Natural Language Engineering **8** (1): 43–68. © 2002 Cambridge University Press
DOI: 10.1017/S1351324901002741 Printed in the United Kingdom

SUMMAC: a text summarization evaluation

INDERJEET MANI,† GARY KLEIN,
DAVID HOUSE, 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

(*Received 7 June 2000; revised 24 May 2001*)

Abstract

The TIPSTER Text Summarization Evaluation (SUMMAC) has developed several new extrinsic and intrinsic methods for evaluating summaries. It has established definitively that automatic text summarization is very effective in relevance assessment tasks on news articles. 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 accuracy. Analysis of feedback forms filled in after each decision indicated that the intelligibility of present-day machine-generated summaries is high. Systems that performed most accurately in the production of indicative and informative topic-related summaries used term frequency and co-occurrence statistics, and vocabulary overlap comparisons between text passages. However, in the absence of a topic, these statistical methods do not appear to provide any additional leverage: in the case of generic summaries, the systems were indistinguishable in accuracy. The paper discusses some of the tradeoffs and challenges faced by the evaluation, and also lists some of the lessons learned, impacts, and possible future directions. The evaluation methods used in the SUMMAC evaluation are of interest to both summarization evaluation as well as evaluation of other 'output-related' NLP technologies, where there may be many potentially acceptable outputs, with no automatic way to compare them.

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 US government completed the TIPSTER Text Summarization Evaluation (SUMMAC[1]),

† Also at the Department of Linguistics, Georgetown University, Washington, DC 20037, USA.
‡ 202 Burlington Rd., Bedford, MA 01730, USA.
[1] *SUMMAC* stands for First *Summa*rization *C*onference.

https://doi.org/10.1017/S1351324901002741 Published online by Cambridge University Press

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

Automatic summarization is the process of distilling the most important information from a source or set of sources to produce an abridged version for particular users and tasks (Maybury 1995). Examples of naturally occurring summaries include news headlines, scientific abstracts, movie previews and reviews, meeting minutes, TV guides, weather bulletins, drastically condensed books, etc. Since abridgment is crucial, an important parameter to summarization is the level of *compression* (ratio of summary length to source length) desired.

A traditional distinction (Borko and Bernier 1975) is between *indicative* summaries, which provide a reference or 'alerting' function for selecting documents for more in-depth reading, and *informative* summaries, which can stand in place of the source. Summaries can also offer a critique of the source (the *evaluative* function) (Lancaster 1991), (Sparck-Jones 1999). Summaries can be tailored to a reader's interests and expertise, yielding *topic-related* summaries, or else they can be aimed at a broad readership, as in the case of so-called *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. Summaries can span one or more documents (Radev and McKeown 1998), (Mani and Bloedorn 1999); here we are concerned primarily with single-document summaries. Finally, it is possible to compose together translation and summarization in some order; here we deal simply with English summaries of English texts.

1.2 Summarization evaluation methods

Methods for evaluating text summarization can be broadly classified into two categories, based on the intrinsic/extrinsic distinction of Sparck-Jones and Galliers (1996).

Intrinsic evaluation tests the summarization system in itself. This can involve assessing *coherence* of the summary in terms of subjective grading of readability, lapses in grammaticality, presence of dangling anaphors (a common problem when extracting sentences out of context), or ravaging of structured environments like lists or tables (Brandow, Mitze and Rau 1995; Minel, Nugier and Piat 1997; Saggion and Lapalme 2000). Coherence in itself is not a sufficient test of summarization capability; it is certainly possible to have a well-written but atrocious summary. Another approach involves assessing the *informativeness* of the summary. This can be based on to what extent key information from the source is preserved in the system summary at different levels of compression (Paice 1990; Brandow, Mitze and Rau 1995). Informativeness can also be assessed in terms of how much information

https://doi.org/10.1017/S1351324901002741 Published online by Cambridge University Press

in an *ideal* (or 'reference') summary is preserved in the system summary, where the summaries being compared are at similar levels of compression (Edmundson 1969).

The problem with matching a system summary against an ideal summary is that the set of reference summaries is necessarily incomplete: there is always the possibility of a system generating a summary that is quite different from any reference summary, but is still a good summary. This complication, which is also a problem for other NLP technologies such as machine translation, text generation, and speech synthesis where there may be many potentially acceptable outputs, is most extreme in the case of generated abstracts. However, it can occur even in the simple case of extracts where referential expressions are substituted (e.g. a proper name for a pronoun). The construction of reference summaries can be quite problematic in practice; there have been several reports of low agreement among judges when they are asked to construct reference summaries by extracting sentences, e.g. (Rath, Resnick and Savage 1961; Salton, Singhal, Mitra and Buckley 1997), although judges may agree more on the most important sentences to include (Jing, Barzilay, McKeown and Elhadad 1998; Marcu 1999). The comparison between system and reference summaries is usually measured in the case of extracts in terms of 'sentence recall' (how many of the reference summary sentences the machine summary contains), 'sentence rank correlation' (comparing the summarizer's ranking of sentences for extract-worthiness with a corresponding ranking by a human), or a 'content-based measure' (which compares the semantic content of system and reference summaries) (Donaway, Drummey and Mather 2000).

Content-based measures, which offer some advantages over the others, including the ability to address both extracts and abstracts (Donaway, Drummey and Mather 2000), are usually approximated by an automatically scored 'vocabulary overlap measure', generated by first filtering out function words in a stop-list, and then computing overlap using a measure such as cosine similarity (Salton and McGill 1983). Vocabulary overlap measures can also be enhanced to take word order into account. While they are relatively insensitive to differences in meaning, e.g., between "The experiments provide evidence in favor of the hypothesis" and "The experiments *don't* provide evidence in favor of the hypothesis", since they have been used mostly for comparing extracts, or comparing abstracts that have a high degree of cut-and-paste relationship with the source, this problem does not manifest itself as much.

The second category of evaluation, an *extrinsic* evaluation, tests the summarization based on how it affects the completion of some other task. There have been a number of extrinsic evaluations involving question-answering and comprehension tasks (Morris, Kasper and Adams 1992), as well as tasks which measure the impact of summarization on determining the relevance of a document to a topic (Brandow, Mitze and Rau 1995; Mani and Bloedorn 1997; Jing, Barzilay, McKeown and Elhadad 1998; Tombros and Sanderson 1998). In applications involving information dissemination (e.g., distributing abstracts of scientific literature), one can also measure how much effort is required ('post-edit' time) to get the summary into some task-dependent acceptable state.

Extrinsic evaluations are especially attractive to a program like TIPSTER, because

https://doi.org/10.1017/S1351324901002741 Published online by Cambridge University Press

they can model tasks of interest to the TIPSTER funding agencies. Accordingly, the SUMMAC evaluation is mainly an extrinsic evaluation, focused on relevance assessment tasks. However, in order to explore relationships between the two kinds of evaluation, we also conducted an intrinsic evaluation Q&A task described below, which measures summary informativeness in terms of the extent to which key information from the source is preserved in the system summary at different levels of compression.

