for variable-length summaries.

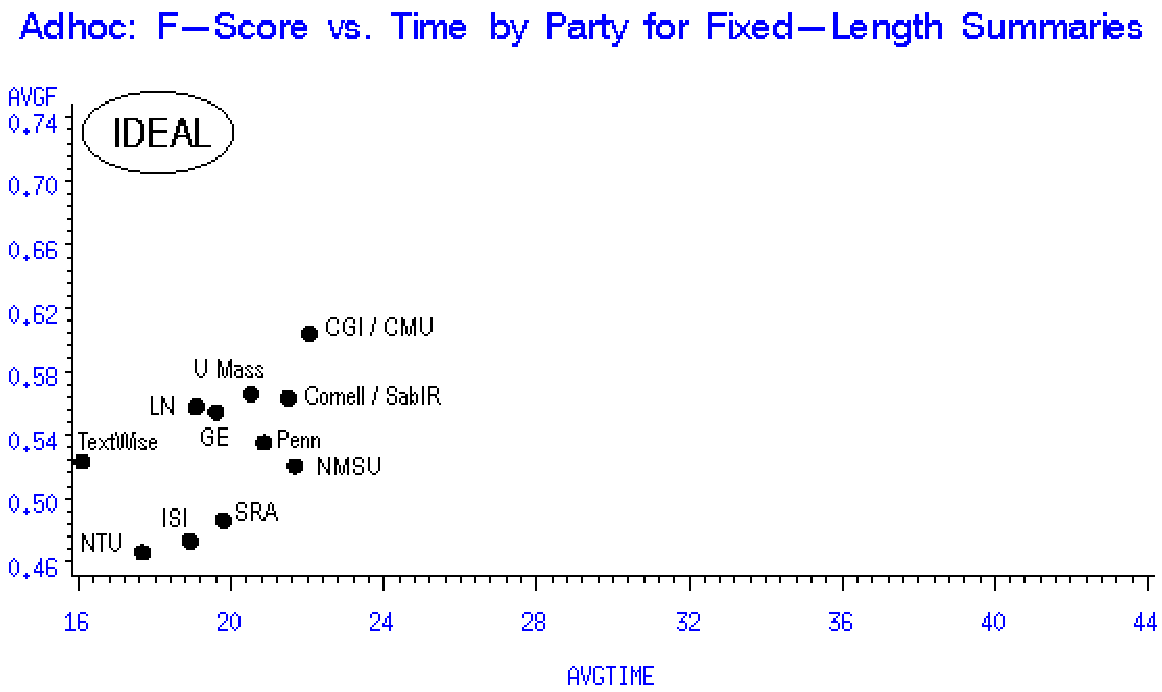


Figure 5: Adhoc F-score versus Time by Participant (fixed-length summaries). No significant differences in F-score, or in Time.

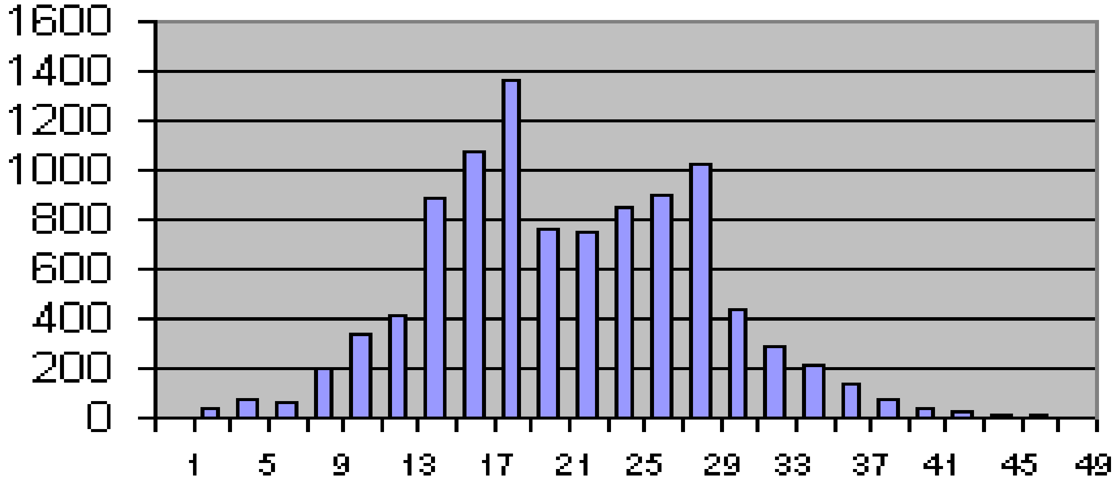


Figure 6: Adhoc: Compression Rate Histogram (variable-length summaries).

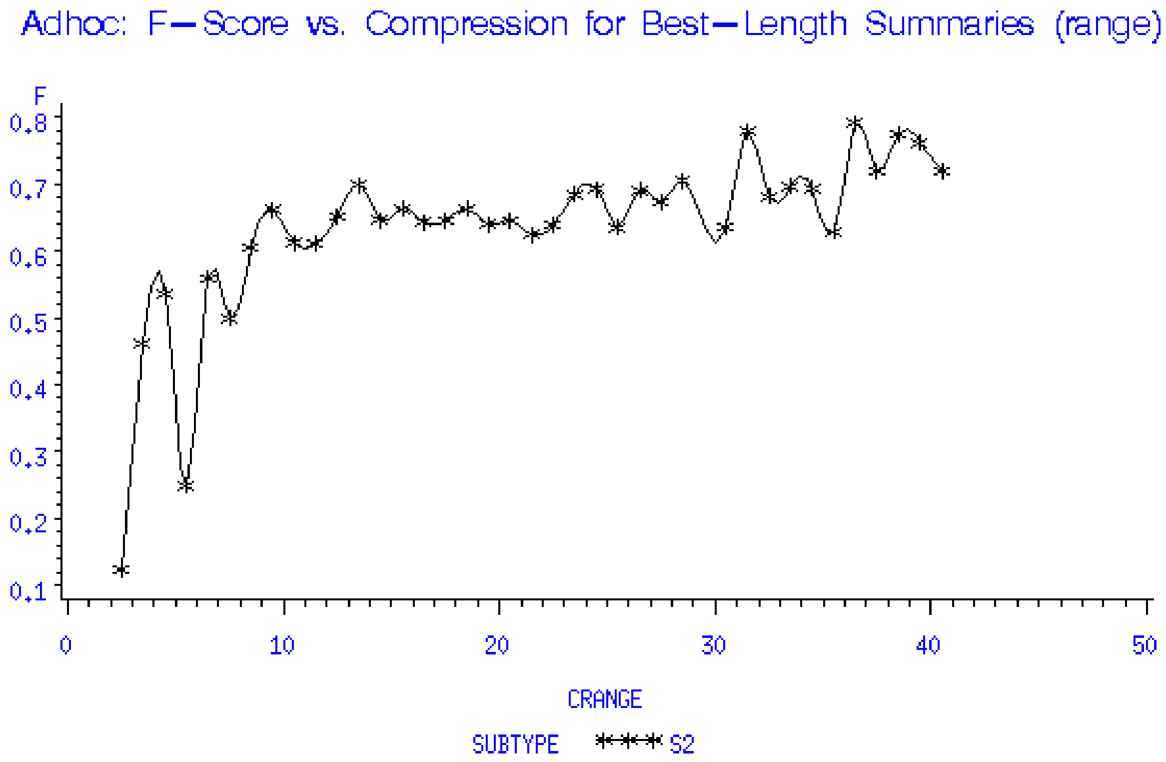


Figure 7: Adhoc F-score versus Compression (variable-length summaries).

balanced against the increased chance of including irrelevant information as well. As can be seen from Figure 7, for summaries less than 10% of the full source length, there is a sharp increase in the F-score (.1-.6) as the summaries become a longer fraction of the full-source length. However, between 10%-40% compression, there is a lesser increase in F-score (.6 to .8). In other words, relatively low compression rates (e.g., 17%) are already achieving relatively high performance. Similar trends are reflected in the compression curves for the individual participants shown in Figure 8. We note that the Group I systems are the most accurate at all except one compression band.

7.1.4 Agreement between Subjects

The agreement between the 14 adhoc subjects from the "consistency" adhoc task on full-text is shown in Table 15. Of the 1000 decisions each subject made, the subjects were unanimous about the outcome of only

Percent subjects agreeing	Agree Relevant	Agree Irrelevant	Total	Running Total
100	52	114	166	166
90-99	63	89	152	318
80-89	72	95	167	485
70-79	107	94	201	686
60-69	73	73	146	832
50-59	94	74	168	1000

Table 15: Agreement between Subjects for Adhoc Task

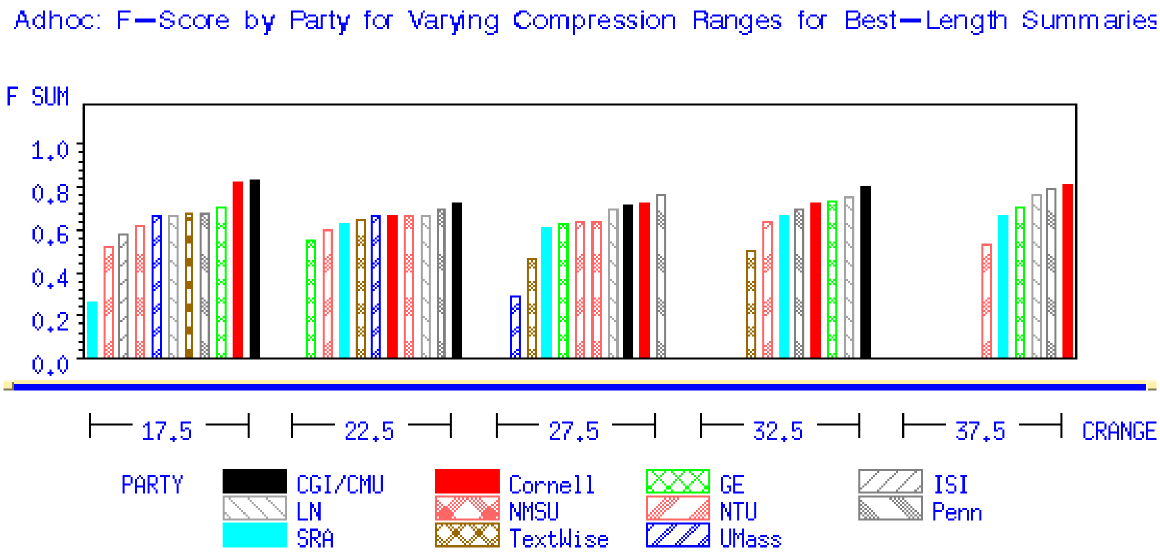


Figure 8: Adhoc F-score versus Compression by Participant (variable-length summaries).

166 decisions, indicating 16.6% agreement on their decisions. The average pairwise agreement was 69.1%.

There are various ways of testing whether the disagreement is significant, including Cohen's effect size [Cohen 1969] discussed earlier and the Kappa measure [Siegel 1988].

The effect size e of the disagreement in the categorial decisions among subjects is 0.07, which is considered to be a "medium" effect size (here the criterion for small is < 0.05, for large > 0.10). This means that there is a moderate amount of disagreement among subjects.

Next we compute the Kappa measure. Kappa is defined as:

$$K = \frac{P(A) - P(E)}{1 - P(E)} \tag{8}$$

where $P(A)$ is the proportion of times the judges agree and $P(E)$ is the proportion of time we expect the judges to agree, corrected for chance agreement. $K = 1$ if there is complete agreement among subjects, and $K = 0$ if there is no agreement other than what is expected by chance.

To compute K (see [Siegel 1988], pp. 284-291 for details), we construct a table with a row for each document, and two columns, one indicating how many judges marked the document relevant, and the second indicating how many marked it irrelevant. The proportion of documents assigned to the jth category (relevant/irrelevant) is $p_j = C_j/(N * k)$ where C_j is the total for column j, N is the number of documents and k is the number of judges. The expected proportion of agreement on category j is p_j^2, assuming subjects assign documents to categories at random. $P(E)$ is thus the sum of the expected proportion of agreement on relevant and the expected agreement on irrelevant.

