Summarization Evaluation Methods: Experiments and Analysis

Hongyan Jing
Dept. of Computer Science
Columbia University
New York, NY, 10027, U.S.A.
(hjing@cs.columbia.edu)

Regina Barzilay
Dept. of Computer Science
Ben-Gurion University
Be'er-Sheva, 84105, Israel
(regina@cs.bgu.ac.il)

Kathleen McKeown
Dept. of Computer Science
Columbia University
New York, NY, 10027, U.S.A.
(kathy@cs.columbia.edu)

Michael Elhadad
Dept. of Computer Science
Ben-Gurion University
Be'er-Sheva, 84105, Israel
(elhadad@cs.bgu.ac.il)

Abstract

Two methods are used for evaluation of summarization systems: an evaluation of generated summaries against an "ideal" summary and evaluation of how well summaries help a person perform in a task such as information retrieval. We carried out two large experiments to study the two evaluation methods. Our results show that different parameters of an experiment can dramatically affect how well a system scores. For example, summary length was found to affect both types of evaluations. For the "ideal" summary based evaluation, accuracy decreases as summary length increases, while for task based evaluations summary length and accuracy on an information retrieval task appear to correlate randomly. In this paper, we show how this parameter and others can affect evaluation results and describe how parameters can be controlled to produce a sound evaluation.

Motivation

The evaluation of an NLP system is a key part of any research or development effort and yet it is probably also the most controversial. In the field of automatic summarization, most papers address evaluation by first stating how hard the problem is and then by applying methods that the developers consider appropriate for the task. Not only does this kind of individual system evaluation make comparison across systems impossible, but the evaluation results often fail to serve their purpose: readers can not infer how well a system performs by examining the results from a non-standard evaluation administered by the developers. With summarization emerging as a fast-developing new research area, there is an urgent need for a good evaluation methodology.

This paper is an analysis of current and proposed evaluation techniques. For each approach, we aim to identify the factors that can possibly affect the final evaluation results and how they affect the results, the appropriateness of the measures used to evaluate the performance, and advice for future experiments. We conducted two relatively large experiments to explore the problems, collecting 16 summaries for each of 40 documents selected. Thus, our evaluation testbed includes 640 summaries in total; among them, 400 were built by human subjects, 240 were generated by automatic summarization systems. We also used 13 human subjects to evaluate the performance of the 640 summaries in an information retrieval task. As previous research (Jones 1994) indicated: there is no easy way to evaluate systems, nor any magic numbers that can encapsulate performance. The goal of this research is not to judge which is the best evaluation method, but to understand more about each methodology so that future evaluations will be based on sounder assumptions and we'll be able to chose methods that most suit our task.

Evaluation of summarization systems can be intrinsic or extrinsic (Jones & Galliers 1996). Intrinsic methods measure a system's quality; extrinsic methods measure a system's performance in a particular task. Most evaluations of summarization systems to date are intrinsic: the quality of the summaries are judged by direct human judgment of, for example, informativeness, coverage, or fluency, or by comparing them with an "ideal" summary. Recently, a few task-based evaluations were also presented. The performance of the summarization system was evaluated in an information retrieval or news analysis task.

Most previous evaluations, however, used only a small set of data. The only two substantial evaluations we are aware of are (Edmunson 1969) and (Brandow, Mitze, & Rau 1995). Edmunson (Edmunson 1969) selected 200 documents in chemistry and compared the abstracts produced by professional abstractors with the outputs of 4 extraction methods. Brandow et al

(Brandow, Mitze, & Rau 1995) asked professional news analysts to score the summaries of 250 documents by automatic summarizers and "lead summaries" (only the first portion of the text), and unexpectedly found that "lead summaries" outperformed the "intelligent" summaries significantly.