1.3 Related evaluations

There have been several other extrinsic summarization evaluations related to the task of relevance assessment. In the evaluation by Tombros *et al.* (1998), the subjects were asked to find as many relevant documents as possible for a query within five minutes, with a comparison made between topic-related and generic summaries. Subjects were also allowed to reference the full-text. While this study is very interesting (with topic-related summaries resulting in higher accuracy and speed of relevance assessment compared to generic summaries), 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. Jing *et al.* (1998) found that deciding on the relevance of a summary at 20% of the source length took a little over half the time in comparison with using the full source, with only a 7% loss of accuracy. However, this was a pilot study which used a small number of topics (4) and documents (10 per topic). Mani and Bloedorn (1997) also found that summaries led to faster relevance assessment without significant loss of accuracy, but they too used only four topics (but with 75 documents per topic); their study in addition lacked a baseline summary for comparison. The extrinsic evaluation of Brandow *et al.* (1995) differed from the others in the following way. Instead of selecting the set of documents for a query based on an information retrieval (IR) system's search against a full-text index, and then varying which instantiation (e.g. full-text or summary) each subject saw, Brandow *et al.* (1995) selected multiple sets of documents for each query, based on varying between a full-text and a summary index. Their results showed that summaries led to a dramatic loss in recall, with a significant gain in precision. However, their summarizer fared no better than a baseline method where just the initial ('leading') text of the document was used.

2 Summarization technologies

2.1 Overview

We now briefly describe the technologies used in text summarization. In general, summarization systems go through three stages: analysis of the source text to build an internal representation, transformation of the internal representation into a summary representation, and synthesis of output text (absent from systems which produce only extracts). Key operations carried out involve selection (filtering of elements), aggregation (merging of elements), and generalization (substituting more general elements for more specific ones).

https://doi.org/10.1017/S1351324901002741 Published online by Cambridge University Press

There are two broad classes of approaches to text summarization (Mani and Maybury 1999). The *knowledge-rich* approaches build a formal representation of the meaning of the source text, either as a set of logical forms (Reimer and Hahn 1988), or as a template describing some key concept (e.g. a main event in a document about that event) in the text (Paice 1990; Radev and McKeown 1998). In some cases, the system starts from a source of structured data, e.g. (Robin 1994; Maybury 1995; Radev and McKeown 1998), where there is no associated full-text source (though in some of these instances, the structured data could have been constructed from a full-text source or sources as a result of an information extraction process).

The *surface-oriented* approaches select material from the source based on weights for features such as (1) position in some structural representation of the document (Edmundson 1969; Kupiec, Pedersen and Chen 1995; Hovy and Lin 1999), (2) presence of cue phrases (Edmundson 1969; Pollock and Zamora 1975) like "in summary" or "in conclusion", as well as terms like "excellent" (higher weight), and "unimportant" (lower weight), (3) presence of background terms from the title, headings in the text, the initial part of the text, or a user's query, and (4) presence of statistically salient terms, also called 'keywords' (Luhn 1958; Edmundson 1969). These surface-oriented approaches have been limited to selecting extracts of source material (sometimes accompanied by topic or named entity lists), which are recombined by concatenation or syntactic rearrangement (Mani, Gates and Bloedorn 1999) to meet the compression rate requirement.

These are pure examples of the approaches, and many systems, including some of those evaluated here, adopt a hybrid approach. In recent research, corpus-based approaches have become quite dominant, including the machine learning of feature combinations (Kupiec, Pedersen and Chen 1995; Mani and Bloedorn 1998; Aone, Okurowski, Gorlinsky and Larsen 1999; Teufel and Moens 1999), as well as the learning of sentence-shortening rules (Knight and Marcu 2000). There has also been an interesting body of work exploiting a representation of the discourse structure of the text, based on Halliday's notions (Halliday and Hasan 1996) of text cohesion (Morris and Hirst 1991; Hearst 1997; Mani and Bloedorn 1999; Barzilay and Elhadad 1999; Boguraev and Kennedy 1999; Aone, Okurowski, Gorlinsky and Larsen 1999) and text coherence (Miike, Itoh, Ono and Sumita 1994; Marcu 1999). For a more detailed introduction to the field, see Hovy and Marcu (1998), Mani and Maybury (1999) and Mani (2001).

2.2 Participant systems

To simplify the evaluation, it was designed as a black-box evaluation, which looked at summarization systems as a whole rather than their internal components. To keep things simple, the evaluation did not consider the type of technology used in the system as an independent variable; no attempt was made to systematically classify the different technologies to study their effect on performance. Nevertheless, in keeping with many successful black-box evaluations such as the Message Understanding Conference (MUC) (Grishman and Sundheim 1996) and the Text Retrieval

https://doi.org/10.1017/S1351324901002741 Published online by Cambridge University Press

Table 1. *Participant summarization method features. tf: term frequency; loc: location; disc:discourse; coref: coreference; co-occ: co-occurrence; syn: synonyms.*

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	+	-	-	-	+	-

Conference (TREC) (Harman and Voorhees 1996), we wanted to offer feedback to developers by comparing different participants' performances on these tasks.

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 Massachusetts (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 participant systems[2]. Most participants confined their summaries to extracts of passages from the source text; TextWise, however, extracted combinations of passages, phrases,

[2] The discourse feature in Table 1 covers a variety of different uses of discourse models, ranging from document structure parsing to analysis of cohesion patterns in the text.

https://doi.org/10.1017/S1351324901002741 Published online by Cambridge University Press

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. To offer a little more detail in terms of the type of technique used, a breakdown per participant system is as follows:

- 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 were 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.
- 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 used proprietary algorithms to analyze a document and produce a summary intended 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 term-frequency derived 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, 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, named

https://doi.org/10.1017/S1351324901002741 Published online by Cambridge University Press

entity 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 assigned subject field codes to documents (using a thesaurus) 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 information retrieval 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.

3 SUMMAC summarization tasks

3.1 Overview

To address the goals of the evaluation, two extrinsic evaluation tasks and one intrinsic task were defined, based on activities typically carried out by information analysts in the US Government. In the extrinsic *ad hoc 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 judgments of the full-text source relevance, which were separately obtained from the TREC conferences. Thus, an indicative summary would be 'accurate' if it accurately reflected the relevance or irrelevance of the corresponding source.

In the extrinsic *categorization task*, the evaluation sought to find out whether a *generic* indicative 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, an intrinsic *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',

https://doi.org/10.1017/S1351324901002741 Published online by Cambridge University Press

namely, the degree to which it contained answers found in the source document to a set of topic-related questions.

