The TIPSTER SUMMAC Text Summarization Evaluation

Inderjeet Mani
David House
Gary Klein
Lynette Hirschman*
The MITRE Corporation
11493 Sunset Hills Rd.
Reston, VA 22090
USA

Therese Firmin
Department of Defense
9800 Savage Rd.
Ft. Meade, MD 20755
USA

Beth Sundheim
SPAWAR Systems Center
Code D44208
53140 Gatchell Rd.
San Diego, CA 92152
USA

Abstract

The TIPSTER Text Summarization Evaluation (SUMMAC) has established definitively that automatic text summarization is very effective in relevance assessment tasks. Summaries as short as 17% of full text length sped up decision-making by almost a factor of 2 with no statistically significant degradation in F-score accuracy. SUMMAC has also introduced a new intrinsic method for automated evaluation of informative summaries.

1 Introduction

In May 1998, the U.S. government completed the TIPSTER Text Summarization Evaluation (SUMMAC), which was the first large-scale, developer-independent evaluation of automatic text summarization systems. The goals of the SUMMAC evaluation were to judge individual summarization systems in terms of their usefulness in specific summarization tasks and to gain a better understanding of the issues involved in building and evaluating such systems.

1.1 Text Summarization

Text summarization is the process of distilling the most important information from a set of sources to produce an abridged version for particular users and tasks (Maybury 1995). Since abridgment is crucial, an important parameter to summarization is the level of *compression* (ratio of summary length to source length) desired. 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 source content (the *indicative* function). In addition, summaries can also be used to stand in place of the source (the *informative* function). They can even offer a critique of the source (the *evaluative* function) (Sparck-Jones 1998). Often, summaries are tailored to a reader's interests and expertise, yielding *topic-related* summaries, or else they can be aimed at a broad readership community, as in the case of *generic* summaries. It is also useful to distinguish between summaries which are *extracts* of source material, and those which are *abstracts* containing new text generated by the summarizer.

1.2 Summarization Evaluation Methods

Methods for evaluating text summarization can be broadly classified into two categories.

The first, an *intrinsic* (or normative) evaluation, judges the quality of the summary directly based on analysis in terms of some set of norms. This can involve user judgments of fluency of the summary (Minel et al. 1997), (Brandow et al. 1994), coverage of stipulated "key/essential ideas" in the source (Paice 1990), (Brandow et al. 1994), or similarity to an "ideal" summary, e.g., (Edmundson 1969), (Kupiec et al. 1995).

The problem with matching a system summary against an ideal summary is that the ideal summary is hard to establish. There can be a large number of generic and topic-related abstracts that could summarize a given document. Also, there have been several reports of low inter-annotator agreement on sentence extracts, e.g., (Rath et al. 1961), (Salton et al. 1997), although judges may agree more on the most important sentences to include (Jing et al. 1998).

The second category, an *extrinsic* evaluation, judges the quality of the summarization based on how it affects the completion of some other task. There have been a number of extrinsic evaluations, including question-answering and comprehension tasks, e.g., (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 1997), (Jing et al.

*202 Burlington Rd., Bedford, MA 01730

1998), (Tombros et al. 1998), (Brandow et al. 1994).

1.3 Participant Technologies

Sixteen systems participated in the SUMMAC Evaluation: 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), the University of Southern California-Information Sciences Institute (ISI), 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 Russian Center for Information Research (CIR), and the National Taiwan University (NTU). Table 1 offers a high-level summary of the features used by the different participants. Most participants confined their summaries to extracts of passages from the source text; TextWise, however, extracted combinations of passages, phrases, named entities, and subject fields. Two participants modified the extracted text: Penn replaced pronouns with coreferential noun phrases, and Penn and NMSU both shortened sentences by dropping constituents.