Both experiments aimed at evaluating different systems under the same standard. In contrast, our approach is to evaluate a system using different evaluation methods. By examining evaluation results of the same system by different methods, we try to identify the factors that can influence the result of an evaluation method. We carried out 2 experiments: one to study intrinsic methods, *i.e.*, "ideal" summary based evaluation methods; the other to study extrinsic methods, *i.e.*, task-based evaluation. Three automatic summarizers and 40 documents were used in both experiments. Around 30 human subjects were involved in different tasks at different stages. The first experiment used 16 summaries produced for each document: 5 summaries at 10% length of the full document by human subjects, 5 at 20% length by human subjects, 3 at 10% by automatic summarizers, and 3 at 20% by automatic summarizers. This database of 640 summaries serves as our basis for studying "ideal" summary based evaluation methods. In the second experiment, we compared the performance of the summaries versus full texts in an information retrieval task, with 13 human judges involved. We're not aware of other large-scaled efforts to collect and analyze comparable data for both intrinsic and extrinsic evaluations.

In section 2, we describe the experiment on "ideal" summary based evaluation methods. We introduce previous work, describe the design of the experiment, and present and analyze the results. In section 3, we describe the experiment on task-based evaluation methods. The last section is conclusion.

"Ideal" Summary Based Evaluation

Previous Work

Most evaluations of summarization systems use an intrinsic method (Edmunson 1969; Paice 1990; Kupiec, Pederson, & Chen 1995; Marcu 1997; Salton *et al.* 1997; Ono, Sumita, & Miike 1994). The typical approach is to create an "ideal" summary, either by professional abstractors or by merging summaries provided by multiple human subjects using methods such as majority opinion, union, or intersection. The output of the summarizers is then compared with the "ideal" summary. Precision and recall are used to measure the quality of the summary. The main problem with this method is obvious and is also mentioned in other papers (Edmunson 1969; Paice 1990; Hand 1997): there is no single correct summary. Johnson (Johnson *et al.* 1993) proposed matching a template of manually generated key concepts with the concepts included in an abstract, but again, there is no single correct template of key concepts and matching of concepts is a fuzzy problem too.

Description of the Experiment

One way to reduce the subjectiveness of the "ideal" summary is to introduce a statistical model: instead of using a summary generated by only one human subject, we can ask multiple human subjects to construct summaries and build an "ideal" summary by taking the majority opinion of the human constructed summaries. Agreement among human subjects then becomes an issue (Hatzivassiloglou & McKeown 1993; William, Church, & Yarowsky 1992). Another factor that can influence evaluation results is the length of the summary. The evaluation results of the same summarization system can be significantly different if summaries are cut at different length.

The goal of this experiment is to understand more about agreement among human subjects and how different parameters such as summary length influence evaluation results. To study agreement of human subjects, 40 documents were selected; for each document, 10 summaries were constructed by 5 human subjects using sentence extraction. Each subject constructed 2 summaries of a document: one at 10% length and the other at 20%. For convenience, percent of length was computed in terms of number of sentences. We measured the agreement among human subjects using different models, and compared the result with those presented by other researchers.

To evaluate how length can influence evaluation results, we chose 3 automatic summarizers (we'll call them System A, System B, System C thereafter) and ran them on all 40 documents to generate 10% and 20% summaries for each document. We then evaluated 3 systems using summaries at different length, and analyzed how changes in length can influence final result.

A total of 16 summaries were produced for each document. Table 1 shows a sample summary database for a document. The documents were selected from the TREC collection (Harman 1994) and were used for both this experiment and the following task-based evaluation experiment. They are news articles on computers, terrorism, hypnosis and nuclear treaties. The average length of the articles is 30 sentences and the average number of words per article is 477. Human subjects are graduate students in the Department of Computer Science at Columbia University, Cornell University, and Ben-Gurion University.

Results and Analysis

Agreement Among Human Subjects We measured agreement among human subjects using *percent agreement*, a metric defined by (William, Church, & Yarowsky 1992) for the word sense disambiguation task, but also used in other applications such as discourse segmentation (Passonneau & Litman 1993; Hearst 1994). Percent agreement is the ratio of observed agreements with the majority opinion to possible agreements with the majority opinion. For our

Doc. Num	536075							
Doc. Title	State-Sponsored Death Squads Blocking Third World Development							
sent num	Sub1 10-20	Sub2 10-20	Sub3 10-20	Sub4 10-20	Sub5 10-20	SysA 10-20	SysB 10-20	SysC 10-20
1	+ +	+ +	+ +		+ +	+ +	+ +	+ +
2	+ +	+ +	+		+ +		+ +	
3		+		+ +	+	+		+ +
4								
5						+		
6						+ +		+ +
7			+ +					+
8								+
9								
10							+	
11								
12								
13								
14				+				
15								
16	+			+ +				