3.2 Data selection

In the ad hoc task, 20 TREC topics were selected (see Table 9 for an example of a topic). 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 TREC topics were selected, with 100 documents used per topic. The categorization topics were selected such that they could be grouped into two mutually exclusive classes: *environment* and *global economy*. 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).

In both tasks, the subsets were constructed for each topic 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 subsets had no documents in common. Given the top 200 ranked documents, documents which met the length and relevance criteria were selected; from these, the top 50 (*ad hoc* task) or top 100 (categorization task) were chosen based on the search engine's ranking.

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.

In each task, participants submitted two summaries: a fixed-length ($S_{10\%}$) summary limited to 10% of the length of the source, and a summary which was not limited in length (S_{var}).

4 Experimental hypotheses and method

4.1 Tests and experimental design

In meeting the 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*: to determine whether subjects' relevance assessment performance in terms of time and accuracy was affected by different conditions: full-text (F), fixed-length summaries ($S_{10\%}$), variable-length summaries (S_{var}), and baseline summaries (B). The baselines were comprised of the first 10% of the body of the source text; this baseline was chosen based on the finding by Brandow *et al.* (1995) of the importance of leading text summaries. The second test was a *participant system 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 ad hoc and categorization tasks.

https://doi.org/10.1017/S1351324901002741 Published online by Cambridge University Press

Table 2. *Ad hoc task contingency table. TP = true positive, FP = false positive, TN = true negative, FN = false negative*

Ground truth	**Subject's judgment**	
	Relevant	**Irrelevant**
Relevant is True	TP	FN
Irrelevant is True	FP	TN

Table 3. *Categorization task contingency table. X and Y are distinct categories other than "none of the above", represented as None*

Ground truth	**Subject's judgment**		
	X	**Y**	**None**
X is True	TP	FN	FN
None is True	FP	FP	TN

To arrive at definitive conclusions regarding these comparisons, we assumed a statistical methodology based on statistical significance testing. This methodology is as follows: in each of the tests above, we assume a null hypothesis, e.g. in the first two tests, the null hypothesis is that there is no difference in performance of subjects or systems between different conditions. We apply a statistical test to assess the difference in performance between groups (e.g. summarization conditions or systems), 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 in performance between summarization conditions or systems). Conventions for 'small', 'medium', and 'large' are arrived at from the experimental literature; see Cohen (1969) for details.

In designing our experiment, we conducted a 'dry-run' evaluation on the ad hoc and categorization tasks, which used fewer subjects and systems than the final evaluation. The dry-run indicated that we could expect very large effect sizes for the summarization condition and participant system 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

https://doi.org/10.1017/S1351324901002741 Published online by Cambridge University Press

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 $S_{10\%}$ and 480 to S_{var}. These 960 $S_{10\%}$ and S_{var} documents not only provided for acceptably small α and β for the participant system test, but also allowed us to sample as many of the systems' summaries as possible, which enabled us to offer detailed feedback to each participant.

Based on this, in the formal evaluation, for the ad hoc and categorization tasks, the 1000 documents assigned to a subject for each task were allocated among F, B, $S_{10\%}$, and S_{var} conditions through random selection without replacement (20 F, 20 B, 480 $S_{10\%}$, and 480 S_{var}). The 480 $S_{10\%}$ and S_{var} summaries were each divided uniformly across the 16 participants (with random selection of approximately 30 summaries per participant for each of $S_{10\%}$ and S_{var}). 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 ad hoc task used 21 subjects, the main categorization 24 subjects; the consistency ad hoc task had 14 subjects, the consistency categorization 7 subjects (some subjects from the main task also did a consistency task). The subjects were told they were working with documents that included summaries. In the *ad hoc* task they were told 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; in the categorization task, they were asked, given a document and five topics, to select a topic to which the document was relevant, or else to choose "none of the above"[3].

4.2 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.

4.2.1 Basic measures of accuracy

The contingency table for the ad hoc task is shown in Table 2. This contingency table analysis is ideally suited to problems of binary classification as in the *ad hoc* task. For the categorization task, where one is assigning a document to exactly one of six categories, we chose to collapse all the outcomes to two possibilities, correct or incorrect, as shown in Table 3.

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

[3] The instructions were only a few pages long.

https://doi.org/10.1017/S1351324901002741 Published online by Cambridge University Press

4.2.2 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[4], 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.

$$Precision = TP/(TP + FP) \tag{1}$$

$$Recall = TP/(TP + FN) \tag{2}$$

$$Fscore = 2 * Precision * Recall/(Precision + Recall) \tag{3}$$

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-score) 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) 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 truly isn't) has a probability of only α.

5 Results: *ad hoc* and categorization tasks

5.1 Performance by condition

In the *ad hoc* task, S_{var} summaries (at compressions as low as 17% of full text length) sped up decision-making by almost a factor of 2 (33.12 seconds per decision average time for S_{var} compared to 58.89 for F in Table 4), and were only 4% less accurate than full text (not a statistically significant difference – see Table 4)[5].

In the categorization task, the F-score on full-text was only 0.5, suggesting the task was hard in comparison with the ad-hoc task. Here a time speedup to 25.48 seconds for $S_{10\%}$ compared to 43.11 seconds for F was the only significant difference. None of the differences in accuracy were significant; since news articles are typically written to put the most important information first, in the case of generic summaries, it appears that here the baseline condition is hard to beat.

[4] Another aggregate measure is Predictive Accuracy $(TP + TN)/(TP + FP + TN + FN)$. Precision and Recall values tend to agree with Predictive Accuracy when the number of TP, TN, FP and FN are about equal; however, since Precision and Recall allow us to distinguish between the effects of FPs and FNs, we will use those instead.

[5] The significance level $\alpha < 0.05$ is used throughout this paper, unless noted otherwise.

https://doi.org/10.1017/S1351324901002741 Published online by Cambridge University Press

Table 4. *Ad hoc 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 $S_{10\%}$ (HSD=9.8). All F-score differences are significant, except between F (Full-Text) and S_{var} (HSD=.10). Precision (P) differences are not significant. All Recall (R) differences between conditions are significant, except between F and S_{var} (HSD=0.12)*

Condition	Time	F-score	TP	FP	FN	TN	P	R
F	58.89	0.67	0.38	0.08	0.26	0.28	0.83	0.22
S_{var}	33.12	0.64	0.35	0.08	0.28	0.28	0.80	0.23
$S_{10\%}$	19.75	0.53	0.27	0.07	0.35	0.31	0.79	0.19
B	23.15	0.42	0.18	0.05	0.41	0.35	0.81	0.12

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[6]. This indicates that improved tailoring of the summary to the topic is required. In both, there were fewer gains in F-score above a 20% compression rate[7]. Finally, both tasks reveal a high variability in the time per decision taken by subjects in all conditions (a standard deviation of 35.45 for *ad hoc* and 38.79 for categorization).