The Kappa value for agreement in this task is 0.38. According to [Carletta et al. 1997], based on conventions derived from the literature, K must be at least .8 for good replicability, with $0.67 < K < 0.8$ "allowing tentative conclusions to be drawn" (although they do point out that in the medical literature, $.21 < K < .40$ is considered "fair" agreement).

The unanimous agreement on just 16.6% of the decisions (with 14 subjects) is lower than reported in experiments in TREC [Harman and Voorhees 1996], where all of 3 subjects had unanimous agreement on 71.7% of the decisions. It is also lower than found in the dry-run adhoc task, where all of 3 subjects had unanimous agreement of 59.1% and average pairwise agreement of 72.73%. The lower agreement in the current study

Condition	TREC	60% TIPSTER
F	.67	.76
S2	.64	.75
S1	.53	.73
B	.42	.54

Table 16: Adhoc Accuracy by Condition under Different Ground Truth Schemes

Condition	F-score
F	0.50
S2	0.50
S1	0.43
B	0.03

Table 17: Categorization F-score by Condition. Only the F-score of B is significantly less than the others (HSD=0.09, F-ratio=74.64).

compared to the dry-run and TREC might indicate that the number of subjects (3) in those previous studies may be too small to provide a representative analysis.

7.1.5 Revising Ground Truth

Given the lack of good agreement among subjects compared to TREC, it becomes difficult to use relevance judgments derived from the consistency task in place of TREC relevance judgments. Nevertheless, we investigated the use of a "majority agreement" criterion based (arbitrarily) on 60% agreement between subjects. In other words, we consider just the 832 full-text documents for which there was at least 60% agreement, and assign ground truth of "relevant" to a query given one of those documents if at least 60% of the consistency adhoc subjects say it's relevant to that query, and assign ground truth of "irrelevant" if at least 60% say it's irrelevant. The disadvantage of course, is that we are throwing away "minority" judgments, lessening the amount of valuable data we have. Moreover, decisions which are hard to agree on may be those where summarization could help, and therefore these should perhaps not be excluded.

With these caveats in mind, based on this 60% scheme, the F-scores for different conditions were recomputed. This resulted in boosting of the F-scores for different conditions. Since we limited ourselves to documents which are "obviously" relevant, we would expect more TPs and less FNs, thereby raising the F-score. Although the basic ordering of conditions was relatively unchanged, the differences between full-text and non-baseline summaries became blurred. This is seen in Table 16. In short, the low levels of agreement we have do not yield a useful substitute for TREC judgments.

7.2 Categorization tasks

7.2.1 Performance by Condition

The performance in terms of F-score and Time for the different categorization conditions, averaged across subjects, is shown in Tables 17 and 18. The S2 and S1 summaries are not significantly different in accuracy from full-text. The very low F-scores for the baselines can be explained by a bug discovered in the randomization step in the assignment of baselines to participants which resulted in the same 20 baselines being offered to each subject. However, the variable-length summaries do not significantly differ in time from the full-text. All other time differences except between fixed-length and baselines are also significant at the $\alpha = 0.05$ level.

In the categorization task, for the other accuracy measures (Table 19), we note again that performance in terms of recall is much poorer (excluding the problematic baseline) than performance in terms of precision: the

Condition	Time Mean	Time Standard Deviation
S2	43.15	42.16
F	43.11	52.84
B	27.36	30.35
S1	25.48	29.81

Table 18: Categorization Time by Condition. All time differences are significant except between F and S2, and between B and S1 (HSD=15.6, F-ratio=456.82).

Condition	TP	FP	FN	TN	Precision	Recall	Fallout
F	24.3	13.3	28.5	33.9	0.63	0.45	0.31
S2	19.3	10.5	36.9	33.3	.68	.42	.31
S1	27.1	10.7	30.9	31.3	.68	.34	.23
B	7.5	11.9	52.5	28.1	0.04	0.02	0.12

Table 19: Categorization: Other Accuracy Measures by Condition. Here TP, FP, FN, TN are expressed as percentage of totals in all four categories. The Precision and Recall of B is significantly less than the others: Precision (HSD=0.11, F-ratio=95.5); Recall: (HSD=0.11, F-ratio=40.26). The Fallout of B is significantly less than F and S2, but the Fallout of S1 and B are indistinguishable (HSD=0.14, F-ratio= 5.12).

main source of error in all conditions is FNs, not FPs. Here too, the proportion of FNs is significantly higher than that of the FPs across conditions ($\chi^2 = 70.58$). This means that subjects are not recognizing relevant documents. As in the adhoc task, the FPs are not significantly larger a source of error in summaries than they are in full-text. As in the adhoc task, the main weakness with these summaries compared to full-text is that they are more often missing relevant information from the source (resulting in higher FNs). So, in these generic summaries, improved coverage of salient information in the source is required.

Figure 9 shows the time taken as a function of summary length. Each point on the graph represents the set of variable-length summaries at that summary size range (sets of size less than 20 aren't plotted). The curve here is similar to the adhoc case; the time taken steadily increases with summary length in a more or less linear manner until (in this case) about 2000 bytes, after which there is less linearity and fewer data points.

7.2.2 Performance by Participant

The F-score differences for participants on their variable-length and fixed-length summaries are shown in Tables 20 and 21, which are sorted by decreasing F-score. Even more so than the adhoc task, the systems here are astonishingly close in accuracy: for both variable-length and fixed-length summaries, the systems have no significant differences in F-score accuracy whatsoever. We may surmise that the systems are all finding the categorization task equally hard, with no particular technique for producing generic summaries standing out (in fact, the participants' accuracy at the level of TP, FP, etc. is also rather similar.) Since performance on the full-text isn't much better, we may assume that the task is hard to begin with. This latter point is further confirmed by examining the full-text performance data from the consistency task (shown later in Table 25).

Figures 10 and 11 show the categorization accuracy (F-score) versus time for different participants. As can be seen, for variable-length summaries (Figure 10), GE and IA are the closest in Cartesian distance from the ideal. For fixed-length summaries (Figure 11), ISI and SRA are closest in Cartesian distance from the ideal, but the accuracy and time differences between these and the rest aren't significant.

7.2.3 Effect of Compression Rate

Figure 12 shows the number of summaries at different compression rates. (Recall once more that the compression is defined as the number of characters in the summary divided by number of characters in the source.) The

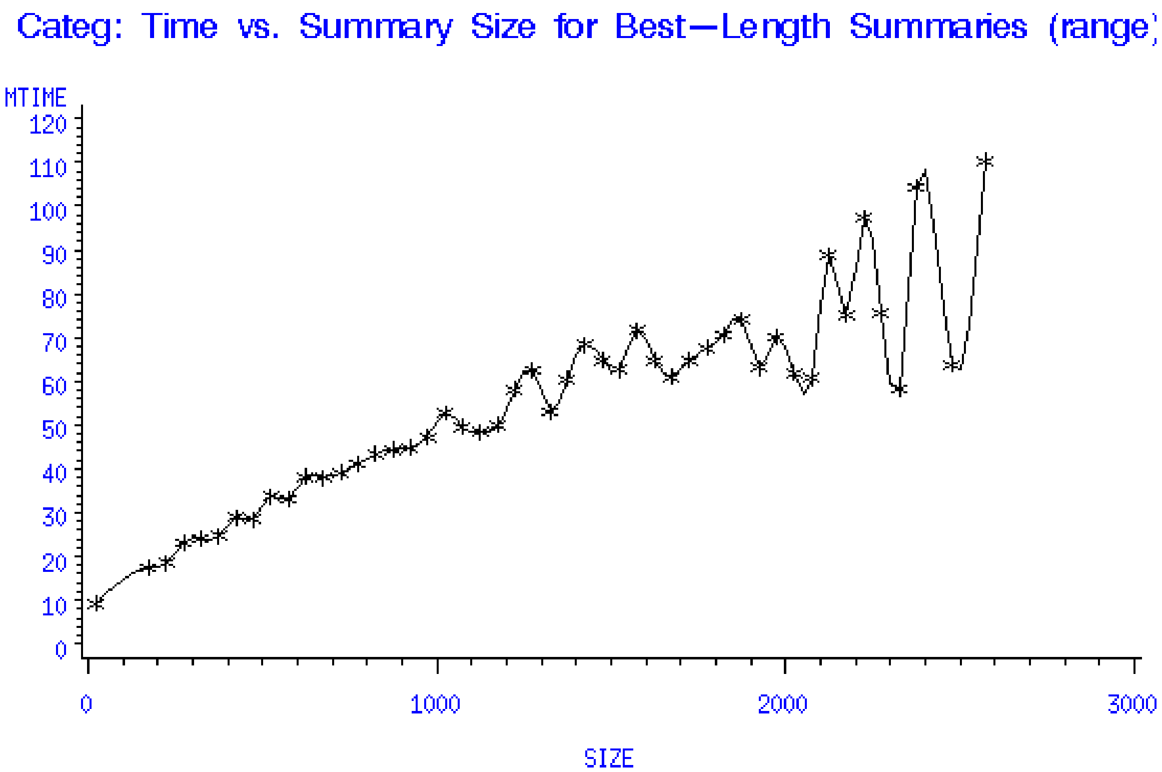


Figure 9: Categorization Time versus Summary Size in bytes (variable-length summaries).

Party	Precision	Recall	Fallout	F-score
CIR	.71	.47	.33	.54
IBM	.68	.47	.30	.51
NMSU	.69	.46	.35	.51
Surrey	.69	.43	.33	.51
Penn	.70	.42	.28	.50
ISI	.71	.42	.27	.49
IA	.69	.42	.28	.49
BT	.63	.43	.34	.48
NTU	.66	.41	.30	.48
SRA	.65	.42	.35	.48
LN	.68	.41	.34	.47
Cornell/SabIR	.66	.40	.29	.47
GE	.69	.40	.30	.47
CGI/CMU	.74	.39	.26	.47

Table 20: Categorization Accuracy (variable-length) by Participant. No significant differences on any of the measures.

Party	Precision	Recall	Fallout	F-score
LN	.68	.37	.25	.45
SRA	.73	.37	.21	.45
IBM	.63	.37	.25	.44
ISI	.71	.35	.20	.44
NMSU	.69	.34	.26	.43
NTU	.68	.33	.22	.43
CIR	.68	.35	.25	.43
Cornell/SabIR	.62	.36	.29	.42
CGI/CMU	.69	.33	.25	.42
GE	.69	.33	.22	.42
IA	.67	.33	.23	.41
BT	.70	.33	.23	.41
Surrey	.69	.31	.22	.39
Penn	.66	.29	.21	.38

Table 21: Categorization Accuracy (fixed-length) by Participant. No significant differences on any of the measures.

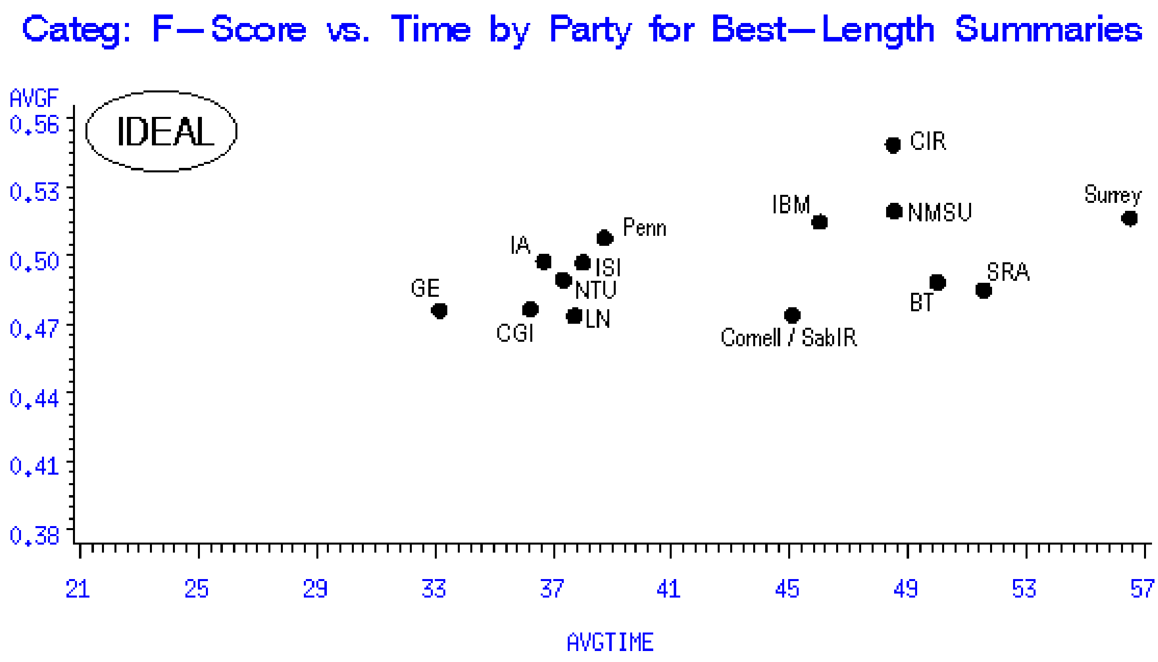


Figure 10: Categorization F-score versus Time by Participant (variable-length summaries). F-scores are not significantly different. HSD(Time) = 17.23, F-ratio(Time)=3.93. GE is significantly faster than SRA and Surrey. The latter two are also significantly slower than Penn, ISI, LN, NTU, IA, and CGI/CMU.