2 SUMMAC Summarization Tasks

In order to address the goals of the evaluation, two main extrinsic evaluation tasks were defined, based on activities typically carried out by information analysts in the U.S. Government. In the *adhoc task*, the focus was on indicative summaries which were *tailored to a particular topic*. This task relates to the real-world activity of an analyst conducting full-text searches using an IR system to quickly determine the relevance of a retrieved document. Given a document (which could be a summary or a full-text source - the subject was not told which), and a topic description, the human subject was asked to determine whether the document was relevant to the topic. The accuracy of the subject's relevance assessment decision was measured in terms of "ground-truth" judgments of the full-text source relevance, which were separately obtained from the Text Retrieval (TREC) (Harman and Voorhees 1996) conferences. Thus, an indicative summary would be "accurate" if it accurately reflected the relevance or irrelevance of the corresponding source.

In the *categorization task*, the evaluation sought to find out whether a *generic* summary could effectively present enough information to allow an analyst to quickly and correctly categorize a document. Here the topic was not known to the summarization system. Given a document, which could be a generic summary or a full-text source (the subject was not told which), the human subject would choose a single category out of five categories (each of which had an associated topic description) to which the document was relevant, or else choose "none of the above".

The final task, a *question-answering task*, was intended to support an information analyst writing a report. This involved an *intrinsic* evaluation where a topic-related summary for a document was evaluated in terms of its "informativeness", namely, the degree to which it contained answers found in the source document to a set of topic-related questions.

3 Data Selection

In the adhoc task, 20 topics were selected. For each topic, a 50-document subset was created from the top 200 ranked documents retrieved by a standard IR system. For the categorization task, only 10 topics were selected, with 100 documents used per topic. For both tasks, the subsets were constructed such that 25%-75% of the documents were relevant to the topic, with full-text documents being 2000-20,000 bytes (300-2700 words) long, so that they were long enough to be worth summarizing but short enough to be read within the time-frame of the experiment.

The documents were all newspaper sources, the vast majority of which were news stories, but which also included sundry material such as letters to the editor. Reliance on TREC data for documents and topics, and internal criteria for length, relevance, and non-overlap among test sets, resulted in the evaluation focusing mostly on short newswire texts. We recognize that larger-sized texts from a wider range of genres might challenge the summarizers to a greater extent.

In each task, participants submitted two summaries: a fixed-length (S1) summary limited to 10% of the length of the source, and a summary which was not limited in length (S2).

4 Experimental Hypotheses and Method

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

Participant	tf	loc	disc	coref	co-occ	syn
BT	+	+	-	+	+	-
CGI/CMU	+	+	-	-	+	-
CIR	+	+	-	-	-	+
Cornell/SabIR	+	-	-	-	+	-
GE	+	+	+	+	+	-
IA	+	-	-	-	+	-
IBM	+	+	-	-	-	-
ISI	+	+	-	-	-	+
LN	+	-	-	-	+	-
NMSU	+	-	+	+	-	-
NTU	+	-	+	+	-	-
Penn	-	+	-	+	-	-
SRA	+	+	-	+	-	+
Surrey	+	-	+	-	+	+
TextWise	+	-	-	+	+	+
UMass	+	-	-	-	+	-

Table 1: Participant Summarization Features. tf: term frequency; loc: location; disc:discourse (e.g., use of discourse model); coref: coreference; co-occ: co-occurrence; syn: synonyms.

Ground Truth	Subject's Judgment	
	Relevant	Irrelevant
Relevant is True	TP	FN
Irrelevant is True	FP	TN

Table 2: Adhoc Task Contingency Table. TP=true positive, FP = false positive, TN= true negative, FN=false negative.

Ground Truth	Subject's Judgment		
	X	Y	None
X is True	TP	FN	FN
None is True	FP	FP	TN

Table 3: Categorization Task Contingency Table. X and Y are distinct categories other than None-of-the- above, represented as None.

The first test was a *summarization condition test*: to determine whether subjects' relevance assessment performance in terms of time and accuracy was affected by different conditions: full-text (F), fixed-length summaries (S1), variable-length summaries (S2), and baseline summaries (B). The latter were comprised of the first 10% of the body of the source text.

The second test was a *participant technology test*: to compare the performance of different participants' systems.