5.2 Performance by participant

In the *ad hoc* task, the systems were all very close in accuracy for both summary types (Table 5). Three groups of systems were evident in the *ad hoc* S_{var} F-score accuracy data: Group I (CGI/CMU and Cornell/SabIR) is significantly more accurate than Group III (ISI), with no significant differences in accuracy between either of these groups and Group II (the rest); within groups, no significant differences in accuracy were found[8]. Interestingly, the Group I systems both relied on statistics based on term frequency and co-occurrence (Table 1) and vocabulary overlap comparisons between text passages. For the S_{var} 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; considering both accuracy and time, TextWise, GE and Penn (equivalent in accuracy) were the closest in terms of Cartesian distance from the ideal performance. For $S_{10\%}$ 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*

[6] By using Precision and Recall rather than Predictive Accuracy, the fact that none of the summaries (including baselines) differ from the source in Precision, although some differ in Recall, is brought to light. This means that the summaries are not erroneously including information that would lead to a false judgment of relevance.

[7] The S_{var} summaries in both tasks had similar average compression rates, 22.24% *ad hoc* and 23.58% categorization.

[8] To show the differences, the rows are sorted by decreasing S_{var} F-scores.

https://doi.org/10.1017/S1351324901002741 Published online by Cambridge University Press

Table 5. *Ad hoc accuracy by participant. For variable-length (S_{var}): 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 ($S_{10\%}$), there are no significant differences on any of the measures*

System	S_{var}			$S_{10\%}$		
	P	**R**	**F-score**	**P**	**R**	**F-score**
CGI/CMU	0.82	0.66	0.72	0.76	0.52	0.60
Cornell/SabIR	0.78	0.67	0.70	0.79	0.47	0.56
GE	0.78	0.60	0.67	0.77	0.45	0.55
LN	0.78	0.58	0.65	0.81	0.45	0.55
Penn	0.81	0.57	0.65	0.76	0.45	0.53
UMass	0.80	0.54	0.63	0.81	0.47	0.56
NMSU	0.80	0.54	0.63	0.80	0.40	0.52
TextWise	0.81	0.51	0.61	0.79	0.41	0.52
SRA	0.82	0.49	0.60	0.79	0.37	0.48
NTU	0.80	0.49	0.59	0.82	0.34	0.46
ISI	0.80	0.46	0.56	0.82	0.36	0.47

TP resulted in CGI/CMU, GE, LN, UMass and Cornell/SabIR clustering together on both $S_{10\%}$ and S_{var} 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[9] had no significant differences in F-score accuracy whatsoever (Table 6, Figures 3 and 4). In this task, in the absence of a topic, the statistical salience systems which performed relatively more accurately in the *ad hoc* 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, Mitze and Rau 1995): the generic (categorization) summaries had a higher percentage of selections of first sentences from the source than the ad hoc summaries (35% of $S_{10\%}$ and 41% of S_{var} for categorization, compared to 21% $S_{10\%}$ and 32% S_{var} for *ad hoc*). 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.

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

https://doi.org/10.1017/S1351324901002741 Published online by Cambridge University Press

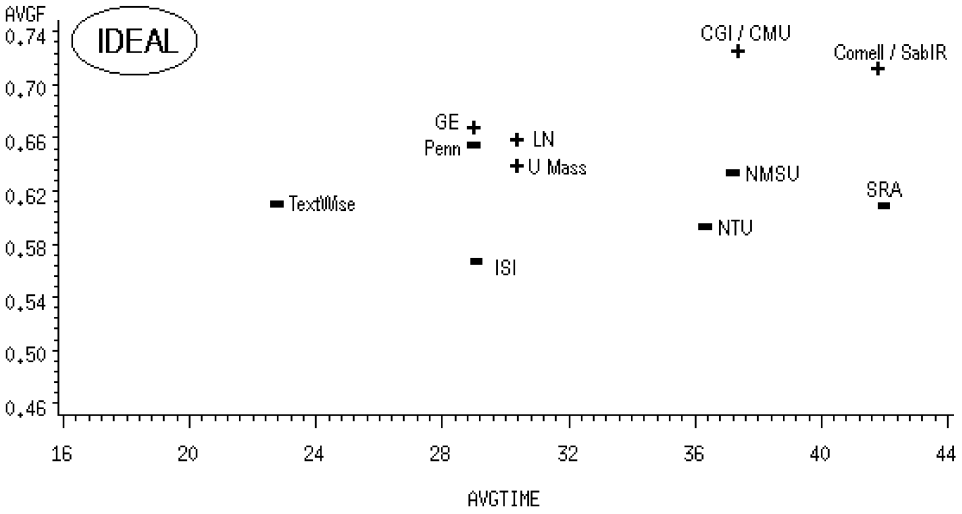


Fig. 1. *Ad hoc* F-score versus time by participant (variable-length summaries). HSD(F-score) is 0.13. HSD(Time) = 12.88. Decisions based on summaries from GE, Penn and TextWise are significantly faster than based on SRA and Cornell/SabIR.

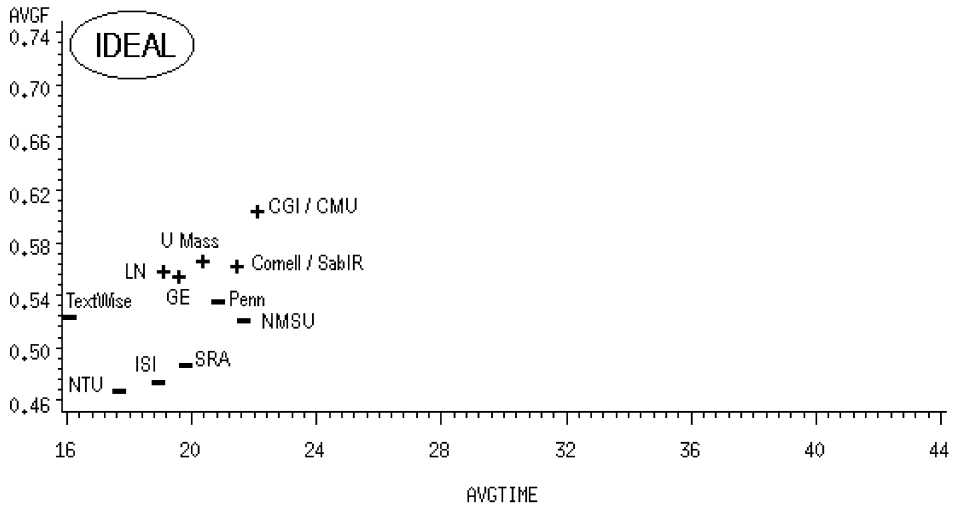


Fig. 2. *Ad hoc* F-score versus time by participant (fixed-length summaries). No significant differences in F-score, or in time.