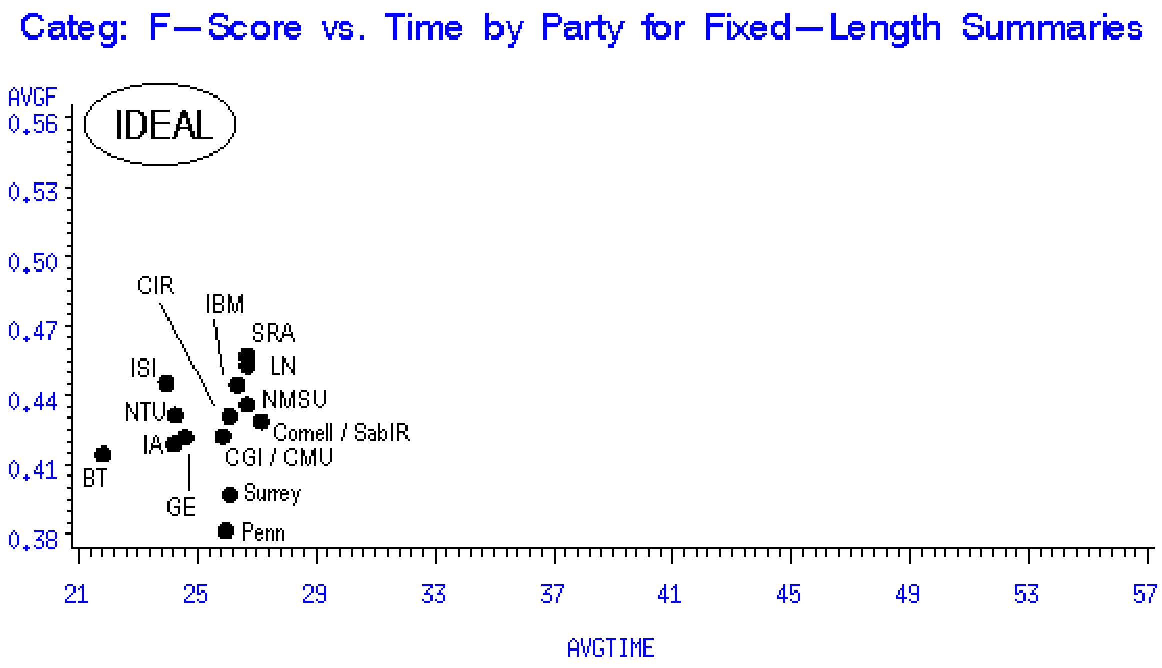


Figure 11: Categorization F-score versus Time by Participant (fixed-length summaries). F-scores are not significantly different, and neither are time differences.

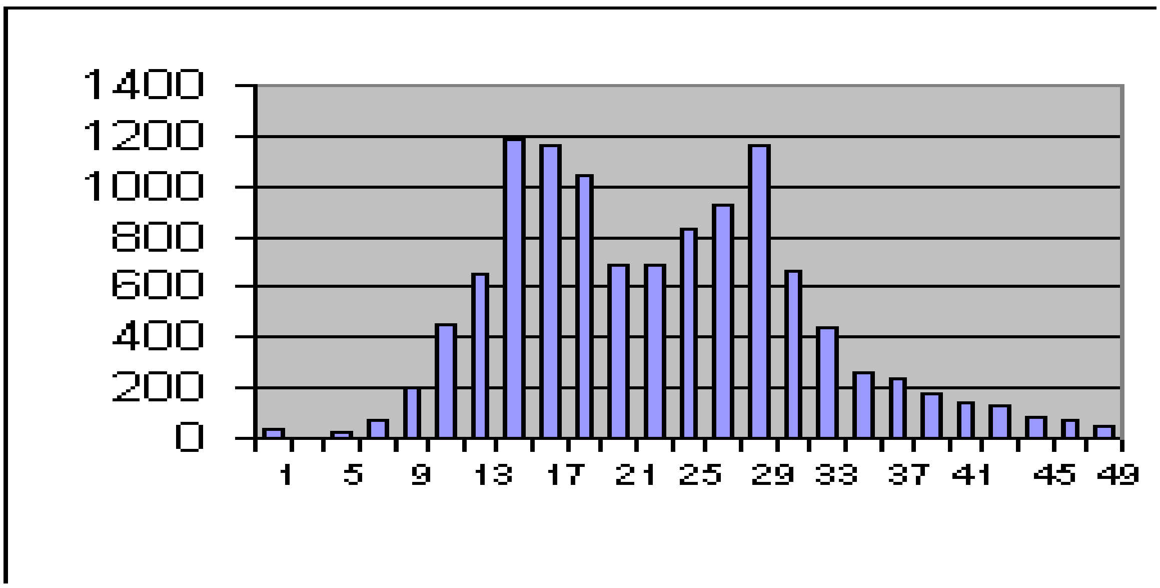


Figure 12: Categorization: Compression Rate Histogram (variable-length summaries).

Participant	Mean Compression	Compression Standard Deviation
Surrey	37.77	9.87
BT	30.43	3.45
CIR	28.05	13.56
SRA	27.62	4.34
NMSU	27.39	2.56
IBM	24.71	4.09
Cornell/SabIR	24.58	8.62
LN	20.90	8.85
Penn	20.76	8.91
NTU	20.04	9.29
CGI/CMU	17.71	4.41
IA	17.33	6.26
ISI	17.14	2.36
GE	15.82	3.21

Table 22: Categorization: Compression by Participant for variable-length summaries.

Percent subjects agreeing	Agree Relevant	Agree Irrelevant	Total	Running Total
100	47	148	195	195
80-99	107	175	282	477
70-79	36	128	164	641
60-69	70	89	159	800
50-59	94	106	200	1000

Table 23: Agreement between Subjects for Categorization Task

histogram here shows peaks at a wider range of compression rates than in the adhoc task. The mean and standard deviation of the compression for each participant are shown in Table 22. We see that some participants' summaries on this task are a longer fraction of the source than anyone's on the adhoc task, although on the average the categorization summaries have a similar average compression rate (23.58%) as the adhoc task (22.24%). The systems which were relatively closer to the ideal in Figure 10 (GE and IA) also had mean compressions at the lower end of the range (less than 18%).

We now turn to the impact of compression on F-score, shown in Figure 13 (where each point on the graph represents the set of variable-length summaries at that compression). This too shows a logarithmic relation between F-score and compression rate. As in the adhoc task, relatively low compression rates (e.g., here as low as 10%) are already achieving fairly high performance. Similar trends are reflected in the compression curves for the individual participants shown in Figure 14.

7.2.4 Agreement between Subjects

As shown in Table 23, unanimous agreement between subjects in the "consistency" categorization task on full-text is only on 19.5% of the data. The average pairwise agreement is 56.38%, with a value for Kappa of 0.29 and Cohen's e of 0.14, which means a large disagreement among subjects. Given this large disagreement, we decided to forego a revision of ground truth based on majority opinions in the categorization consistency task.

7.3 Query "Hardness"

When examining the accuracy data per query in Tables 24 and 25, we again see tremendous differences in accuracy across queries. Summarizers found certain queries "hard" to decide accurately in terms of F-score,

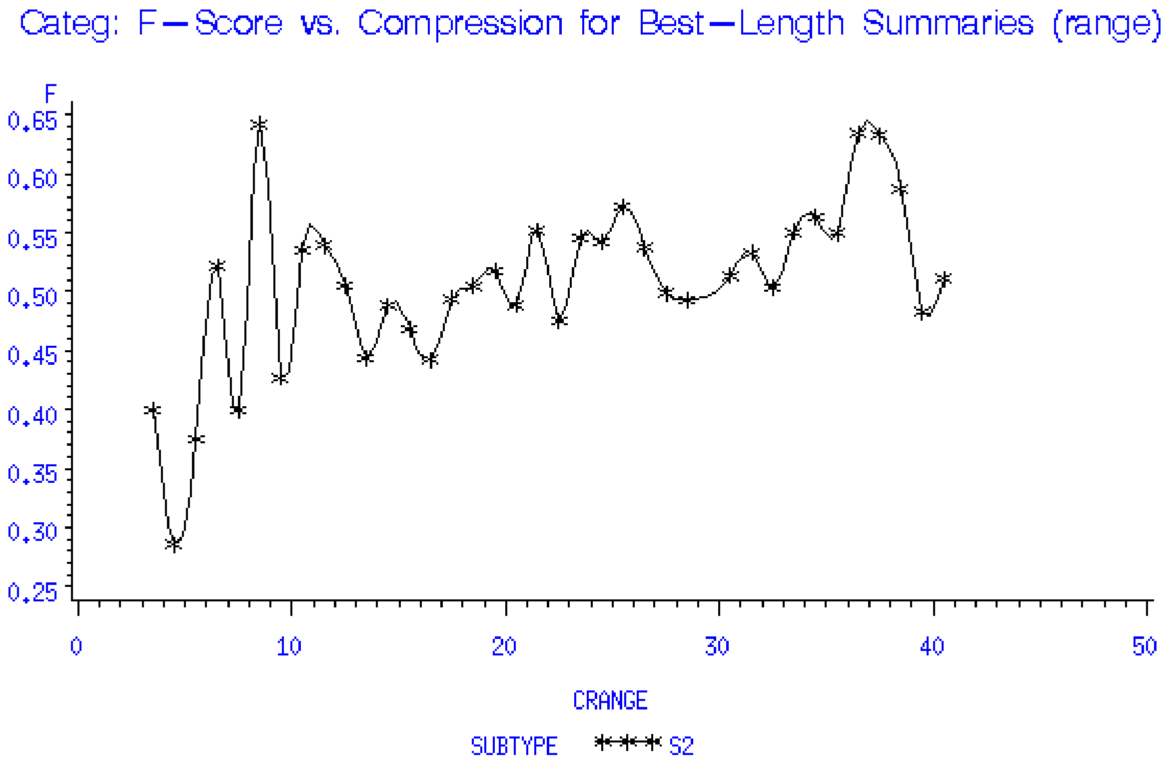


Figure 13: Categorization F-score versus Compression (variable-length summaries).