Table 1: A sample summary database for a document

Length	Avg. Agreement	Std Dev	Max	Min
10%	96%	0.020	100%	87%
20%	90%	0.022	100%	83%

Table 2: Agreement among 5 human subjects for 40 documents

experiments, agreement among 3 or more subjects is a majority opinion. The total possible agreements with the majority opinion is the number of human subjects times the number of sentences in a document. Observed agreement equals the number of times that a subject's decision agrees with the majority opinion, including both the decision to extract the sentence and not to extract the sentence. The results are shown in Table 2.

We can draw two conclusions from the above results: (1) when human subjects are required to summarize an article within the *same* short length, they are quite consistent with what should be included for certain genres of articles, like newswire articles used in the experiement. This is indicated by the high percentage of agreement for the 10% length summaries. (2) The degree of agreement among human subjects tends to decrease as the length of summary increases. This is shown by lower percent agreement and higher standard deviation among *same* human subjects when summary length increases from 10% to 20% in the experiment.

The above observation agrees with those patterns noticed by other researchers (Johnson 1970): human subjects are quite consistent with respect to what they perceive as being the most important and the most unimportant but less consistent with respect to what they perceive as being less important.

The percent agreement in our experiment is surprisingly high compared to results presented by other researchers. Marcu (Marcu 1997) found percent agreement of 13 judges over 5 texts from Scientific America is 71%. Rath (Rath, Resnick, & Savage 1961) found that extracts selected by four different human judges had only 25% overlap. Salton (Salton *et al.* 1997) found that the most important 20% paragraphs extracted by 2 subjects have only 46% overlap. Two main reasons for this high percent agreement are the style of TREC articles and our restriction on uniform length. The 40 documents used in the experiment have very similar text structure: an introduction followed by details, examples, facts, then a secondary point, so on and so forth. The introduction is usually a good short summary of the article, while for other genres of text, this may not be the case. The other reason is our restriction on uniform summary length. This eliminates differences due to perception of optimal summary length; in most other experiments, researchers did not require human judges to create a summary of a given length. Only (Salton *et al.* 1997) has a strict restriction in terms of length, but the extracted unit is a paragraph instead of a sentence, so the result is not comparable.

Statistical Significance Using the same methodology in (Passonneau & Litman 1993; Hearst 1994; Marcu 1997), we applied Cochran's test to our data. For our application, Cochran's test evaluates the null hypothesis that the total number of human subjects extracting the same sentence is randomly distributed. Cochran's statistic Q approximates the χ^2 distribution with $j-1$ degrees of freedom, where j is the number of elements in the data set, for our application, the number of sentences in a document. Our results show that the agreement among subjects is highly significant. That is, the probability that human subjects extract the same sentence is much higher than would be expected by chance. For all 40 documents, the probability is very low: $p < 10^{-6}$.

Parameters We're interested in how different factors can affect final evaluation results. Summaries from human and automatic summarizers were combined in different ways as "ideal" and were evaluated in the experiment to simulate real evaluation experiments.

Summary Length We found that precision and recall are very sensitive to summary length. We evaluated 3 selected summarization systems using summaries at 10% and 20% length respectively. Precision and recall differ greatly as shown in Table 3. The "ideal" summary was constructed by taking the majority opinion of 5 human summaries at the same length.

The table is to show that the evaluation results of the *same* system can change dramatically when evaluated at different length. The comparision of precision and recall of different systems at the same length is not meaningful. One reason is that due to the special text structure of the documents we used in the experiment, a summarizer which always chooses the first couple of sentences has advantage. Another reason is that different methods in computing length of summary in different systems introduce noise.

While one system uses notion of number of sentences as we did in contruction of human summaries, the other

	System A		System B		System C	
	Prec	Recall	Prec	Recall	Prec	Recall
10%	33	37	61	67	46	64
20%	32	39	47	64	36	55

Table 3: Evaluation results of summaries at the same length

	System A		System B		System C	
	Prec	Recall	Prec	Recall	Prec	Recall
10%	26	34	50	69	36	66
20%	18	49	27	82	24	75

Table 4: Evaluation results of human summaries at different length

two use number of words or number of clauses. Thus for the same length summary, the number of sentences extracted by each summarizer may be slightly different, as shown in Table 1, this will subsequently influence precision and recall.