5.3 Agreement between subjects

The percentage agreement between a pair of subjects is computed by examining how often a relevance assessment decision made by one subject in the pair is identical with the decision made by the other subject. We get the 'pairwise agreement' by averaging the percentage agreement across all pairs of subjects. The percentage of

https://doi.org/10.1017/S1351324901002741 Published online by Cambridge University Press

Table 6. *Categorization accuracy by participant. No significant differences on any of the measures*

System	S_{var}			$S_{10\%}$		
	P	**R**	**F-score**	**P**	**R**	**F-score**
BT	0.63	0.43	0.48	0.70	0.33	0.41
CGI/CMU	0.74	0.39	0.47	0.69	0.33	0.42
CIR	0.71	0.47	0.54	0.68	0.35	0.43
Cornell/SabIR	0.66	0.40	0.47	0.62	0.36	0.42
GE	0.69	0.40	0.47	0.69	0.33	0.42
IA	0.69	0.42	0.49	0.67	0.33	0.41
IBM	0.68	0.47	0.51	0.63	0.37	0.44
ISI	0.71	0.42	0.49	0.71	0.35	0.44
LN	0.68	0.41	0.47	0.68	0.37	0.45
NMSU	0.69	0.46	0.51	0.69	0.34	0.43
NTU	0.66	0.41	0.48	0.68	0.33	0.43
Penn	0.70	0.42	0.50	0.66	0.29	0.38
SRA	0.65	0.42	0.48	0.73	0.37	0.45
Surrey	0.69	0.43	0.51	0.69	0.31	0.39

Table 7. *Percentage of decisions subjects agreed on when viewing full-text (consistency tasks)*

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

times *all* subjects make identical decisions is called the 'unanimous agreement'. As indicated in Table 7, the unanimous agreement of just 16.6% and 19.5% in the *ad hoc* and categorization tasks respectively is low: the agreement data has Kappa (Carletta, Isard and others 1997) of 0.38 for *ad hoc* and 0.29 for categorization[10].

[10] 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 *ad hoc* task.

https://doi.org/10.1017/S1351324901002741 Published online by Cambridge University Press

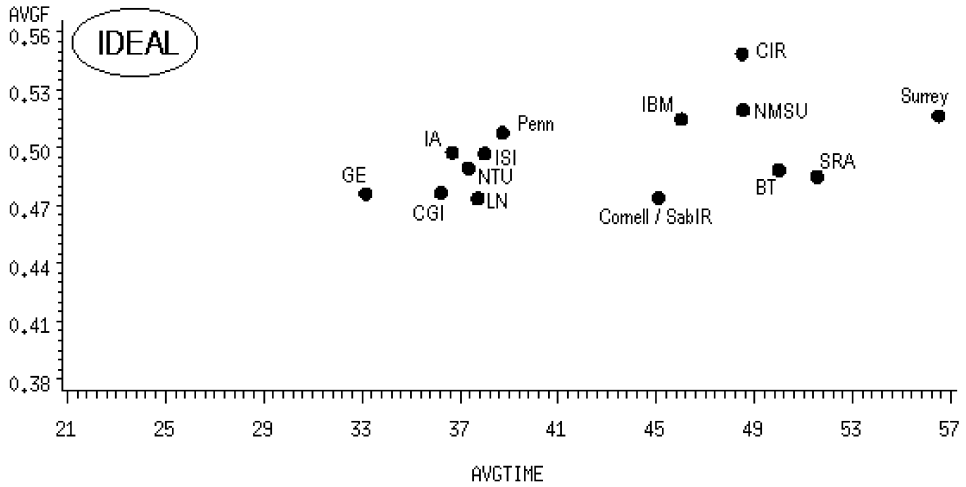


Fig. 3. Categorization F-score versus time by participant (variable-length summaries). F-scores are not significantly different. HSD(Time) = 17.23. GE is significantly faster than SRA and Surrey. The latter two are also significantly slower than Penn, ISI, LN, NTU, IA and CGI/CMU.

Table 8. *Agreement on relevant documents alone (consistency tasks)*

Task	**Pairwise**	**3-way**
Ad hoc	52.9	36.9
Categorization	45.9	29.7
TREC	44.7	30.1

The *ad hoc* pairwise and 3-way agreement (i.e. unanimous agreement among groups of three subjects) is consistent with a 3-subject dry-run *ad hoc* consistency task carried out earlier. However, the three-way agreement is much lower than the 3-way agreement of 71.7% reported in 3-subject studies of agreement in TREC relevance assessment (Harman and Voorhees 1996).

One possible explanation (evidenced also by the high variance in the time for each relevance assessment decisions) 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 the (well recognized problem of) lack of agreement in relevance judgments. Further, she found when only relevant documents were considered, 44.7% pairwise agreement and 30.1% 3-way agreement with 3 subjects, which is comparable to our scores, as shown in Table 8.

https://doi.org/10.1017/S1351324901002741 Published online by Cambridge University Press

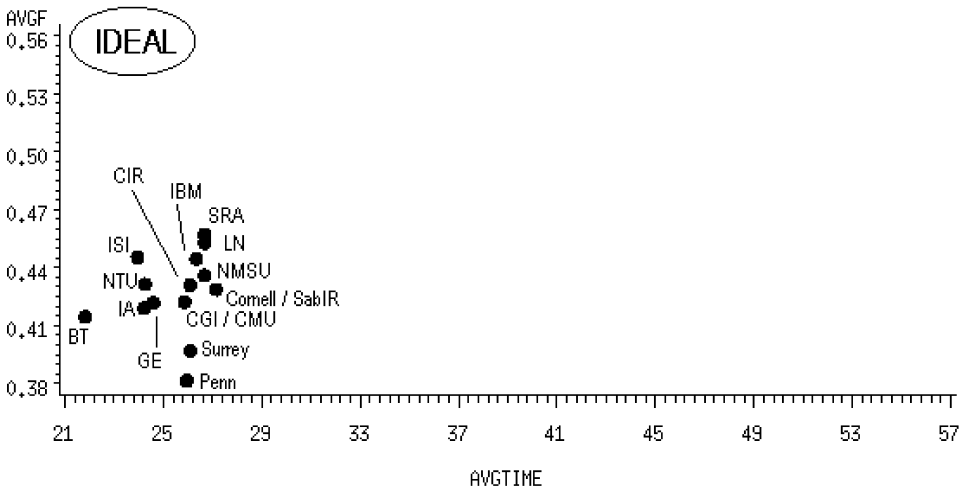


Fig. 4. Categorization F-score versus time by participant (fixed-length summaries). F-scores are not significantly different, and neither are time differences.

6 Question-answering (Q&A) task

In this task, the summarization system, given a topic and a document relevant to the topic, needed to produce an informative, topic-related summary. Here an accurate summary would contain the answers found in that document to a set of topic-related questions. These questions covered 'obligatory' information that would 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?". Assuming that a human annotator has devised a set of such questions, the answers to the questions, when found in a relevant document, are marked up by that annotator in that document. A document marked up with answers is called an 'answer key'. An accurate summary should contain the same information as found in the answers in the answer key.