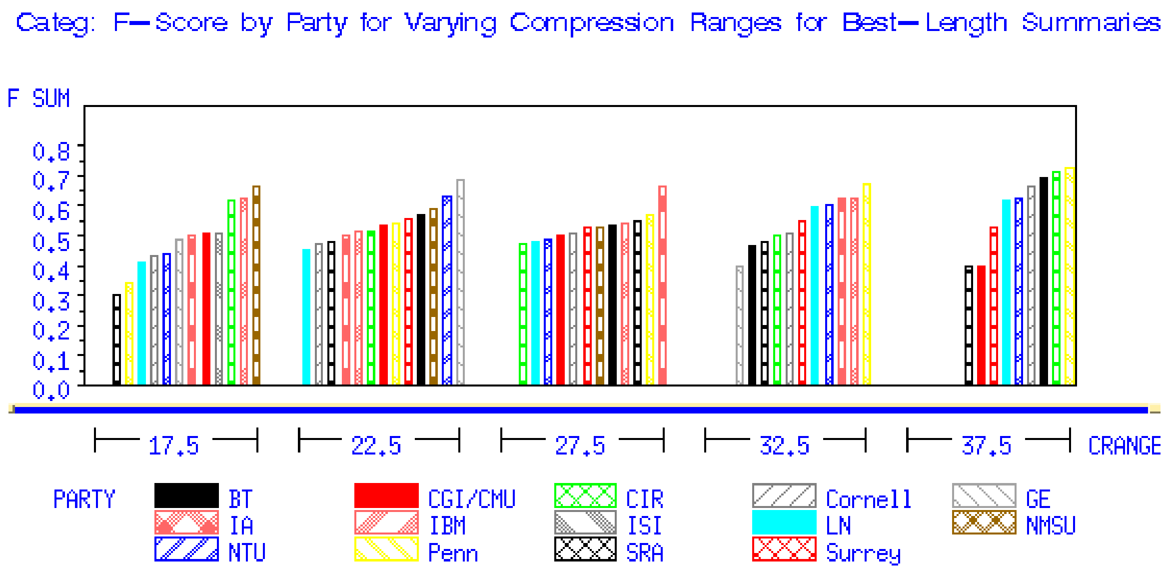


Figure 14: Categorization F-score versus Compression by Participant (variable-length summaries).

Query	Best	Fixed	Full-text	Percent Relevant
297	.81	.62	.81	.70
258	.80	.66	.79	.60
210	.77	.71	.79	.62
204	.77	.66	.71	.58
266	.74	.62	.78	.66
274	.73	.64	.42	.74
254	.72	.60	.74	.68
282	.69	.57	.76	.64
255	.68	.59	.72	.74
207	.67	.60	.72	.74
289	.67	.59	.62	.58
215	.62	.49	.64	.72
271	.60	.52	.58	.60
268	.55	.43	.71	.48
222	.55	.47	.48	.58
211	.55	.41	.48	.62
257	.52	.42	.64	.60
226	.46	.35	.51	.46
219	.41	.37	.46	.34
273	.38	.37	.42	.22

Table 24: Adhoc: F-score by Query for summaries and full-text. Many pairs are significantly different in accuracy (For best, HSD = .207, F-ratio = 11.66; for fixed, HSD =.235, F-ratio = 4.89). Full-text data from consistency task.

irrespective of the strategy used. There is a very strong positive correlation between query accuracy for fixed-length and query accuracy for variable-length summaries (Pearson product-moment correlation $r = 0.95$ in Table 24 and $r = 0.97$ in Table 25, $\alpha < 0.0001$).

In the categorization task, the hardness of queries is intrinsic to the task itself; for categorization the correlation between the accuracy for queries on summaries and accuracy for queries on full-text (the latter from the consistency task, where all 1000 full-text documents were judged by each subject) is also very strong (Pearson $r = 0.96, \alpha < 0.0001$ for both fixed-length and variable-length summaries). In other words, the observed ordering of the queries in terms of hardness is more or less independent of summarization in the case of the categorization task.

In the adhoc task, however, the hardness of queries for summaries is not simply a reflection of the hardness for full-text: the correlation between query accuracy for fixed-length summaries and query accuracy for full-text

Query	Best	Fixed	Full-text	Percent Relevant
291	.82	.76	.84	.68
290	.66	.63	.63	.50
284	.65	.58	.65	.46
269	.61	.44	.52	.62
237	.53	.46	.63	.69
208	.39	.37	.42	.56
228	.39	.34	.37	.34
249	.32	.24	.35	.38
243	.28	.18	.33	.67
251	.22	.17	.19	.60

Table 25: Categorization: F-score by Query for summaries and full-text. Many pairs are significantly different in accuracy (For best, HSD = .154, F-ratio = 27.6; for fixed, HSD =.235, F-ratio = 4.89). Full-text data from consistency task.

Condition	Confidence	Length	Intelligibility
	High/Med/Low	Longer/Just Right/Shorter	Fine/OK/Low
Full-text	69/27/4	1/28/71	44/55/1
Best	66/29/5	17/69/14	35/58/7
Fixed	63/29/8	45/54/1	28/55/17
Baseline	68/27/5	24/74/2	38/58/4

Table 26: Adhoc task: Assessor qualitative judgments expressed as percentages of total documents.

Condition	Confidence	Length	Intelligibility
	High/Med/Low	Longer/Just Right/Shorter	Fine/OK/Low
Full-text	63/33/4	1/45/54	45/54/1
Best	48/44/8	13/63/24	33/62/5
Fixed	46/43/11	47/50/3	26/58/16
Baseline	63/32/5	28/63/7	29/61/10

Table 27: Categorization task: Assessor qualitative judgments expressed as percentages of total documents.

is Pearson $r = 0.73, \alpha < 0.005$, and the correlation between query accuracy for variable-length summaries and query accuracy for full-text is Pearson $r = 0.66, \alpha < 0.005$. It is possible that since these summaries are tailored to each query, the hardness of a query in this task on summaries is related to the summarization techniques, and therefore doesn't align that closely with the hardness on full-text. However, harder queries in this task are associated with a relatively higher percentage of FP errors (Pearson $r = 0.82, \alpha < 0.0001$).

The reason for this is due to the "digging out" effect of summaries in the adhoc task. One might hypothesize that "harder" queries are associated with relatively high FPs, since a lower percentage of relevant documents for a query can mean the summary is "digging out" relevant passages, making the summary either truly relevant or apparently relevant, potentially increasing the FP rate. This hypothesis is confirmed by the data. Higher summary accuracy is to some extent positively correlated with the percentage of relevant full-text documents (i.e., judged relevant in TREC). There is a moderately strong positive correlation between the percentage of relevant full-text documents for the query and the query accuracy for variable length summaries (Pearson $r = 0.74$ in Table 24, $\alpha < 0.0001$) and somewhat less positive correlation for the fixed length summaries (Pearson $r = 0.66$, $\alpha < 0.001$ in Table 24). That this is an effect due to the summary is seen from the fact that the correlation between the percentage of relevant full-text documents for the query and the query accuracy for full-text is much less (Pearson $r = 0.50$, $\alpha < 0.02$ in Table 24).

Now, if the correlation between query accuracy and percentage relevant documents were entirely due to the "digging out" effect boosting the number of false positives, one would not expect such an effect in the categorization task, where the systems are unaware of the topics at the time of generating their generic summaries. This is borne out by the fact that on the categorization task, there is no correlation between query accuracy and percentage of relevant full-text documents for any of the summary or full-text conditions.

7.4 Analysis of Subjects' Feedback

As they evaluated each document for relevance to a topic, the assessors were asked to rate each document according to Degree of Confidence (Low/Medium/High), Preferred Length (Shorter/Just Right/Longer), and Intelligibility (Low/OK/Fine). The totals, expressed as percentages of total documents read, are included in Tables 26 and 27. There were approximately 400 full-text and baseline documents, and 10,000 fixed and best summaries for each task.

The findings for the main adhoc and categorization tasks are very similar. The confidence with which the

Confidence	Percentage Correct (TP or TN)	Total Documents
Low	43.23	2216
Medium	50.08	10172
High	58.87	10977

Table 28: Categorization task: Confidence in Decision and Correctness (all conditions).

Intelligibility	Percentage Correct (TP or TN)	Total Documents
Low	56.42	2423
OK/Fine	61.08	18151

Table 29: Adhoc task: Intelligibility and Correctness (all conditions).

assessor assigns relevance (roughly 90% medium to high confidence in all conditions) falls slightly as the length of the text condition decreases, although there is a slight increase for the baseline summaries (lead 10%) over the automatically generated summaries. In each task, differences between each pair of confidence categories across text conditions were significant, but the effect size was small (e.g., $e = 0.003$ overall for adhoc and 0.006 for categorization, $\alpha < 0.001$). On the categorization task, there is slight evidence that more confident judgments were more correct as shown in Table 28 (e.g., Low:Medium: $e = 0.01$, Medium:High: $e = 0.004, \alpha < 0.001$), but this effect wasn't there for the adhoc task.

A rating of "Fine" or "OK" for intelligibility was assigned over 80% of the time in each text condition; a "poor" rating was assigned more frequently as the length of the text condition decreased. Again, the baseline summaries did not share this trend, with only 4% of the adhoc baseline summaries rated "Poor" for intelligibility compared to 17% of the fixed length summaries. Since these baseline summaries consisted only of the initial 10% of the document, it is not surprising that they would rate higher for intelligibility than 10% selected from a broader spectrum of the document, as was the case with the fixed length summaries (however, it should be noted that the effect size here is small). In general, in each task, differences between each pair of intelligibility categories across text conditions were significant, but with a small effect size (e.g., $e = 0.03$ overall for adhoc and for categorization, $\alpha < 0.001$). Finally, as shown in Tables 29 and 30, there is also slight evidence that more intelligibility leads to more correct judgments, but the effect size is very small ($e = 0.0009$ for adhoc, .0002 for categorization, $\alpha < 0.001$).

The "preferred length" of the document showed great variation across different summarization conditions. When evaluating the full text, assessors would have preferred shorter documents over 70% of the time for the adhoc task. For the categorization task, assessors preferred shorter documents only 54% of the time when reading the full text, possibly indicating that lack of sufficient information in the summary was a problem more often than in the adhoc task (the additional information being needed, presumably, to draw a distinction between closely-related topic areas). In each task, differences between each pair of preferred length categories across text conditions were significant with a large effect size ($e = 0.16$ overall for adhoc and 0.19 for categorization, $\alpha < 0.001$). Further, as shown in Tables 31 and 32, documents which were judged to be acceptable in length or too long tended to be slightly more correctly judged than documents judged to be too short, but the effect size is very small ($e = 0.001$ for adhoc and 0.0007 for categorization, $\alpha < 0.001$).

Intelligibility	Percentage Correct (TP or TN)	Total Documents
Low	50.98	2297
OK/Fine	53.84	21068

Table 30: Categorization task: Intelligibility and Correctness (all conditions).

Preferred	Percentage Correct (TP or TN)	Total Documents
Just Right	62.01	12591
Shorter	60.79	170
Longer	57.46	6213

Table 31: Adhoc task: Length and Correctness (all conditions).

Length Preference	Percentage Correct (TP or TN)	Total Documents
Just Right	54.60	13196
Shorter	53.91	3322
Longer	51.38	6847

Table 32: Categorization task: Length and Correctness (all conditions).

8 Acceptability task

In this task, subjects were asked to judge the acceptability of generic summaries in relation to the full-text. Subjects read the full-text document and a sample of the variable-length summaries of all the systems participating in the categorization task. A sample of 30 full-text documents was chosen, 3 each selected randomly from each of the 10 categorization-task topics. The full-text document and the variable-length summary from each of the 14 categorization task participants was provided to each of four subjects. Each subject saw exactly the same summaries as the others (but in a different order), with 420 summaries in all. The subjects were instructed to read the full-text document and each of the summaries and to determine whether the summary was acceptable ("Yes") or not ("No"). Rather than provide explicit criteria for acceptability, which might be hard to define, it was left to the subject to apply whatever criteria they decided upon, but they were asked to apply them evenly to every summary.

The results, grouped by participant, are shown in Table 33. To take into account the varying compression for these summaries, the "Norm" column normalizes the percentage of "Yes" responses by the average compression rate for the participant. It is worth noting that while there were no significant differences in accuracy across participants' summaries in the categorization task, the acceptability of those summaries (i.e., in the acceptability task) varies quite a bit for different participants, ranging from 11% of a participant's summaries being acceptable to 71% of a participant's summaries being acceptable. In other words, summaries which are equally effective in indicating the content of a full-text source to support accurate categorization decisions, are not equally "acceptable" as summaries for the full-text.