The third test was a *consistency test*: to determine how much agreement there was between subjects' relevance decisions based on showing them only full-text versions of the documents from the main adhoc and categorization tasks. In the adhoc and categorization tasks, the 1000 documents assigned to a subject for each task were allocated among F, B, S1, and S2 conditions through random selection without replacement (20 F, 20 B, 480 S1, and 480 S2[1]). For the consistency tasks, each subject was assigned full-text versions of the same 1000 documents. In all tasks, the presentation order was varied among subjects. The evaluation used 51 professional information analysts as subjects, each of whom took approximately 16-20 hours. The main adhoc task used 21 subjects, the main categorization 24 subjects; the consistency adhoc task had 14 subjects, the consistency categorization 7 subjects (some subjects from the main task also did a different consistency task). The subjects were told they were working with documents that included summaries, and that their goal, on being presented with a topic-document pair, was to examine each document to determine if it was relevant to the topic. The contingency tables for the adhoc and categorization tasks are shown in Tables 2 and 3.

We used the following aggregate accuracy metrics:

$$Precision = TP/(TP + FP) \qquad (1)$$

$$Recall = TP/(TP + FN) \qquad (2)$$

$$Fscore = 2 * Precision * Recall/(Precision + Recall) \qquad (3)$$

5 Results: Adhoc and Categorization Tasks

5.1 Performance by Condition

In the adhoc task, summaries at compressions as low as 17% of full text length were not significantly

[1] This distribution assures sufficient statistical sensitivity for expected effect sizes for both the summarization condition and the participant technology tests.

Condition	Time	Time SD	F-score	TP	FP	FN	TN	P	R
F	58.89	56.86	.67	.38	.08	.26	.28	.83	.22
S2	33.12	36.19	.64	.35	.08	.28	.28	.80	.23
S1	19.75	26.96	.53	.27	.07	.35	.31	.79	.19
B	23.15	21.82	.42	.18	.05	.41	.35	.81	.12

Table 4: Adhoc Time and Accuracy by Condition. TP, FP, FN, TN are expressed as percentage of totals observed in all four categories. All time differences are significant except between B and S1 (HSD=9.8). All F-score differences are significant, except between F (Full-Text) and S2 (HSD=.10). Precision (P) differences aren't significant. All Recall (R) differences between conditions are significant, except between F and S2 (HSD=.12). "SD" = standard deviation.

Condition	Time	Time SD	F-score	TP	FP	FN	TN	P	R
F	43.11	52.84	.50	24.3	13.3	28.5	33.9	.63	.45
S2	43.15	42.16	.50	19.3	10.5	36.9	33.3	.68	.42
S1	25.48	29.81	.43	27.1	10.7	30.9	31.3	.68	.34
B	27.36	30.35	.03	7.5	11.9	52.5	28.1	.04	.02

Table 5: Categorization Time and Accuracy by Condition. Here TP, FP, FN, TN are expressed as percentage of totals in all four categories. All time differences are significant except between F and S2, and between B and S1 (HSD=15.6).Only the F-score of B is significantly less than the others (HSD=.09). Precision (P) and Recall (R) of B is significantly less than the others: HSD(Precision)=.11; HSD(Recall)=.11.

different in accuracy from full text (Table 4), while speeding up decision-making by almost a factor of 2 (33.12 seconds per decision average time for S2 compared to 58.89 for F in 4). Tukey's Honestly Significant Difference test (HSD) is used to compare multiple differences[2].

In the categorization task, the F-score on full-text was only .5, suggesting the task was very hard. Here summaries at 10% of the full-text length were not significantly different in accuracy from full-text (Table 5) while reducing decision time by 40% compared to full text (25.48 seconds for S1 compared to 43.11 for F in 5). The very low F-scores for the Bs can be explained by a bug which resulted in the same 20 relatively less-effective B summaries being offered to each subject. However, in this task, summaries longer than 10% of the full text, while not significantly different in accuracy from full-text, did not take less time than full-text. In both tasks, the main accuracy losses in summarization came from FNs, not FPs, indicating the summaries were missing topic-relevant information from the source.

5.2 Performance by Participant

In the adhoc task, the systems were all very close in accuracy for both summary types (Table 6). Three groups of systems were evident in the adhoc S2 F-score accuracy data, as shown in Table 8. Interestingly, the Group I systems both used only term frequency and co-occurrence (Table 1), in

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

Table 8: Adhoc Accuracy: Participant Groups for S2 summaries. Groups I and III are significantly different in F-score (albeit with a small effect size). Accuracy differences within groups and between Group II and the others are not significant.