Length of Summaries by Human Subjects Many previous evaluations of summarization systems did not strictly restrict summary length as we did in the experiment. Human subjects stop at the length which they think optimal. To stimulate the evaluation, we take 2 subjects' 10% summary and the other 3 subjects' 20% summaries to build an "ideal" summary. Results in Table 4 show human subjects' perception about optimal length can change precision and recall significantly.

Different Human Subjects Another two experiments used summaries written by different subjects to build an "ideal" summary. The first experiment used 3 summaries built by human subjects 1 to 3, the second by subjects 3 to 5. Table 5 and Table 6 show that the results don't change much due to high percent agreement among human subjects.

The above 6 comparative experiments show that precision and recall measures are very sensitive to summary length, both the length of the summaries to be evaluated and the length of the summaries used to construct "ideal" summaries. In the case of high percent agreement among human subjects, changes of human subjects do not influence the result significantly. Most previous evaluations did not restrict the length of the summaries produced by human subjects; the length of the summaries to be evaluated are different across systems too. It is then hard for the reader to infer the performance of the system from the absolute precision and recall measures given by the system developers. How to standardize the evaluation process is yet to be explored.

	System A		System B		System C	
	Prec	Recall	Prec	Recall	Prec	Recall
10%	32	35	59	64	46	53
20%	32	40	48	63	37	54

Table 5: Evaluation results of summaries by different human subjects (subject 1, 2, 3)

	System A		System B		System C	
	Prec	Recall	Prec	Recall	Prec	Recall
10%	32	36	62	70	45	65
20%	31	38	46	57	39	56

Table 6: Evaluation results of summaries by different human subjects (subject 3, 4, 5)

Precision and Recall Measures

Precision and recall measures are used extensively in "ideal" summary based evaluation of summarization systems. We argue, however, that they are not appropriate for the summarization task due to their binary nature and the fact that there is no single correct summary. An example of how the binary precision and recall measures fail the task is as follows: suppose there are two sentences which are interchangeable in terms of producing a summary. If 5 human subjects extract sentences to build a summary, 2 subjects chose sentence 1, 2 chose sentence 2, and the 5th chose sentence 1 by chance. By majority method, sentence 1 will be an "ideal" sentence. In evaluation, if a summarization system chooses sentence 1, it wins; if it chooses sentence 2, it loses, although sentence 1 and 2 are interchangeable. This is obviously not a good measure. One solution to the problem is to identify these interchangeable sentences by their distribution or other clues and treat them as the same entity, and then apply precision and recall measures. Yet another approach is to extend the precision and recall definition and introduce fractions (Hatzivassiloglou & McKeown 1993). In the extended model, the agreement between two objects is not presented by 0 or 1, but by the degree they agree. For the above example, the summarization system which chooses sentence 1 will get a score of 3/5 and the one which chooses sentence 2 will get score 2/5, instead of 1 or 0 respectively. This model is more suitable for the summarization task since alternative sentences are possible in an "ideal" summary.

Task Based Evaluation

Previous Work

Extrinsic methods evaluate the performance of a summarization system in a given task, such as GMAT test (Morris, Kasper, & Adams 1992), news analysis (Miike *et al.* 1994) and information retrieval (Mani & Bloedorn 1997). TIPSTER III will also introduce task-based evaluation of text summarization. Their evaluation will include approximately three tasks intended to judge the utility and appropriateness of the generated summaries, and to provide a way to measure improvement consistently (TIPSTER Homepage `http://tipster.org/`).

(Hand 1997) describes more details of the proposed task-based evaluation under TIPSTER. A "categorization" task will be used to evaluate generic summarizers; systems will be scored on how well the summaries, in lieu of full text, can help users in categorizing documents into different topics. An ad hoc information retrieval task will be used to evaluate user-directed summarization systems. Time and accuracy will be used to measure system performance.