In Table 34, we perform a statistical analysis of this data using χ^2 (see Equation 1 and its discussion), where Exp-Yes (the expected values for "Yes" responses for a cell) for a Participant is 40% of the participant's total number of observed "Yes" and "No" responses (likewise, Exp-No is 60% of the total observed for the Participant). Here the overall percentage of "Yes" responses is 40%, and the overall percentage of "No" responses is 60%. The χ^2 of 204.97 reveals there are significant differences between participants. Thus, overall, 60% of the summaries were found to be "unacceptable". However, looking closer, we see that BT, Surrey, IBM, SRA, and CIR have a majority of their summaries being acceptable, whereas the direction changes for the rest, which have a majority of their summaries being unacceptable. Finally, we see that there are two outliers in this data: BT, which is relatively more acceptable than the rest (BT's χ^2 is relatively higher than the rest), and ISI, which is relatively less acceptable than the rest (ISI's χ^2 is the next highest, and relatively much higher than the remaining).

In short, the participants' generic summaries, while equally indicative of content in the categorization task, were not equally informative of content, concise, readable, intelligible, coherent, etc., depending on what criteria of acceptability were used by the subjects in the acceptability task. It appears that judgments of "acceptability" in this task are somewhat related to judgments of "intelligibility" in the categorization task, given that there is a slight negative correlation between a participant's getting a "Low" intelligibility score on the categorization

Participant	Yes	No	"Yes" %	Compression	Norm
BT	85	34	71	.36	.45
Surrey	72	48	60	.45	.33
IBM	71	49	59	.29	.42
SRA	67	52	56	.33	.38
CIR	61	59	51	.30	.35
NMSU	54	66	45	.34	.30
Penn	51	69	42	.27	.31
NTU	50	70	42	.29	.29
CGI/CMU	39	80	32	.24	.24
LN	35	85	29	.26	.21
GE	31	88	26	.21	.21
Cornell/SabIR	25	95	14	.27	.15
IA	23	96	14	.24	.14
ISI	14	106	11	.25	.08

Table 33: Acceptability task: Percentage of Yes and No responses out of 420 generic variable-length summaries, by Participant.

Participant	Obs-Yes	Obs-No	Exp-Yes	Exp-No	χ^2
BT	85	34	48	71	47.80
Surrey	72	48	49	71	18.24
IBM	71	49	49	71	16.69
SRA	67	52	48	71	12.60
CIR	61	59	49	71	4.96
NMSU	54	66	49	71	0.86
Penn	51	69	49	71	0.13
NTU	50	70	49	71	0.03
CGI/CMU	39	80	48	71	2.82
LN	35	85	49	71	6.76
GE	31	88	48	71	10.09
Cornell/SabIR	25	95	49	71	19.86
IA	23	96	48	71	21.82
ISI	14	106	49	71	42.25
TOTALS	678	997	681	994	204.97
Percent	0.40	0.60			

Table 34: Acceptability task: Observed, Expected, and χ^2 for each Participant. BT, Surrey, IBM, SRA, and CIR have a majority of their summaries being acceptable, whereas the rest have a majority of their summaries being unacceptable.

Percent subjects agreeing	Agree Acceptable	Agree Unacceptable	Total	Running Total
100	38	114	152	152
75	75	85	160	312
50	108	0	108	420

Table 35: Agreement between Subjects for Acceptability Task

task and the "Yes percentage" for that participant in the acceptability task (Pearson $r = -0.58$, $\alpha < 0.02$).

The agreement between the subjects on this task is shown in Table 35. The unanimous agreement between these 4 subjects is only on 36% of the data (i.e., 152/420). The Kappa of 0.24 means there is a large disagreement among these 4 subjects overall as to the acceptability of summaries. However, given that each subject had an open-ended definition of "acceptability", it is unclear whether this disagreement is due to different interpretations of "acceptability" or due to similar interpretations with different judgments.

9 Question-answering (Q&A) task

As mentioned earlier in Section 2, we also conducted an intrinsic evaluation which measures the extent to which a summary provides answers to a set of questions that must be satisfied for a document to be judged relevant to a topic. The design of this evaluation task is intended to reflect the nature of a potential real-life task involving technologies for information retrieval and topic-specific summarization. The analyst might first query a text database for documents on the report topic, use summaries produced by an automatic summarization system for highly-ranked documents to identify the available information that is most pertinent to the query, and use that information in composing her report.

The challenge for the systems is to understand the topic in relation to each document and to produce an "informative" user-focused summary that covers all *obligatory aspects* of the topic in as short a summary as possible. An *obligatory aspect* of a topic is information that must be present in the document for the document to be relevant to the topic. Thus, a summary would be more accurate if it correctly answered a higher percentage of the questions that represented the obligatory aspects of the topic.

For example, an obligatory aspect of a long topic concerning overcrowded prisons would be that the document must contain information about prisons that are overcrowded; the *implicit question* that would have to be answerable by documents that someone had judged relevant to the topic might be, "What is the name and/or location of each correction facility where the reported overcrowding exists?" In this implicit question, "prison" has been expanded to "correction facility", the identity of such a facility has been specified as meaning the "name and/or location", and the "overcrowding" term has been loosened to "reported overcrowding". Since the wording of the implicit questions reflects the actual manifestation of obligatory aspects of the topic in relevant documents, they provide a more precise basis for evaluation than the topic itself.

Thus, the overall objective of the Q&A task is to determine how accurate summaries are in answering questions. We wanted to compare the accuracy of different participants' summarization systems on this task. This required that we have a methodology for constructing questions, constructing answers, and scoring summaries against the answers. We decided to specify the answers in terms of an in-text *answer key*: i.e., each passage in the full-text source which constituted an answer to a question was marked up to indicate which question it answered (thus the answer key would be instantiated for each full-text source document). To score the summarization system's performance on this task, this passage-level representation of the answer is compared against the summary. In other words, the accuracy of the summarization system is related to how many of the correct passages are recalled.

However, determining accuracy is not just a matter of counting the number of matching passages. For example, a document may:

- provide the same information in more than one place (in which case the summary may be accurate if it contains just one relevant passage);

- provide less- and more-specific versions of the same information (in which case the summary may be judged more accurate if it contains the more specific passage);
- provide separate passages that enumerate instances of the information (e.g., separate passages identifying different overcrowded prisons, in which case the summary may be judged accurate only if contains all the passages).

Even in the simplest case, where the document contains exactly one passage that answers a question, it may not be enough to judge a summary as accurate for that question if it contains that one passage. It may be necessary for the summary to contain additional sentences that provide context for that passage, so that the reader realizes that the passage is actually answering the question. For example, if one sentence gives the name of a prison, and a separate sentence identifies that prison as being overcrowded, it will not be sufficient for the summary to contain just one of those sentences, even though the answer key will have the prison name in the first sentence marked as containing the answer to the question. The "fact" marked in the answer key must be supplemented by context provided by other material from either the same sentence or another sentence in order to constitute a correct answer.

Now, if a system simply extracts passages, then it is trivial to automatically score the summaries in terms of which passages got extracted. However, this would not necessarily address the issues outlined above that require judgments about interpreting the key and about providing context. If the passages are modified in some way, or if the summarizer doesn't extract passages (e.g., generating an abstract, or a list of named entities, etc.), then the comparison of the summary against the answer key is somewhat more difficult. Since we wanted to be able to evaluate a variety of summary types and since we believed that accurate scoring would require a significant amount of human judgment (especially given the impoverished nature of the annotation used in the answer key), we developed and carried out a manual scoring procedure. Complementary to our work, however, an automatic scoring program was developed by Chris Buckley from the Cornell/SabIR group. The results from his scoring program (which was applicable to all except one of the Q&A participants - the exception being one which didn't extract passages) are also reported here, and compared against the manual scoring results.

It should be stressed that the evaluation conducted on this task was very preliminary in nature. Due to resource limitations, it was not carried out on a large enough scale to permit statistical analyses of the results.

9.1 Experimental Design

The topics we chose (257, 258, and 271) were a subset of the 20 adhoc TREC topics shown in Table 2. Topic 257 (Cigarette Consumption) was more statistics-oriented (requiring data on cigarette consumption), Topic 258 (Computer Security) had to do with a particular type of event (requiring identification of instances of illegal entry), and Topic 271 (Solar Power) was about a general trend (requiring information on extent of use of solar power worldwide). For each topic, 30 relevant documents were chosen as the source texts for user-focused summarization. The documents in the subcollections did not overlap across topics.

As discussed in the previous section, a set of questions was defined for each topic, each question reflecting an obligatory aspect of the topic. (The title, description and narrative fields in the TREC topic definition were all considered to be part of the topic.) An answer key was then created: every document in the 90-document subcollection was annotated to identify answers to the questions. The summaries generated by the participants (who were given the topics and the documents to be summarized, but not the questions) were then scored against the answer key.

One person per topic was designated the evaluator for the topic. The principal tasks of each evaluator were to prepare the questions and answer keys and to score the system summaries. The rationale for having one person responsible for both task definition and assessment is based on the hypothesized real-life situation described in the previous section, in which it is a single information analyst who conducts the document search

Title : Computer Security	
Description : Identify instances of illegal entry into sensitive computer networks by nonauthorized personnel.	
Narrative : Illegal entry into sensitive computer networks is a serious and potentially menacing problem. Both 'hackers' and foreign agents have been known to acquire unauthorized entry into various networks. Items relative this subject would include but not be limited to instances of illegally entering networks containing information of a sensitive nature to specific countries, such as defense or technology information, international banking, etc. Items of a personal nature (e.g. credit card fraud, changing of college test scores) should not be considered relevant.	
Questions	
1)Who is the known or suspected hacker accessing a sensitive computer or computer network?	
2) How is the hacking accomplished or putatively achieved?	
3) Who is the apparent target of the hacker?	
4) What did the hacker accomplish once the violation occurred? What was the purpose in performing the violation?	
5) What is the time period over which the breakins were occurring?	

Table 36: Q&A Topic 258

and determines the relevant content that was central to the intention of the search topic. This approach leads to evaluation tasks that are "individualistic", and in order to combat that sort of individual bias, a Q&A evaluation of the sort we have defined would have to be conducted on a much larger scale (as done in TREC). For our evaluation, we created a set of guidelines for writing questions, creating answer keys, and scoring summaries that were intended to minimize variability in the methods used by the different evaluators. In addition, we conducted a small study that would help us assess the comparability of the scores across topics; this interassessor study is described in a later section.

To construct questions, each evaluator read his/her TREC topic definition and the set of 30 relevant documents. For each document, the evaluator noted down what made each document relevant to the topic. (There were a few documents which question-developers read and deemed totally irrelevant to the topic; these were replaced until all were deemed relevant by the evaluator.) The evaluator then drafted no more than 5 questions that pertained to the obligatory aspects of the topic. We ended up with 4 questions on Topic 257 and 5 questions on each of the other two topics. Each question had to be answerable by a sizeable number of documents (preferably by more than half of them). It turned out that on average, each question had an answer in approximately two-thirds of the documents. The evaluator was encouraged to include different kinds of questions, e.g., who/what/where/when/how/how much. The evaluator was also allowed to compare his/her questions with those of the other two evaluators and to revise her questions if desired. Finally, the evaluator kept notes on the intended interpretation of questions.

To construct the answer key, each evaluator marked off any passages in the text that provided an answer to a question. The passage was marked off with a tag indicating the question number it answered; the tag provided no additional information, e.g., concerning the relative quality of the tagged passage compared to any other passages that were tagged with the same question number. In the course of answer annotation, question revision was permitted, e.g., to allow a question's answer to come from more documents. The answer annotation was carried out using MITRE's Alembic Workbench [Day et al. 1997], [Day et al. 1998].

As an example, Topic 258 and the constructed questions are shown in Table 36. An article relevant to that topic is shown with its answer annotations in Figure 15. Each topic ended up with a different number of answer tags; Topic 257 had the sparsest set of answers, and Topic 258 the richest set of answers. As a measure of the number of questions answered by a document, there were an average of 2.5 for Topic 257, 3.6 for Topic 258, and 3.2 for Topic 271. As a measure of the number of tagged passages per question per document, there were an average of 2.6 tags for Topic 257, 3.3 tags for Topic 258, and 3.6 tags for Topic 271.