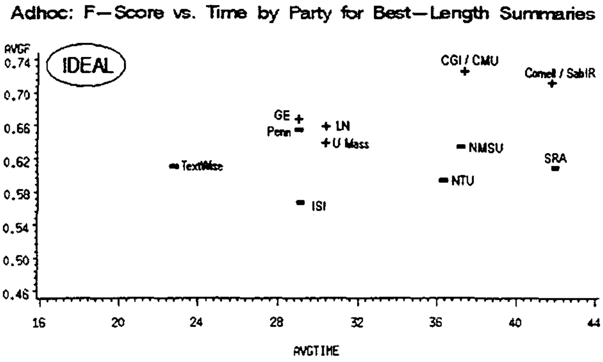

Figure 1: Adhoc F-score versus Time by Participant (variable-length summaries). HSD(F-score) is 0.13. HSD(Time) = 12.88. Decisions based on summaries from GE, Penn, and TextWise are significantly faster than based on SRA and Cornell/SabIR.

[2] The significance level $\alpha < .05$ throughout this paper, unless noted otherwise.

System	S2			S1		
	P	R	F-score	P	R	F-score
CGI/CMU	.82	.66	.72	.76	.52	.60
Cornell/SabIR	.78	.67	.70	.79	.47	.56
GE	.78	.60	.67	.77	.45	.55
LN	.78	.58	.65	.81	.45	.55
Penn	.81	.57	.65	.76	.45	.53
UMass	.80	.54	.63	.81	.47	.56
NMSU	.80	.54	.63	.80	.40	.52
TextWise	.81	.51	.61	.79	.41	.52
SRA	.82	.49	.60	.79	.37	.48
NTU	.80	.49	.59	.82	.34	.46
ISI	.80	.46	.56	.82	.36	.47

Table 6: Adhoc Accuracy by Participant. For variable-length: Precision (P) differences aren't significant; CGI/CMU and Cornell/SabIR are significantly different from SRA, NTU, and ISI in Recall (R) (HSD=0.17) and from ISI in F-score (HSD=0.13). For fixed-length, no significant differences on any of the measures.

System	S2			S1		
	P	R	F-score	P	R	F-score
CIR	.71	.47	.54	.68	.35	.43
IBM	.68	.47	.51	.63	.37	.44
NMSU	.69	.46	.51	.69	.34	.43
Surrey	.69	.43	.51	.69	.31	.39
Penn	.70	.42	.50	.66	.29	.38
ISI	.71	.42	.49	.71	.35	.44
IA	.69	.42	.49	.67	.33	.41
BT	.63	.43	.48	.70	.33	.41
NTU	.66	.41	.48	.68	.33	.43
SRA	.65	.42	.48	.73	.37	.45
LN	.68	.41	.47	.68	.37	.45
Cornell/SabIR	.66	.40	.47	.62	.36	.42
GE	.69	.40	.47	.69	.33	.42
CGI/CMU	.74	.39	.47	.69	.33	.42

Table 7: Categorization Accuracy by Participant. No significant differences on any of the measures.

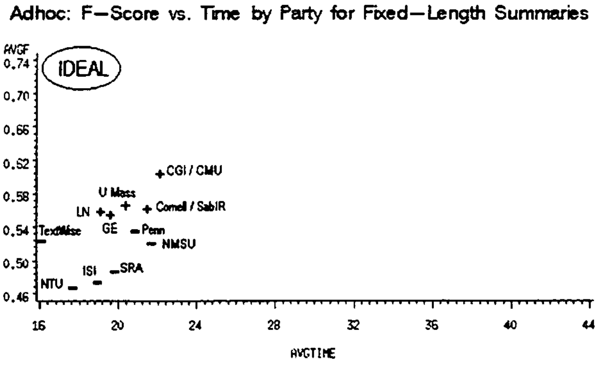

Figure 2: Adhoc F-score versus Time by Participant (fixed-length summaries). No significant differences in F-score, or in Time.