6.1 Experimental design

Three topics were chosen from the 20 ad hoc TREC topics. For each topic, 30 relevant documents from the ad hoc task corpus were chosen as the source texts for topic-related summarization. The principal tasks of each evaluator (one evaluator per topic, three evaluators in all) were to prepare the questions and answer keys and to score the 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 9).

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 keys. The evaluators used a common set of guidelines for writing questions, creating

https://doi.org/10.1017/S1351324901002741 Published online by Cambridge University Press

Table 9. *Q&A Topic 258, topic-related questions, and part of a relevant source document showing answer key annotations*

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 to 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?

Annotated Source Fragment

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

answer keys, and scoring summaries that were intended to minimize variability across evaluators in the methods used[11].

Eight of the *ad hoc* 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 general, if the response summary included the one sentence that was identified in the key as the answer

[11] We also had each of the evaluators score a portion of each others' test data. On two of the three topics, the subjects had similar *ARAs* (mean *ARA* of 0.51 with a standard deviation of 0.034).

https://doi.org/10.1017/S1351324901002741 Published online by Cambridge University Press

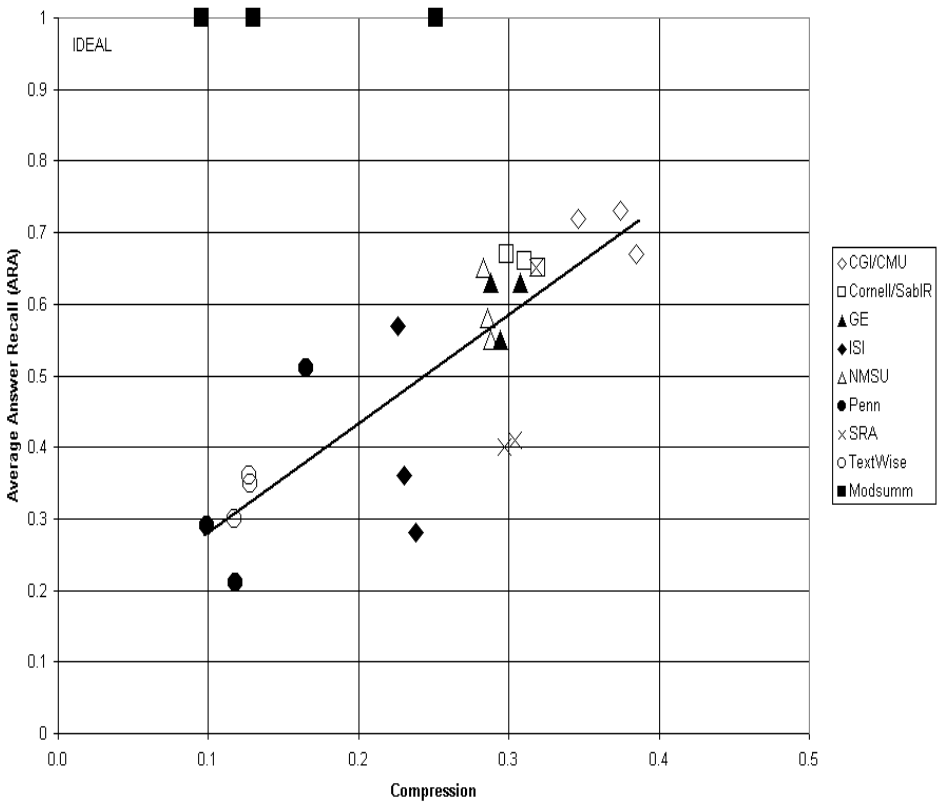


Fig. 5. ARA versus compression by participant. 'Modsumms' are model summaries.

to a question, the assessors would have assigned a score of Correct, without much consideration as to the coherence of that response sentence in the overall summary context. (The fact that the response was a complete sentence was generally regarded as sufficient context for the answer that was contained in that sentence.) However, it occasionally happened that the answer in the key spanned more than one sentence. In such a case, if the response included only one of those sentences, the assessors would usually have assigned a score of Partially Correct, or even Missing. 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: *ARS* (Answer Recall Strict) and *ARL* (Answer Recall Lenient):

$$ARS = n1/n3 \tag{4}$$

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

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. *ARL* is a more lenient measure than *ARS* because it allows

https://doi.org/10.1017/S1351324901002741 Published online by Cambridge University Press

partial credit for Partially Correct answers. 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, where compression is summary length/source length). A line fitted through the data (for just the system summaries) has a slope of 1.52, meaning that the current state of the art performance, expressed as an informativeness ratio of accuracy to compression, is about 1.5. However, because of the very small-scale nature of the Q&A task, no definitive conclusions can be drawn about the differences between individual systems.

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 $r = 0.97$, $\alpha < 0.0001$). The Q&A evaluation is therefore promising as a new methodology for automated evaluation of informative summaries.

7 Conclusions

The main conclusions from SUMMAC are as follows:

- SUMMAC has established definitively in a large-scale evaluation that automatic text summarization is very effective in relevance assessment tasks on newpaper articles. Summaries at relatively low compression rates (generic summaries at 10% of source length, and topic-related summaries as short as 17% of source length) reduced relevance assessment time by 40% (for generic summaries) to 43% (for topic-related summaries), with no statistically significant degradation in accuracy.
- Systems that performed most accurately in the production of indicative and informative topic-related summaries used statistical methods involving term frequency and co-occurrence statistics, and vocabulary overlap comparisons between text passages. Overall, the five highest scoring systems in the *ad hoc* task represented similar features of the source text and extracted similar sets of sentences from it. However, in the absence of a topic, these statistical methods did not provide any additional leverage: in the case of generic summaries, the systems (which relied in this case on inclusion of the first sentence of the source) were indistinguishable in accuracy.
- Analysis of feedback forms filled in after each decision indicated that the coherence of present-day machine-generated summaries is high, due to use of sentence extraction and coherence 'smoothing' based on anaphora resolution and presentation of additional context in the summary. However, the intelli-

https://doi.org/10.1017/S1351324901002741 Published online by Cambridge University Press

bility of these summaries[12] is somewhat less than the full-text and the baseline (the latter in conformance with the result of Brandow *et al.* (1995), who found that that leading-text summaries were more coherent than summaries based on a statistical method).

- A promising new method has been developed for automatically scored intrinsic evaluation of the informativeness of topic-related summaries, as demonstrated in the Q&A task.

8 Lessons learned and future directions

Evaluations involving live subjects can be costly to set up, and in our case called for an experimental design where no group of subjects had to read very long documents. Further, extrinsic evaluations involving relevance assessment require relevance judgments to be available; these are hard to come by, and when made available, potential problems of lack of agreement in relevance judgments, as we discovered, may need to be addressed. Our 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 other kinds of texts might challenge the summarizers to a greater or lesser extent.