To allow estimation of the degree of compression that would be achievable without sacrificing accuracy, model summaries were created by the evaluators. Using the answer keys for reference, each evaluator created a new set of tagged documents, where each tag indicated a sentence that was part of a model summary. These sentences had to cover the material in the answer key (where an answer might be a portion of a sentence or could span several sentences), with redundant sentences eliminated, and with "contextual" sentences introduced to make the sentence extracts more coherent (e.g., to avoid coherence gaps and dangling anaphors). No compression requirements were specified for construction of the model summaries. The model summaries had to be 100% correct, using the ARS accuracy metric discussed in the next section.

The participants in the Q&A evaluation consisted of 8 of the 11 adhoc task participants. They were allowed to generate fresh summaries for this task. Each participant got 90 articles per topic, which contained the 30 relevant articles per topic discussed above (the participants were not informed that only 30 summaries per topic would be used). Prior to the evaluation, each participant was provided with a topic, a set of 5 questions, 50 relevant articles, and answer keys, for use as training data, and to familiarize themselves with the task. The participants were allowed to generate variable-length summaries, as we wanted to study the effect of compression in this task. The 30 summaries per topic were then scored against the answer keys. One person per topic was responsible for preparing questions and answer keys and for scoring summaries (3 assessors in all).

9.2 Scoring

We now explain the manual scoring method. The answer to each question was judged Correct, Partially Correct, or Missing based on a comparison of the summary of a document against the set of tagged passages for that question in the answer key for that document. Note that if a key for a particular document did not contain any answer at all for a question, the summary for that document didn't contribute any score to that question; i.e., that question did not affect the Answer Recall measures.

The evaluators were provided with guidelines, including the following:

1. If a summary contained a passage marked as the (sole) answer in the answer key, the answer to the question was to be judged Correct if there was sufficient context in the passage or elsewhere in the summary to make that judgment.

2. If the summary contained the marked passage but there wasn't sufficient context for the passage to be clearly understood, the summary's answer to the question was to be judged Partially Correct; if needed context was totally lacking or if the context was misleading, the summary's answer to the question was to be judged Missing.

3. If the summary didn't contain a Correct or Partially Correct answer to the question, then the summary's answer to the question was to be judged Missing.

4. Finally, when multiple passages were marked in the answer key as an answer to a question, the scoring was to be carried out based on the above criteria, with additional judgments made on a case by case basis concerning the scoring category to be assigned when the summary did not contain *all* the passages tagged in the key. The key served to indicate all the passages that participated in making the judgment, but it did not indicate which passages were redundant with others, which were "better" (e.g., more specific) than others, nor which passages were part of an enumeration along with others[3].

[3]It may be possible to extend the markup formalism to capture these various cases, but in addition to the increased complexity of the markup, adequate definition of the guidelines would require an intensive collaborative effort in order to ensure replicability.

```
<DOCNO> AP881209-0184 </DOCNO>
<FILEID>AP-NR-12-09-88 0806EST</FILEID>
<1ST_LINE>a a PM-AnotherLook:ComputerVirus Adv12   12-09 0567</1ST_LINE>
<2ND_LINE>PM-Another Look: Computer Virus, Adv 12,0584</2ND_LINE>
<NOTE>$Adv12</NOTE>
<NOTE>For Release Monday PMs, Dec. 12, and Thereafter</NOTE>
<HEAD>Student Linked to `Virus' Teaching Lawyer, Declining Offers</HEAD>
<BYLINE>By WILLIAM KATES</BYLINE>
<BYLINE>Associated Press Writer</BYLINE>
<DATELINE>SYRACUSE, N.Y. (AP) </DATELINE>
<TEXT>
   As a federal grand jury decides whether he
should be prosecuted, <Q1>a graduate student</Q1> linked to a ``virus'' that
disrupted computers nationwide<Q5> last month </Q5>has been teaching his
lawyer about the technical subject and turning down offers for his
life story.
 <Q1>  Robert Morris Jr.</Q1> has declined offers from movie producers, book
publishers, magazines and television shows, said his lawyer, Thomas
Guidoboni.
   ``Until we get the potential criminal matter resolved,''
Guidoboni said, the<Q1> 23-year-old Cornell University graduate student</Q1>
will remain silent, despite heavy pressure to talk.
....
   No charges have been filed against <Q1>Morris</Q1>, who reportedly told
friends that he designed the virus that temporarily clogged about
<Q3>6,000 university and military computers</Q3><Q2> linked to the Pentagon's
Arpanet network.
</Q2> .....
   According to search warrants authorities executed at Harvard and
Cornell universities in the days following the<Q5> Nov. 2 episode</Q5>, FBI
investigators are reviewing statutes concerning computer fraud, wire
fraud, malicious mischief and unlawful access to stored
communications.
....
   <Q1>Morris</Q1>, whose father, Robert T. Morris Sr., is the chief
scientist at the National Computer Security Center in Bethesda, Md.,
reportedly wrote the virus as an intellectual experiment and did not
intend any harm.
   The program apparently was intended to live undetected in the
Arpanet system in which it was first introduced and secretly make
copies that would move from computer to computer. However,<Q4> a design
error caused it to replicate out of control in computers linked to
the system.
</Q4> ....
</TEXT>}
```

Figure 15: Parts of a relevant document showing answers to questions on Topic 257

	257			258			271		
Party	**ARS**	**ARL**	**ARA**	**ARS**	**ARL**	**ARA**	**ARS**	**ARL**	**ARA**
CGI/CMU	69	77	73	63	71	67	66	78	72
Cornell/SabIR	61	68	65	64	70	67	60	73	66
GE	49	60	55	58	67	63	56	70	63
NMSU	52	63	58	62	69	65	48	63	55
SRA	36	46	41	61	69	65	35	45	40
ISI	30	42	36	55	59	57	22	34	28
TextWise	27	34	30	32	40	36	29	41	35
Penn	17	24	21	48	55	51	24	34	29

Table 37: Answer Recall by Participant.

Topic	**Mean**	**Standard Deviation**	**Variance**
257	47.37	18.14	329.41
258	58.87	10.77	116.12
271	48.5	17.59	309.42

Table 38: Answer Recall (ARA) by Topic.

Two accuracy metrics were defined to capture the accuracy of a summary in answering questions, ARL (Answer Recall Lenient) and ARS (Answer Recall Strict):

$$ARL = (n1 + (.5 * n2))/n3 \tag{9}$$

where $n1$ is the number of Correct answers in the summary, $n2$ is the number of Partially Correct answers in the summary, and $n3$ is the number of questions answered in the key. Note that the definition of the denominator means that the range of values will be between 1 and 4 for Topic 257 and between 1 and 5 for Topics 258 and 271.

$$ARS = n1/n3 \tag{10}$$

where $n1$ and $n3$ are as above.

A third measure ARA (Answer Recall Average) was defined as the average of ARL and ARS and was commonly used as a succinct way of reporting the results of the evaluation, in combination with the compression measure. There were no scoring categories for "spurious" or "incorrect" answers, as it was unclear how to define them in the context of question-answering by summarization; therefore, there was no precision metric (and consequently, no F-score).

9.3 Results

9.3.1 Basic Results

Table 37 shows the three ARA measures for different participants. The ARA ordering of the systems on topics 257 and 271 is strongly correlated (Pearson $r = 0.95$, $\alpha < 0.003$). It is worth noting that the systems which are in Group I in accuracy on the adhoc task (CGI/CMU and Cornell/SabIR) are at the top of the ordering. Further, as shown in Table 38, the ARA on Topic 258 is higher and varies less than on other topics. However, in the absence of knowledge of statistical significance of the differences, we can't arrive at any definitive conclusions about accuracy differences among participants on this task.

Figure 16 shows a plot of the ARA against compression. The model summaries were able to achieve compression rates of 10% on topic 271, whereas for the other summaries the compression rate was closer to

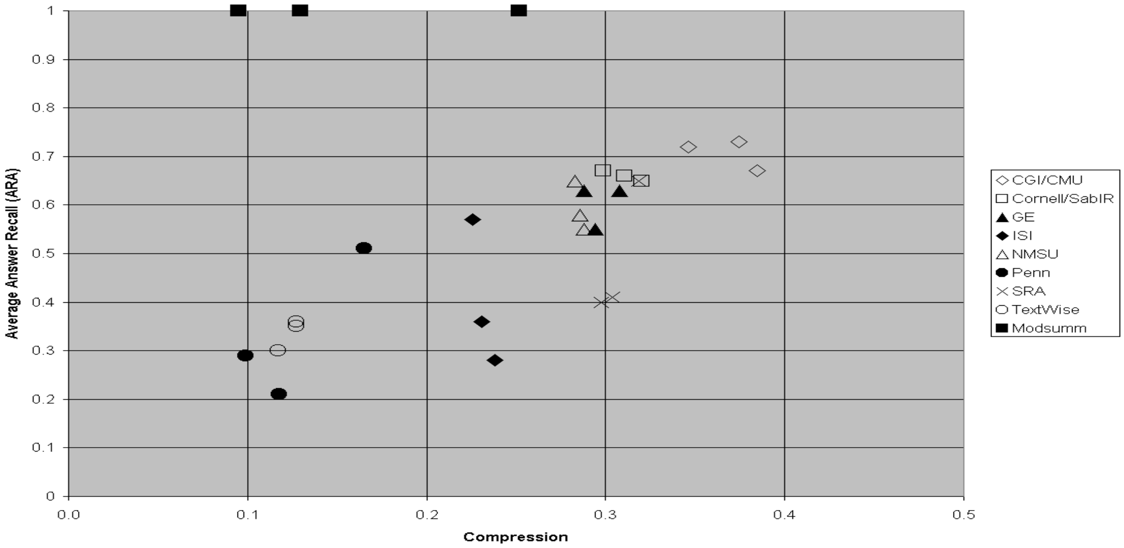


Figure 16: ARA versus Compression by Participant.

Overlap Metric	Definition
V4	full credit if the text spans for all tagged key passages are found in their entirety in the summary
V5	full credit if the text spans for all tagged key passages are found in their entirety in the summary; half credit if the text spans for all tagged key passages are found in some combination of full or truncated form in the summary
V6	full credit if the text spans for all tagged key passages are found in some combination of full or truncated form in the summary
V7	percentage of credit assigned that is commensurate with the extent to which the text spans for tagged key passages are present in the summary

Table 39: Overlap Measures for Q&A Automatic Scoring of each question that has an answer according to the key.

25%. The highest *ARA* is associated with the least reduction (35-40% compression), indicating a tradeoff between accuracy and compression. As pointed out earlier, scores on Topic 258 are more tightly grouped (and have less variance than) scores for other topics. This is because, as will be explained in Section 9.3.3, most of the documents on 258, in contrast to the other topics, were "on-topic".

9.3.2 Comparison with automatic scoring results

Chris Buckley of Cornell/SabIR developed an automatic scoring program for this task, which measured the overlap between summaries and answers in the answer key. Four of the overlap measures are shown in Table 39. As with the Answer Recall measures, these overlap measures are expressed in terms of the percentage of questions that can be answered, according to the key. The program computes the degree to which the summaries align with the tagged passages found in the key. For a given document, measures V4, V5 and V6 give summaries no credit if they are missing any of the text spans tagged in the key (for questions where the key contains more than one tagged passage); V4 is strictest, followed by V5, then V6. V7 will give some credit if at least one of several tagged segments in the key is found (in its entirety or in truncated form) in the summary.

Participant	ARS	ARL	ARA	V4	V5	V6	V7
CGI/CMU	0.66	0.75	0.71	31.0	54.2	77.66	53.66
Cornell/SabIR	0.62	0.71	0.66	26.33	48.86	71.53	48.73
GE	0.55	0.66	0.60	24.06	46.66	69.33	44.06
NMSU	0.54	0.65	0.59	19.66	45.46	71.0	44.33
SRA	0.44	0.53	0.49	14.8	34.86	54.6	34.06
ISI	0.36	0.45	0.40	10.13	30.66	51.26	27.53
Penn	0.30	0.38	0.34	8.2	24.13	40.2	20.8

Table 40: Answer Recall and Overlap by Participant.