particular, exploiting similarity computations between text passages. For the S2 summaries (Figure 1), the Group I systems (average compression 25% for CGI/CMU and 30% for Cornell/SabIR) were not the fastest in terms of human decision time; in terms of both accuracy and time, TextWise, GE and Penn (equivalent in accuracy) were the closest in terms of Cartesian distance from the ideal performance. For S1 summaries (Figure 2), the accuracy and time differences aren't significant. Finally, clustering the systems based on *degree of overlap between the sets of sentences they extracted for summaries judged TP* resulted in CGI/CMU, GE, LN, UMass, and Cornell/SabIR clustering together on both S1 and S2 summaries. It is striking that this cluster, shown with the "+" icon in Figures 1 and 2, corresponds to the systems with the highest F-scores, all of whom, with the exception of GE, used similar features in analysis (Table 1).

In the categorization task, by contrast, the 14 participating systems[3] had no significant differences in F-score accuracy whatsoever (Table 7,

[3] Note that some participants participated in only one of the two tasks.

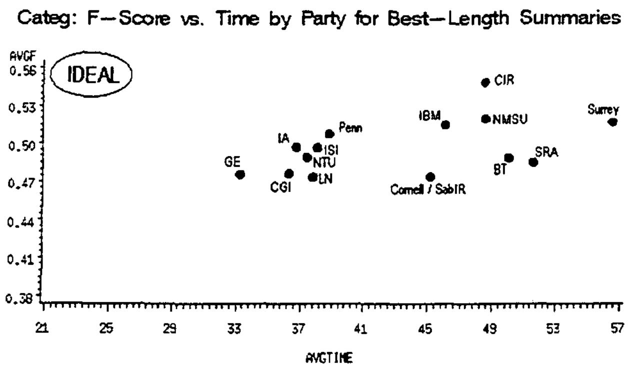

Figure 3: Categorization F-score versus Time by Participant (variable-length summaries). F-scores are not significantly different. HSD(Time) = 17.23. GE is significantly faster than SRA and Surrey. The latter two are also significantly slower than Penn, ISI, LN, NTU, IA, and CGI/CMU.

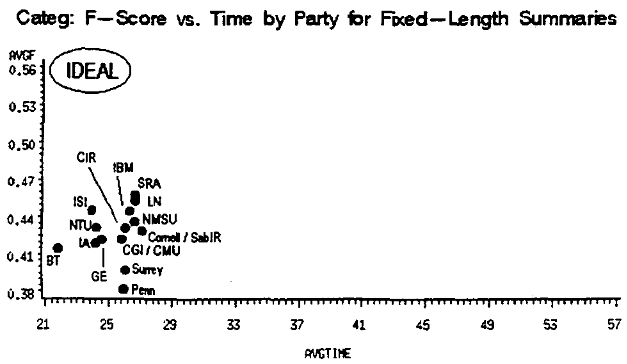

Figure 4: Categorization F-score versus Time by Participant (fixed-length summaries). F-scores are not significantly different, and neither are time differences.

Figures 3 and 4). In this task, in the absence of a topic, the statistical salience systems which performed relatively more accurately in the adhoc task had no advantage over the others, and so their performance more closely resemble that of other systems. Instead, *the systems more often relied on inclusion of the first sentence of the source* - a useful strategy for newswire (Brandow et al. 1994): the generic (categorization) summaries had a higher percentage of selections of first sentences from the source than the adhoc summaries (35% of S1 and 41% of S2 for categorization, compared to 21% S1 and 32% S2 for adhoc). We may surmise that in this task, where performance on full-text was hard to begin with, the systems were all finding the categorization task equally hard, with no particular technique for producing generic summaries standing out.