SUMMAC was an extremely labor-intensive evaluation. While the results are definitive, developers need repeatable, automatically-scorable evaluations (Hirschman and Mani 2001). This can be achieved by developing annotated corpora which can provide reference data, whether in the form of a representation of the input source document, as in the Q&A annotated mini-corpus, or in the form of reference summaries linked to their sources (Jing and McKeown 1999; CMP-LG 1999; Marcu 1999; Goldstein, Kantrowitz, Mittal and Carbonell 1999), etc. In addition to general issues of corpus creation such as size, heterogeneity, annotation standards, etc., adequate document lengths and compression rates of summaries are critical.

Another challenge is designing evaluations which exploit features unique to summarization. The SUMMAC tasks did not cover the full spectrum of single-document summarization: they required only extracts, rather than abstracts (although participants could have submitted abstracts, none did.) In addition, sophisticated presentation and interaction strategies may make a substantial difference in the effectiveness of summarization, and could substantially challenge the synthesis component of summarizers. We chose to ignore these presentation issues to simplify the evaluation.

Summarization evaluation is a very active field, with many ongoing efforts in the US, Europe, and Japan. Japan's National Institute of Informatics is in the midst of conducting an evaluation of text summarization systems, called the Text Summarization Challenge (NII 2001). The evaluation, which was due to be completed in March 2001, has an intrinsic evaluation component as well as an extrinsic one, with the latter being based on the SUMMAC *ad hoc* task.

[12] On the *ad hoc* task, 99% of F were judged 'intelligible', as were 93% S_{var}, 96% B, 83% $S_{10\%}$; similar data for categorization.

https://doi.org/10.1017/S1351324901002741 Published online by Cambridge University Press

The Q&A evaluation data has been used in small-scale evaluations to evaluate other summarization systems, including Mani *et al.* (1999) and Lin (1999). The Q&A evaluation's emphases on approximate matching and indexing into the right place in the source text has also influenced the design of the 'event99' information extraction metrics (Hirschman *et al.* 1999). The evaluation also demonstrates that metrics based on vocabulary overlap can be quite effective in determining the informativeness of summaries, and we expect that such metrics will enjoy wide use in related evaluations, e.g. of question-answering capabilities of systems.

In the future, new areas such as multi-document summarization and multi-lingual summarization will assume increasing importance, posing new challenges for evaluations (see Baldwin, Donaway and others (2000) for a roadmap for future summarization evaluations associated with the Document Understanding Conference (DUC 2001)). The 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 and the 16 participant system providers. 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 relevance assessment tool. We would also like to thank several anonymous referees for their comments.

References

Aone, C., Okurowski, M. E., Gorlinsky, J. and Larsen, B. (1997) A trainable summarizer with knowledge acquired from robust NLP techniques. In: Mani, I. and Maybury, M., editors, *Advances in Automatic Text Summarization*, pp. 71–80. MIT Press.

Baldwin, N., Donaway, R., Hovy, E., Liddy, E., Mani, I., Marcu, D., McKeown, K., Mittal, V., Moens, M., Radev, D., Sparck Jones, K., Sundheim, B., Teufel, S., Weischedel, R. and White, M. (2000) An Evaluation Roadmap for Summarization Research. http://www-nlpir.nist.gov/projects/duc/roadmapping.html

Barzilay, R. and Elhadad, M. (1999) Using lexical chains for text summarization. In: Mani, I. and Maybury, M., editors, *Advances in Automatic Text Summarization*, pp. 111–121. MIT Press.

Borko, H. and Bernier, C. (1975) *Abstracting Concepts and Methods*. Academic Press.

Brandow, R., Mitze, K. and Rau, L. (1995) Automatic condensation of electronic publications by sentence selection. *Infor. Process. Manage.* **31**(5): 675–685. (Reprinted in Mani, I. and Maybury, M., editors, *Advances in Automatic Text Summarization*, pp. 293–303. MIT Press.)

Boguraev, B. and Kennedy, C. (1999) Salience-based content characterization of text documents. In: Mani, I. and Maybury, M., editors, *Advances in Automatic Text Summarization*, pp. 99–110. MIT Press.

CMP-LG Annotated Corpus (1999) http://www.itl.nist.gov/div894/894.02/related_projects/tipster_summac/index.html

https://doi.org/10.1017/S1351324901002741 Published online by Cambridge University Press

Carletta, J., Isard, A., Isard, S., Jowtko, J. C., Doherty-Sneddon, G. and Anderson, A. H. (1997) The reliability of a dialogue structure coding scheme. *Computational Linguistics*, **23**(1): 13–32.

Cohen, J. (1969) *Statistical Power Analysis for the Behavioral Sciences*. Academic Press.

Donaway, R. L., Drummey, K. W. and Mather, L. A. (2000) A comparison of rankings produced by summarization evaluation measures. *Proceedings ANLP-NAACL'2000 Workshop on Automatic Summarization*, pp. 69–78.

Document Understanding Conference (2001) http://www.nlp-ir.nist.gov/projects/duc/2001.html

Edmundson, H. P. (1969) New methods in automatic abstracting. *J. ACM*, **16**(2): 264–285. (Reprinted in Mani, I. and Maybury, M., editors, *Advances in Automatic Text Summarization*, pp. 21–42. MIT Press.)

Goldstein, J., Kantrowitz, M., Mittal, V. and Carbonell, J. (1999) Summarizing text documents: sentence selection and evaluation metrics. *Proceedings 22nd International Conference on Research and Development in Information Retrieval (SIGIR'99)*, pp. 121–128.

Hahn, U. and Mani, I. (2000) The challenges of automatic summarization. *IEEE Computer*, **33**(11): 29–36.

Halliday, M. and Hasan, R. (1996) *Cohesion in Text*. Longman.

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

Hearst, M. (1997) TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, **23**(1): 33–64.

Hirschman, L., Robinson, P., Ferro, L., Brown, E., Chinchor, N., Sundheim, B. and Grishman, R. (1999) Event99: Event evaluation for news on demand. Unpublished presentation for HUB4, prepared by MITRE, SAIC, NRaD and NYU.

Hirschman, L. and Mani, I. (2001) Evaluation. In: Mitkov, R., editor, *Handbook of Computational inguistics*. Oxford University Press.

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

Hovy, E. and Lin, C-Y. (1999) Automated text summarization in SUMMARIST. In: Mani, I. and Maybury, M., editors, *Advances in Automatic Text Summarization*, pp. 81–94. MIT Press.

Text Summarization Challenge (2001) *Proceedings Second NTCIR Workshop on Evaluation of Chinese and Japanese Text Retrieval and Text Summarization*. http://galaga.jaist.ac.jp:8000/tsc

Jing, H., Barzilay, R., McKeown, K. and Elhadad, M. (1998) Summarization evaluation methods: Experiments and analysis. *Working Notes of the AAAI Spring Symposium on Intelligent Text Summarization*, Technical Report, pp. 60–68.