Category	Definition
T	The main theme of the document is on the topic.
MT	The document has multiple main themes, and the topic is one of them
AT	The document basically is an anecdote related to the topic
ST	The document is relevant to the topic, but the topic is a subsidiary theme

Table 41: Distribution of Topic Prominence Categories.

The scores for the participants using these automatically scored overlap measures are shown in Table 40, along with the corresponding scores using the manual method. (TextWise's data wasn't analyzed using the overlap measure, since it provided lists of named entities, instead of extracting passages.) The correlation between the participants' scores on each of the four overlap measures and each of the Answer Recall scores in Table 40 is very strong (Pearson $r > .97$, $\alpha < 0.0001$). This means that for the kinds of summaries under consideration, the automatic scoring method will be suitable for this task. The value of an objective, automated method for scoring summaries for such a task cannot be underestimated, even though the manual scoring method captures many subtleties (discussed earlier in Section 9.2) that the automated method doesn't.

9.3.3 Topic Prominence

In this study, all documents in the test set were analyzed according to the degree to which they were "on" topic, using a four-way classification scheme shown in Table 41. (Each annotator judged the documents in his/her own topic's test set, and the judgments were reviewed and modified as needed by one person in order to ensure reasonable consistency in the application of classification guidelines.) The purpose of making the classification was not merely to identify distributional differences across topics, but also to determine whether the different classifications were reflected as differences in the test results. In particular, we wanted to know if any of the following hypotheses was true:

1. The lengths of summaries of documents that are classified as "T" will vary with the length of the full document, but the lengths of other summaries won't vary.
2. Summaries of documents that are classified as "T" (and maybe "MT") are more accurate than those produced for documents classified as "ST".

The results of this analysis for each of the topics are shown in Figure 17. It can be seen that topics 251 and 271 are secondary topics in most of their documents. Almost all the documents (26/30) in topic 258's 30-document set were completely on topic. No documents in these test sets were classified as "AT".

Regarding our hypotheses, the first hypothesis seems false for the system summaries and true for the model summaries. Summaries from six out of eight participants vary with the length of the full document regardless of topic prominence, whereas the remaining two participants' summaries stay the same length regardless of topic prominence. Model summaries tend to be longer on average for longer documents than for shorter documents, but there are exceptions to this trend. The second hypothesis seems false as well: there is similar spread

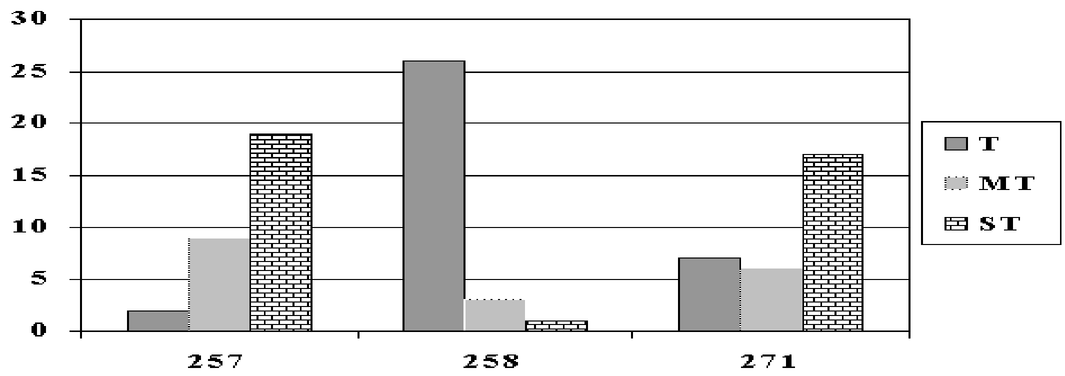


Figure 17: Distribution of On- and Off-Topic Documents.

and density of scores across topics. However, definitive testing of these hypotheses with statistical significance measures will require a larger-scale study.

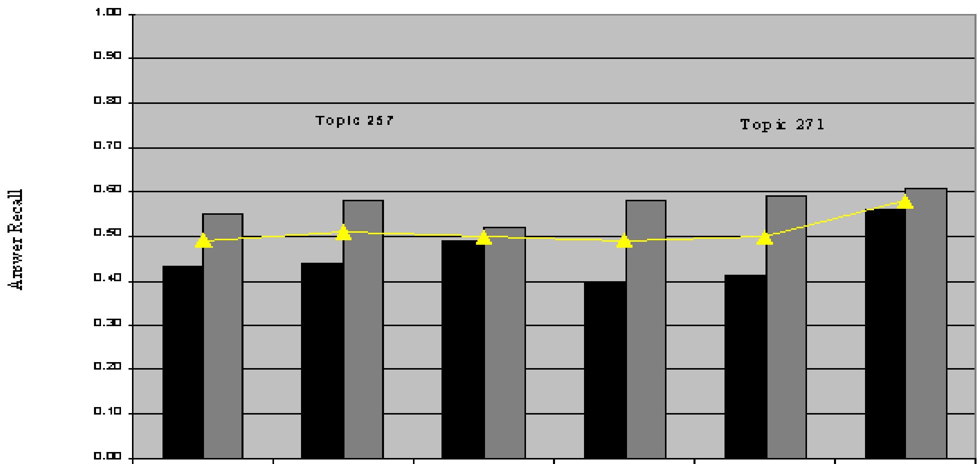
A second breakdown of the test documents was made according to the "genre" of each (with categories for 'news', 'feature', 'opinions', and combinations thereof). The breakdown was intended to reveal genre distribution differences across topics and to help us determine whether answer recall varied from one genre to the next, irrespective of topic. The results were not analyzed in detail, but an initial analysis showed no real effects.

9.3.4 Interassessor Reliability

We also conducted a small experiment to determine the relative stringency of each person who scored the answers. To do this, we had each of the three assessors score a portion of each of the other assessors' test data (i.e., using the other assessors' questions, notes about the intended interpretation of the questions, and answer keys). Each assessor selected 10 out of the 30 documents in his/her own test set for inclusion in the study. The selected documents covered just 4 of the summarization systems. The results are shown in Figure 18 (topic 258 was thrown out because the assessor omitted some annotations that were essential to her method of scoring, and these omissions precluded the possibility of obtaining consistent scoring by other assessors). We see that Subject3 tends to be more lenient on ARS than the others, giving higher scores on each topic. Subject3's ARS and ARL are also closer together than those of the other two assessors (3-8 point difference versus 12-18 point difference). The $ARAs$ across both topics appear to be very similar across all three assessors, with the exception of Subject3's ARA on topic 271.

10 Comparisons with Related Evaluations

As mentioned earlier, there have been several other extrinsic evaluations dealing with the task of relevance assessment [Mani and Bloedorn 1999], [Jing et al. 1998], [Tombros et al. 1998], [Brandow et al. 1994]. As noted in the introduction, the SUMMAC evaluation is the first large-scale developer-independent evaluation of summarization systems. It also explores a variety of detailed analyses not reported in these earlier studies, allowing new



	Subject1	Subject2	Subject3	Subject1	Subject2	Subject3
ARS	0.43	0.44	0.49	0.40	0.41	0.56
ARL	0.55	0.58	0.52	0.58	0.59	0.61
ARA	0.49	0.51	0.5	0.49	0.5	0.58

Figure 18: Answer Recall Across Assessors.

insights to be gained. In addition, those earlier studies differ from SUMMAC in task details and methodology. In the [Tombros et al. 1998] evaluation, the subjects are asked to find as many relevant documents as possible for a query within five minutes, with a comparison made between user-focused and generic summaries. Subjects were also allowed to reference the full-text. While this study is very interesting, it is subject to confounding factors such as "trying to beat the clock" as well as the interference from full-text access in determination of summary relevance. The [Jing et al. 1998] evaluation was a pilot study which used a small number of queries (4) and documents (10 per query). A similar criticism can be made of [Mani and Bloedorn 1999], which in addition lacked a baseline (e.g., leading text) summary for comparison. The [Brandow et al. 1994] extrinsic evaluation differed from the others in the following way. Instead of selecting the set of documents for a query based on an IR system's search against a full-text index, and then varying which instantiation (full-text or summary) each subject saw, [Brandow et al. 1994] selected multiple sets of documents for each query, based on varying between a full-text and a summary index. Their paper also does not include any statistical analysis of results.

There have also been numerous intrinsic evaluations, including judgments of fluency of the summary [Minel et al. 1997], [Brandow et al. 1994], coverage of stipulated "key/essential ideas" [Paice 1990], [Brandow et al. 1994], or similarity to an "ideal" summary, e.g., [Edmundson 1969], [Kupiec et al. 1995]. The acceptability task here is a small-scale evaluation, similar in method to the intrinsic study by [Brandow et al. 1994]; however, the task allows us to determine acceptability of generic summaries used in the categorization task, along with statistical analyses. The Q&A task described here, is quite different from the earlier ones, going beyond the "key-ideas" approach to offer a new information-extraction inspired, answer-key-based methodology for evaluating user-focused summaries as surrogates for full documents.

11 Conclusions

In conclusion, the SUMMAC evaluation has made a number of contributions:

- It has established that automatic text summarization is very effective in relevance assessment tasks, with summaries at compressions over 10% being as effective in supporting efficient relevance assessment as full-text.
- Rather than compare summaries against ideal summaries, which are hard to arrive at and rarely unique, SUMMAC has used two relatively new methods for evaluating performance of indicative user-focused and generic summaries in terms of their impact on relevance assement tasks.
- SUMMAC has also introduced a new method for evaluating user-focused summaries in terms of their informativeness in covering a topic.
- The large-scale nature of the SUMMAC evaluation has made it possible to arrive at more definitive conclusions about differences in performance, with the aid of powerful statistical analyses.
- SUMMAC has also developed a number of software tools and resources for use in text summarization research, including a corpus of full-texts with summarizer sentence selection annotations, utilities for matching sentences, etc.

Overall, in the tasks we have examined here, SUMMAC has yielded the following specific results:

- **Effectiveness:** Variable-length summaries save time and are as accurate as full-text.
- **Intelligibility:** Over 80% of summaries produced by these systems are intelligible.
- **Technologies:** Even though a variety of different techniques are used, the systems are all fairly close in accuracy (three groups in the adhoc task, indistinguishable on the categorization task - which was hard even with full-text). The most accurate group on the adhoc task is not the fastest.

- **Compression Rates:** The systems which are relatively closer to the ideal in terms of accuracy and time on either task have mean compressions of less than 20%. 10% summaries are less accurate than variable-length ones. The F-scores in these tasks appear to vary in a logarithmic relation to the compression rate.
- **Acceptability:** There was quite a lot of variation across participants in terms of the percentage of their generic summaries judged "acceptable" (11%-71%) given the full-text, even though the summaries were very similar in their ability to indicate source content in the categorization task.
- **Relevant Passage Extraction:** Summarizers in the adhoc task are able to "dig out" passages from the document that appear relevant to the topic, which results in the SUMMAC assessors judging the document to be relevant where the TREC assessor (who lacked summarization tools) had not.
- **Subjects' judgments:** Subjects had medium to high confidence in their judgments most of the time. There is a moderate level of disagreement between subjects when they decide relevance of full-text in adhoc and categorization tasks.
- **Evaluation Methods:** In the Q&A task, a new methodology was established for evaluating user-focused summaries as surrogates for full documents. An automatic scoring program produced scores on several different measures that were strongly correlated with those that were manually produced.

12 Lessons Learned

This evaluation occurred at an appropriate time: the level of interest is high in the research community, the user community, and the commercial software industry, and there are no widely agreed-upon benchmarks against which to measure system performance. The evaluation was designed to address the utility of the technology from the user's perspective and was not intended to find the "best" summarization system available. We began with the belief that there wasn't a gold standard summary against which automatically generated summaries should be judged.

The type of evaluation we designed found some differences in system performance in the adhoc task. However, on the categorization task, which evaluated generic summaries, there was very little difference in system performance. It appears that summarizers all found the task to be equally "hard". In the dry-run evaluation, the topics in the categorization task were too distinct, resulting in very few instances of an incorrect judgment other than "none of the above". In the formal evaluation, we corrected that, and as a result, the task was harder, even using full-text alone. A somewhat different design methodology would be needed if the goal is to highlight such differences in performance on a categorization task. It is worth stressing that our reliance on TREC data for documents and topics, and our internal criteria for length, relevance, and non-overlap among test sets has resulted in the evaluation focusing mainly on mostly short newswire texts. This is somewhat unfortunate, as larger sized texts from a wider range of genres would have required more flexible summarization solutions, perhaps challenging the summarizers to a greater extent, and perhaps exposing more radical differences among systems.