5.3 Agreement between Subjects

As indicated in Table 9, the unanimous agreement of just 16.6% and 19.5% in the adhoc and categorization tasks respectively is low: the agreement data has Kappa (Carletta et al. 1997) of .38 for adhoc and .29 for categorization[4]. The adhoc pairwise and 3-way agreement (i.e., agreement between groups of 3 subjects) is consistent with a 3-subject "dry-run" adhoc consistency task carried out earlier. However, it is much lower than reported in 3-subject adhoc experiments in TREC (Harman and Voorhees 1996). One possible explanation is that in contrast to our subjects, TREC subjects had years of experience in this task. It is also possible that our mix of documents had fewer obviously relevant or obviously irrelevant documents than TREC. However, as (Voorhees 1998) has shown in her TREC study, system performance rankings can remain relatively stable even with lack of agreement in relevance judgments. Further, (Voorhees 1998) found, when only relevant documents were considered (and measuring agreement by intersection over union), 44.7% pairwise agreement and 30.1% 3-way agreement with 3 subjects, which is comparable to our scores on this latter measure (52.9% pairwise, 36.9% 3-way on adhoc, 45.9% pairwise, 29.7% 3-way on categorization).

6 Question-answering (Q&A) task

In this task, the summarization system, given a document and a topic, needed to produce an informative, topic-related summary that contained the answers found in that document to a set of topic-related questions. These questions covered "obligatory" information that had to be provided in any document judged relevant to the topic. For example, for a topic concerning prison overcrowding, a topic-related question would be "What is the name of each correction facility where the reported overcrowding exists?"

6.1 Experimental Design

The topics we chose were a subset of the 20 adhoc TREC topics selected. For each topic, 30 relevant documents from the adhoc task corpus were chosen as the source texts for topic-related summarization. The principal tasks of each evaluator (one evaluator per topic, 3 in all) were to prepare the questions and answer keys and to score the

[4] Dropping two outlier assessors in the categorization task - the fastest and the slowest - resulted in the pairwise and three-way agreement going up to 69.3% and 54.0% respectively, making the agreement comparable with the adhoc task.

Task	Pairwise	3-way	All 7	All 14
Adhoc	69.1	53.7	NA	16.6
Categorization	56.4	50.6	19.5	NA
Adhoc Dry-Run	72.7	59.1	NA	NA
TREC	88.0	71.7	NA	NA

Table 9: Percentage of decisions subjects agreed on when viewing full-text (consistency tasks).

system summaries. To construct the answer key, each evaluator marked off any passages in the text that provided an answer to a question (example shown in Table 10).

The summaries generated by the participants (who were given the topics and the documents to be summarized, but not the questions) were scored against the answer key. The evaluators used a common set of guidelines for writing questions, creating answer keys, and scoring summaries that were intended to minimize variability across evaluators in the methods used[5].

Eight of the adhoc participants also submitted summaries for the Q&A evaluation. Thirty summaries per topic were scored against the answer keys.

6.2 Scoring

Each summary was compared manually to the answer key for a given document. If a summary contained a passage that was tagged in the answer key as the only available answer to a question, the summary was judged Correct for that question as long as the summary provided sufficient context for the passage; if there was insufficient context, the summary was judged Partially Correct. If needed context was totally lacking or was misleading, or if the summary did not contain the expected passage at all, the summary was judged Missing for that question. In the case where (a) the answer key contained multiple tagged passages as answer(s) to a single question and (b) the summary did not contain all of those passages, assessors applied additional scoring criteria to determine the amount of credit to assign.

Two accuracy metrics were defined, *ARL* (Answer Recall Lenient) and *ARS* (Answer Recall Strict):

$$ARL = (n1 + (.5 * n2))/n3 \quad (4)$$

$$ARS = n1/n3 \quad (5)$$

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. A third measure, *ARA* (Answer Recall Average), was defined as the average of *ARL* and *ARS*.

6.3 Results

Figure 5 shows a plot of the *ARA* against compression. The "model" summaries were sentence-extraction summaries created by the evaluators from the answer keys but not used to evaluate the summaries. For the machine-generated summaries, the highest *ARA* was associated with the least reduction (35-40% compression). The systems which were in Group I in accuracy on the adhoc task, CGI/CMU and Cornell/SabIR, were at the top of the *ARA* ordering of systems on topics 257 and 271. The participants' human-evaluated *ARA* scores were strongly correlated with scores computed by a program from Cornell/SabIR which measured overlap between summaries and answers in the key (Pearson $r > .97$, $\alpha < 0.0001$). The Q&A evaluation is therefore promising as a new method for automated evaluation of informative summaries.