Jing, H. and McKeown, K. (1999) The decomposition of human-written summary sentences. *Proceedings 22nd International Conference on Research and Development in Information Retrieval (SIGIR'99)*, pp. 129–136.

Kirk, R. E. (1968) *Experimental Design: Procedures for the Behavioral Sciences*. Wadsworth.

Knight, K. and Marcu, D. (2000) Statistics-based summarization – step one: Sentence compression. *Proceedings Seventeenth National Conference on Artificial Intelligence (AAAI'2000)*, pp. 703–710.

Kupiec, J., Pedersen, F. and Chen, F. (1995) A trainable document summarizer. *Proceedings 18th ACM SIGIR Conference (SIGIR'95)*, pp. 68–73. (Reprinted in Mani, I. and Maybury, M., editors, *Advances in Automatic Text Summarization*, pp. 55–60. MIT Press.)

Lancaster, F. W. (1991) Indexing and Abstracting in Theory and Practice. University of Illinois Graduate School of Library and Information Science.

Lin, C-Y. (1999) Training a selection function for extraction. *Proceedings 18th International Conference on Information and Knowledge Management (CIKM'99)*, pp. 1–8.

https://doi.org/10.1017/S1351324901002741 Published online by Cambridge University Press

Luhn, H. P. (1958) The automatic creation of literature abstracts. *IBM J. Research & Development*, **2**: 159–165. (Reprinted in Mani, I. and Maybury, M., editors, *Advances in Automatic Text Summarization*, pp. 15–21. MIT Press.)

Mani, I. and Bloedorn, E. (1997) Multi-document Summarization by graph search and merging. *Proceedings 14th National Conference on Artificial Intelligence (AAAI-97)*, Providence, RI, pp. 622–628.

Mani, I. and Bloedorn, E. (1998) Machine learning of generic and user-focused summarization. *Proceedings 15th National Conference on Artificial Intelligence (AAAI-98)*, Madison, WI, pp. 821–826.

Mani, I. and Bloedorn, E. (1999) Summarizing similarities and differences among related documents. *Infor. Retrieval*, **1**: 35–67.

Mani, I., Gates, B. and Bloedorn, E. (1999) Improving summaries by revising them. *Proceedings 37th Annual Meeting of the ACL*, pp. 558–565.

Mani, I. and Maybury, M. (editors) (1999) *Advances in Automatic Text Summarization*. MIT Press.

Mani, I. (2001) *Automatic Summarization*. John Benjamins.

Marcu, D. (1999) The automatic construction of large-scale corpora for summarization research. *Proceedings 22nd International Conference on Research and Development in Information Retrieval (SIGIR'99)*, pp. 137–144.

Maybury, M. (1995) Generating summaries from event data. *Infor. Process. Manage.* **31**(5): 735–751.

Miike, S., Itoh, E., Ono, K. and Sumita, K. (1994) A full-text retrieval system with a dynamic abstract generation function. *Proceedings 17th International Conference on Research and Development in Information Retrieval (SIGIR'94)*, pp. 152–161.

Minel, J.-L., Nugier, S. and Piat, G. (1997) How to appreciate the quality of automatic text summarization. In: Mani, I. and Maybury, M., editors, *Proceedings ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pp. 25–30.

Morris, A., Kasper, G. and Adams, D. (1992) The effects and limitations of automatic text condensing on reading comprehension performance. *Infor. Syst. Res.*, **3**(1): 17-35. (Reprinted in Mani, I. and Maybury, M., editors, *Advances in Automatic Text Summarization*, pp. 305–323. MIT Press.)

Morris, J. and Hirst, G. (1991) Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, **17**(1): 21–43.

Grishman, R. and Sundheim, B. (1996) Message Understanding Conference-6: A Brief History. *Proceedings COLING-96*, pp. 466–471.

Paice, C. (1990) Constructing literature abstracts by computer: Techniques and prospects. *Infor. Process. Manage.* **26**(1): 171–186.

Pollock, J. J. and Zamora, A. (1975) Automatic abstracting research at chemical abstracts service. *J. Chem. Infor. Comput. Sci.* **15**(4): 226–232. (Reprinted in Mani, I. and Maybury, M., editors, *Advances in Automatic Text Summarization*, pp. 43–49. MIT Press.)

Radev, D. and McKeown, K. (1998) Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, **24**(3): 469–500.

Rath, G. J., Resnick, A. and Savage, T. R. (1961) The formation of abstracts by the selection of sentences. *Am. Documentation*, **12**(2): 139–143. (Reprinted in Mani, I. and Maybury, M., editors, *Advances in Automatic Text Summarization*, pp. 287–292. MIT Press.)

Reimer, U. and Hahn, U. (1988) Text condensation as knowledge base abstraction. *Proceedings 4th IEEE/AAAI Conference on Artificial Intelligence Applications*, pp. 338–344.

Robin, J. (1994) Revision-based generation of natural language summaries providing historical background: corpus-based analysis, design and implementation. *PhD Thesis*, Columbia University.

Saggion, H. and Lapalme, G. (2000) Concept identification and presentation in the context of technical text summarization. *Proceedings ANLP-NAACL'2000 Workshop on Automatic Summarization*, pp. 1–10.

https://doi.org/10.1017/S1351324901002741 Published online by Cambridge University Press

Salton, G. and McGill, M. J. (1983) *Introduction to Modern Information Retrieval*. McGraw-Hill.

Salton, G., Singhal, A., Mitra, M. and Buckley, C. (1997) Automatic text structuring and summarization. *Infor. Process. Manage.* **33**(2): 193–207. (Reprinted in Mani, I. and Maybury, M., editors, *Advances in Automatic Text Summarization*, pp. 341–355. MIT Press.)

Sparck-Jones, K. (1999) Automatic summarizing: factors and directions. In: Mani, I. and Maybury, M., editors, *Advances in Automatic Text Summarization*, pp. 1–12. MIT Press.

Sparck-Jones, K. and Galliers, J. (1996) *Evaluating Natural Language Processing Systems: An Analysis and Review: Lecture Notes in Artificial Intelligence 1083*. Springer-Verlag.

Teufel, S. and Moens, M. (1999) Argumentative classification of extracted sentences as a first step towards flexible abstracting. In: Mani, I. and Maybury, M., editors, *Advances in Automatic Text Summarization*, pp. 155–171. MIT Press.

Tombros, A. and Sanderson, M. (1998) Advantages of query biased summaries in information retrieval. *Proceedings 21st ACM SIGIR Conference (SIGIR'98)*, pp. 2–10.

Voorhees, E. M. (1998) Variations in relevance judgments and the measurement of retrieval effectiveness. *Proceedings 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-98)*, pp. 315–323.

https://doi.org/10.1017/S1351324901002741 Published online by Cambridge University Press