The conduct of this evaluation relied heavily on existing human judgments that were created while evaluating information retrieval systems for TREC, comparing those judgments to ones made by summarization assessors. It would be better for many reasons to re-orient any subsequent evaluations to rely less on new input by human assessors. In an era of declining resources, labor-intensive evaluations such as these are costly to set up and execute. They don't provide the type of independent, repeatable measures that allow system developers to continually test and improve their systems. One rather unexpected result was the extreme level of variability in judgments. Assessors often disgreed with each other regarding the relevance of a particular full-text to a topic. While some level of disagreement was expected, it was much more pronounced in this evaluation than in others in the literature, including a very recent TREC experiment [Voorhees 1998]. There are a number of reasons why this might be so, e.g., differences in the number of subjects, and the instructions given to subjects, etc., but further analysis of the data, including more detailed examination of feedback and comments, would be needed here.

It is important to emphasize for future evaluations the availability of open, shareable resources not subjected to copyright or proprietary restrictions. One prime example of this type of resource sharing is the interest in the co-reference engine developed by the University of Pennsylvania. It is also useful to provide access to summaries generated by candidate systems, and any performance measures or indicators of accuracy of the summaries. Creation and sharing of corpora to support research in summarization are also very important. An open evaluation environment can have a dramatic impact in accelerating progress in text summarization technology.

13 Future Directions

This evaluation has focused mainly on an extrinsic evaluation in tasks of relevance assessment, with more exploratory intrinsic evaluations. These only begin to explore the space of evaluation methods used to determine the usefulness of summarization. The evaluation here has been a black-box evaluation confined to single-document summaries of mostly short, newswire texts. The field of text summarization is currently experiencing unprecendented attention in terms of research and development activities, and we expect there will be considerable changes in the types of summaries available in the future, including generated abstracts, multi-document and multi-lingual summaries. Increased research is also expected to result in increased availability of test collections in different genres, including scientific and technical text, as well as larger sized documents and collections. This may require more powerful capabilities for compression, challenging some of the extraction-based methods in use today. To the extent that systems can automatially create these new varieties of summaries (not to mention summaries such as reviews, digests, previews, reports, etc., that humans routinely create), additional challenges for evaluation will be created.

Acknowledgments

The authors wish to thank John Burger, Barbara Gates, Glenn Iwerks, Sara Shelton, and Sandra Wagner, as well as 51 experimental subjects. We are also grateful to the Linguistic Data Consortium for making the TREC documents available to us, and to the National Institute of Standards and Technology for providing TREC data and the initial version of the ASSESS tool. Last but not least, we are thankful to all the participants, whose technology made this evaluation possible and worthwhile in the first place.

References

[Brandow et al. 1994] Brandow, R., K. Mitze, and L. Rau. 1994. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31(5).

[Carletta et al. 1997] Carletta, J., A. Isard, S. Isard, J. C. Jowtko, G. Doherty-Sneddon, and A. H. Anderson. 1997. The Reliability of a Dialogue Structure Coding Scheme. *Computational Linguistics*, 23, 1, 13-32.

[Cohen 1969] Cohen, J. 1969. Statistical Power Analysis for the Behavioral Sciences. Academic Press, NY.

[Day et al. 1997] Day, D., J. Aberdeen, L. Hirschman, R. Kozierok, P. Robinson, and M. Vilain. 1997. Mixed-Initiative Development of Language Processing Systems. *in Fifth Conference on Applied Natural Language Processing, Association for Computational Linguistics, Washington, D.C., pp. 348-355.*

[Day et al. 1998] Day, D., J. Aberdeen, S. Caskey, L. Hirschman, P. Robinson, and M. Vilain. 1998. Alembic Workbench Corpus Development Tool. http://www.mitre.org/technology/alembic-workbench. *in First International Conference on Language Resources and Evaluation, European Language Resources Association, Granada, Spain, 1021-1028.*

[Edmundson 1969] Edmundson, H.P. 1969. New methods in automatic abstracting. *The Association for Computing Machinery*, 16(2).

[Hahn and Mani 1998] Hahh, U., and Mani, I. 1998. AAAI'98 Tutorial on Automatic Text Summarization. http://www.aaai.org/Conferences/National/1998/Tutorials/sp1.html.

[Harman and Voorhees 1996] Harman, D.K. and E.M. Voorhees. 1996. The fifth text retrieval conference (trec-5). *National Institute of Standards and Technology NIST SP 500-238.*

[Hovy and Marcu 1998] Hovy, E., and Marcu, D. 1998. COLING-ACL'98 Tutorial on Text Summarization. http://www.isi.edu/ marcu/coling-acl98-tutorial.html.

[Jing et al. 1998] Jing, H., R. Barzilay, K. McKeown, and M. Elhadad. 1998. Summarization evaluation methods: Experiments and analysis. *in Working Notes of the AAAI Spring Symposium on Intelligent Text Summarization, Spring 1998, Technical Report, AAAI, 1998.*

[Kirk 1968] Kirk, R.E. 1968. Experimental Design: Procedures for the Behavioral Sciences. Wadsworth, Belmont, CA.

[Kupiec et al. 1995] Kupiec, J. Pedersen, and F. Chen. 1995. A trainable document summarizer. *Proceedings of the 18th ACM SIGIR Conference (SIGIR'95).*

[Mani and Bloedorn 1999] Mani, I. and E. Bloedorn. 1999. Summarizing similarities and differences among related documents. *INFORMATION RETRIEVAL*, 1(1).

[Mani and Maybury 1999] Mani, I. and M. Maybury, eds. 1999. *Advances in Automatic Text Summarization.* MIT Press.

[Marcu 1997] Marcu, D. 1997. From discourse structures to text summaries. *in Mani, I. and Maybury, M., eds., Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization.*

[McKeown and Radev 1995] McKeown, K. and D. Radev. 1995. Generating summaries of multiple news articles. *Proceedings of ACM-SIGIR'95.*

[Minel et al. 1997] Minel, J-L., S. Nugier, and G. Piat. 1997. How to appreciate the quality of automatic text summarization. In Mani, I. and Maybury, M., eds., *Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization.*

[Morris et al. 1992] Morris, A., G. Kasper, and D. Adams. 1992. The Effects and Limitations of Automatic Text Condensing on Reading Comprehension Performance. *Information Systems Research*, 3(1).

[MUC-6 1995] MUC-6. 1995. Proceedings of the sixth message understanding conference (muc-6). Morgan-Kaufmann, San Francisco.

[Paice 1990] Paice, C. 1990. Constructing literature abstracts by computer: Techniques and prospects. *Information Processing and Management*, 26(1).

[Pollock and Zamora 1975] Pollock, J.J. and A. Zamora. 1975. Automatic abstracting research at chemical abstracts service. *Chemical Information and Computer Sciences*, 15(4).

[Press et al. 1988] Press, W.H., Flannery, B. P., Teukolsky, S. A., Vetterling, W. T.. 1988. *Numerical Recipes in C.* Cambridge University Press.

[Rath et al. 1961] Rath, G.J., A. Resnick, and T.R. Savage. 1961. The formation of abstracts by the selection of sentences. *American Documentation*, 12(2).

[Salton et al. 1997] Salton, G., A. Singhal, M. Mitra, and C. Buckley. 1997. Automatic Text Structuring and Summarization. *Information Processing and Management*, 33(2).

[Siegel 1988] Siegel, S., and N. Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill Inc., second edition.

[Sparck-Jones 1998] Sparck-Jones, K. 1998. Summarizing: Where are we now? where should we go? *Mani, I. and Maybury, M., eds., Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization.*

[Tombros et al. 1998] Tombros, A., and M. Sanderson. 1998. Advantages of query biased summaries in information retrieval. *in Proceedings of the 21st ACM SIGIR Conference (SIGIR'98)*, 2-10.

[Voorhees 1998] Voorhees, Ellen M. 1998. Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-98)*, Melbourne, Australia. 315–323.

14 Appendix

To give a flavor as to what the summaries look like, here we show a sample topic, a sample fixed-length adhoc summary, and a sample variable-length adhoc summary of a document.

```
<top> <num> Number: 257
<title> Topic: Cigarette Consumption
<desc> Description:
What data is available on cigarette consumption by country?
<narr> Narrative:
If cigarette smoking is a causative factor in
lung cancer, then countries with higher cigarette
consumption per capita might experience a higher incidence
of lung cancer.  This topic would provide basic data for
such a comparison.
</top>
```

Figure 19: Topic 257

```
<DOC>
<TASKTYPE>adhoc</TASKTYPE>
<SUMMARYTYPE>fixed</SUMMARYTYPE>
<QNUM>257</QNUM>
<DOCNO>FT941-1421</DOCNO>
<TEXT>

In some cancers there is an obvious link with risk factors. The north-south
gradient in lung cancer is caused mainly by the greater prevalence of
smoking...

</TEXT>
</DOC>
```

Figure 20: A Sample Fixed-Length (10%) Adhoc Summary for Document FT941-1421 and Topic 257.

```
<TASKTYPE>adhoc</TASKTYPE>
<SUMMARYTYPE>best</SUMMARYTYPE>
<QNUM>257</QNUM>
<DOCNO>FT941-1421</DOCNO>
<TEXT>
Striking regional variations are revealed by the first atlas of cancer
incidence in England and Wales, published yesterday, Clive Cookson
writes.
They compared these with the geographical distributions of risk
factors such as smoking and occupation.
The north-south gradient in lung cancer is caused mainly by the
greater prevalence of smoking in the north.
The fact that southern women have fewer children on average may be a
partial explanation, Drs Silver and Swerdlow say, because they have
higher levels of the hormones related to these cancers.
</TEXT>
</DOC>
```

Figure 21: A Sample Variable-Length Adhoc Summary for Document FT941-1421 and Topic 257.

```
<DOC>
<DOCNO>FT941-1421</DOCNO>
<PROFILE>_AN-ECYC5ACHFT</PROFILE>
<DATE>940325
</DATE>
<HEADLINE>
FT  25 MAR 94 / Cancer map shows regional contrasts
</HEADLINE>
<BYLINE>
   By CLIVE COOKSON
</BYLINE>
<TEXT>
Striking regional variations are revealed by the first atlas of cancer
incidence in England and Wales, published yesterday, Clive Cookson writes.
The atlas, commissioned by the Cancer Research Campaign, shows that lung
cancer, the most common form of the disease in men, is much more prevalent
in the north than in the south. The reverse is true for breast cancer, the
most common cancer in women.
Dr Isabel Silva and Dr Anthony Swerdlow of the London School of Hygiene and
Tropical Medicine analysed information about 3m new cancer patients between
1968 and 1985 to give a county-by-county variation in cancer risks. They
compared these with the geographical distributions of risk factors such as
smoking and occupation. The figures in the map above are an index of the
number of new cases in each county over the period.
In some cancers there is an obvious link with risk factors. The north-south
gradient in lung cancer is caused mainly by the greater prevalence of
smoking in the north. The authors say that greater industrial exposures to
smoke, dust and toxic fumes in the north are not sufficient to account for
the regional differences.
Malignant melanoma, the most virulent skin cancer, has a strong south/north
gradient - someone living on the south coast is three times more likely to
suffer than someone in northern England. There is a clear correlation with
hours of sunshine.
The reason why breast and ovarian cancers are more common in the south is
not obvious. The fact that southern women have fewer children on average may
be a partial explanation, Drs Silver and Swerdlow say, because they have
higher levels of the hormones related to these cancers.
Atlas of Cancer Incidence in England &amp; Wales. Oxford University Press.
Pounds 170.
</TEXT>
</DOC>
```

Figure 22: A Sample Full-Text Document. Only the $< TEXT >$ portion was summarized and evaluated. To fit this document on one page, trailing tags after the TEXT portion have been omitted.