7 Conclusions

SUMMAC has established definitively in a large-scale evaluation that automatic text summarization is very effective in relevance assessment tasks. Summaries at relatively low compression rates (summaries as short as 17% of source length for adhoc, 10% for categorization) allowed for relevance assessment almost as accurate as with full-text (5% degradation in F-score for adhoc and 14% degradation for categorization, both degradations not being statistically significant), while reducing decision-making time by 40% (categorization) and 50% (adhoc). Analysis of feedback forms filled in after each decision indicated that the intelligibility of present-day machine-generated summaries is high, due to use of sentence extraction and coherence "smoothing"[6].

The task of topic-related summarization, when limited to passage extraction, can be characterized as a passage ranking problem, and as such lends itself very well to information retrieval tech-

[5] We also had each of the evaluators score a portion of each others' test data; the scores across evaluators were very similar, with one exception.

[6] On the adhoc task, 99% of F were judged "intelligible", as were 93% S2, 96% B, 83% S1; similar data for categorization.

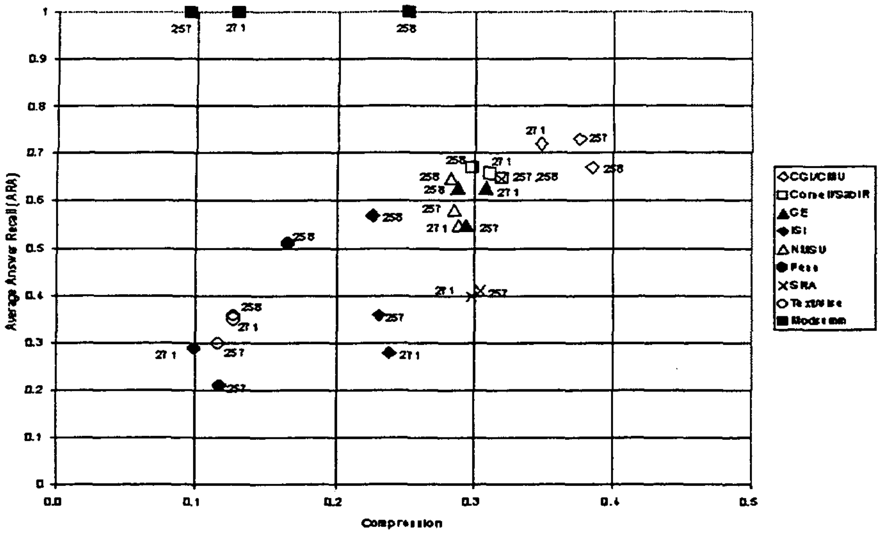

Figure 5: ARA versus Compression by Participant. "Modsumms" are model summaries.

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?

```
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. ....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>. .....
```

Table 10: Q&A Topic 258, topic-related questions, and part of a relevant source document showing answer key annotations.

niques. Summarizers that performed most accurately in the adhoc task used statistical passage similarity and passage ranking methods common in information retrieval. Overall, the most accurate systems in this task used similar features and had similar sentence extraction behavior.

However, for the generic summaries in the categorization task (which was hard even for humans with full-text), in the absence of a topic, the summarization methods in use by these systems were indistinguishable in accuracy. Whether this suggests an inherent limitation to summarization methods which produce extracts of the source, as opposed to generating abstracts, remains to be seen.

In future, text summarization evaluations will benefit greatly from the availability of test sets covering a wider variety of genres, and including much longer documents. The extrinsic and intrinsic evaluations reported here are also relevant to the evaluation of other NLP technologies where there may be many potentially acceptable outputs (e.g., machine translation, text generation, speech synthesis).

Acknowledgments

The authors wish to thank Eric Bloedorn, John Burger, Mike Chrzanowski, Barbara Gates, Glenn Iwerks, Leo Obrst, 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.

References

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

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.

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

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.*

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.*

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

Mani, I. and E. Bloedorn. 1997. Multi-document Summarization by Graph Search and Merging. *Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI-97), Providence, RI, July 27-31, 1997*, 622-628.

Maybury, M. 1995. Generating Summaries from Event Data. *Information Processing and Management, 31,5, 735-751.*

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

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

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

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

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, 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, 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.