Description of the Experiment

Our task-based evaluation follows the summarization evaluation proposal under TIPSTER III (Hand 1997). In particular, we chose the ad hoc information retrieval task. Evaluations of this kind so far are (Mani & Bloedorn 1997), (Brandow, Mitze, & Rau 1995), and the very recent TIPSTER dry-run. To study the methodology, we evaluated summaries from 3 summarizers at 10% and 20% length, summaries by a human subject at 10% and 20% length, keywords, titles, and full text (a total of 11 forms for each document) in an information retrieval task.

The ad hoc retrieval task evaluates a summarization system as a back end to an information retrieval engine. A summary that is tailored to the user query instead of full text will be presented to the user to decide the relevance of the retrieved document to the query. Time and accuracy will be measured to evaluate system performance. The ad hoc retrieval task is intended to evaluate user-directed summaries. However, summarizers that are capable of generating a tailored summary based on the user's input are unavailable in the experiment, so we used generic summarizers. To minimize the influence of this difference, we constrain the choice of queries and documents in a way that the query is relevant to the document main topic as opposed to unimportant topics or facts, which might be included in a query-based summary but not in a generic summary. By this way, the difference between a generic summary and a user-directed summary is reduced.

Material and Methods Four queries were selected from the TREC collection:

203 "Status of nuclear proliferation treaties — violations and monitoring."

215 "What evidence is there of paramilitary activity in the U.S.?"

223 "What are the different techniques used to create self-induced hypnosis?"

234 "What was responsible for the great emergence of 'MICROSOFT' in the computer industry?"

For each query, 10 documents were selected from the documents retrieved by the SMART system (Buckley 1993) (These are the same 40 documents in the first experiment). Human resources do not allow us to do a larger scale experiment. Although 10 is a small number compared to 200 documents per query in TIPSTER proposal, the experiment aims to be a pilot study of the problems invloved in the process instead of making conclusive statements.

	10% SysA	10% SysB	10% SysC	Full
Query1	Subj.-1	2	3	4
Query2	4	1	2	3
Query3	3	4	1	2
Query4	2	3	4	1

Table 7: Distribution of queries and documents among group members

We want to evaluate how the performance of human subjects in deciding the relevancy of the queries and the documents changes when they are presented with summaries from different systems at different lengths versus full-text.

Since each document has 11 forms and each human subject can not read a query-document pair more than once, we did the experiment in 3 groups. Each group has 4 human subjects, all 4 queries, and 40 documents in 4 different forms, with each 10 documents related to a certain query in the same form. Subjects were asked to decide the relevance of the documents to a query and write down the total time they spent on 10 documents for each query.

The document forms in 3 groups are:

Group1: 10% length summaries from 3 summarizers, full text

Group2: 20% length summaries from 3 summarizers, full text

Group3: 10%, 20% length summaries from a human subject, keywords, full-text

Keywords (the 5 most frequent nouns) and a human subject's summaries are used as baselines.

In each group, 4 human subjects rotated on 4 forms of documents and 4 queries. An example of Group 1 is given in Table 7.

We assume that performance of a subject is consistent during the experiment; since each subject contributes to evaluation of 4 forms of documents, the bias of any individual in the group is minimized.

All summaries and full text were presented without a title.

Evaluation Criteria

- Time Required
- Accuracy of decision
- Confidence of decision (additional to TIPSTER proposal)

Results and Analysis

The results are shown in Tables 8 to Table 11. Numbers in parentheses are percentages of those for full-text. To

	System A	System B	System C	Full
Precision	.56 (70%)	.65 (82%)	.45 (60%)	.79
Recall	.52 (66%)	.56 (71%)	.57 (72%)	.79
F-measure	.54 (68%)	.60 (76%)	.50 (63%)	.79
Time (min)	21 (46%)	33 (71%)	31 (67%)	46
Confidence	.77 (83%)	.66 (70%)	.89 (96%)	.93

Table 8: 10% summaries by auto-summarizers

	System A	System B	System C	Full
Precision	.68 (71%)	.75 (78%)	.89 (93%)	.95
Recall	.41 (48%)	.51 (60%)	.79 (93%)	.85
F-measure	.51 (56%)	.60 (67%)	.84 (93%)	.90
Time (min)	22 (44%)	29 (58%)	28 (56%)	50
Confidence	.64 (78%)	.67 (82%)	.71 (87%)	.82

Table 9: 20% summaries by auto-summarizers

account for different average abilities of subjects in different groups, we used the relative percentage in the parentheses for evaluation. Precision measures for all documents that are considered relevant by a reader, the percentage of documents that are really relevant. Recall measures for all documents that are relevant, the percentage that are considered relevant by a reader.

Statistical Significance Using the same technique as mentioned in section 2.3.2, we applied Cochran's test to our data. For this task, Cochran's test evaluates the null hypothesis that the total number of human subjects deciding a document is relevant to a query is randomly distributed. Our results show that the agreement among subjects is significant, i.e, the probability that human subjects consider a document as relevant is much higher than would be expected by chance. For all four queries, the probability $P < 10^{-5}$.

Cutoff Length Figure 1 shows how performance measures change with length. Figure 1.A shows precision of 3 summarizers at 10% and 20% length. We first standardized the notion of percentage of original text. In the summary extraction process, different criteria were used to compute percentage of original text: System A computes percentage of words, System B computes percentage of number of sentences, and System C computes percentage of number of clauses Human summary constructors are instructed to use percentage of number of sentences for convenience. By counting the total number of words in all summaries provided by a

	Human-10	Human-20	Keywords	Full
Precision	.60 (81%)	.47 (63%)	.55 (74%)	.74
Recall	.68 (100%)	.44 (65%)	.59 (86%)	.68
F-measure	.64 (90%)	.45 (63%)	.57 (80%)	.71
Time (min)	27 (40%)	30 (45%)	15 (23%)	66
Confidence	.86 (113%)	.63 (82%)	.72 (97%)	.76

Table 10: Summaries by human subject, keyword

	Full	Title
Average precision	.82	.50 (61%)
Average recall	.77	.25 (32%)
F-measure	.79	.33 (41%)
Average time (min)	54	4 (7%)

Table 11: Summaries using titles only

	Criteria used	10%	20%
System A	num. of words	10%	20%
System B	num. of sentences	12%	23%
System C	num. of clauses	14%	23%
Human	num. of sentences	17%	26%

Table 12: Length normalization

system and dividing it by the total number of words in all full texts, we converted percentage of number of sentences and clauses to percentage of words. For example, for a 10% summary, System B always extracts 10% sentences of original text, but the average number of words is actually 12% of the original text. The conversion table is shown in Table 12 (it shows that sentences extracted by human subjects are much longer than average).

Figure 1.A shows precision of each system at different length. It indicates that precision of each system is sensitive to the length of summary. While human generated summaries have a much higher precision for shorter summaries, some systems have a much higher precision for longer summaries (*e.g.*, System C). Figure 1.B shows that recall is also sensitive to the length of summaries. It can either increase or decrease while the length of summary increases.

The time required for different length of summaries is shown in Figure 1.C. This indicates that the time is not proportional to the length of the summary. A 10% summary does not require only 10% of the time as required by full text. In this experiment, people save most time when they read human generated summaries, with 17% of original words and 40% percent of time. The figure also shows that for some systems the time spent can decrease as the length of summaries increases. This result goes against the intuition that longer text will take longer time to read. One possible reason is that auto summarizers are still not at the stage of generating cohesive and coherent summaries, so it's possible that when a summary contains more sentences, it is more readable for human judges and thus they actually use less time. Evidence shows that although human generated summaries are longer, they are generally more coherent and as shown in Figure 1.C, they take the least amount of time to read. Another reason is that the length difference in our experiment is small. 10% and 20% summaries don't differ much in length since the articles are short. The recent TIPSTER workshop shows that when the length increases in larger amounts, time

	Query 1		Query 2		Query 3		Query 4	
	Avg	Stdev	Avg	Stdev	Avg	Stdev	Avg	Stdev
Precision	100	00	57	32	58	38	50	13
Recall	68	20	59	31	61	43	59	27

Table 13: Query performance statistics

went up for all systems. Our results show that when the length difference is small, time can possibly decrease as length increases.

Figure 1.D shows people's confidence in making decisions can also increase or decrease with length of summaries. According to the above four figures, precision, recall, time and confidence can all increase or decrease as the length of summary increases in a small amount.

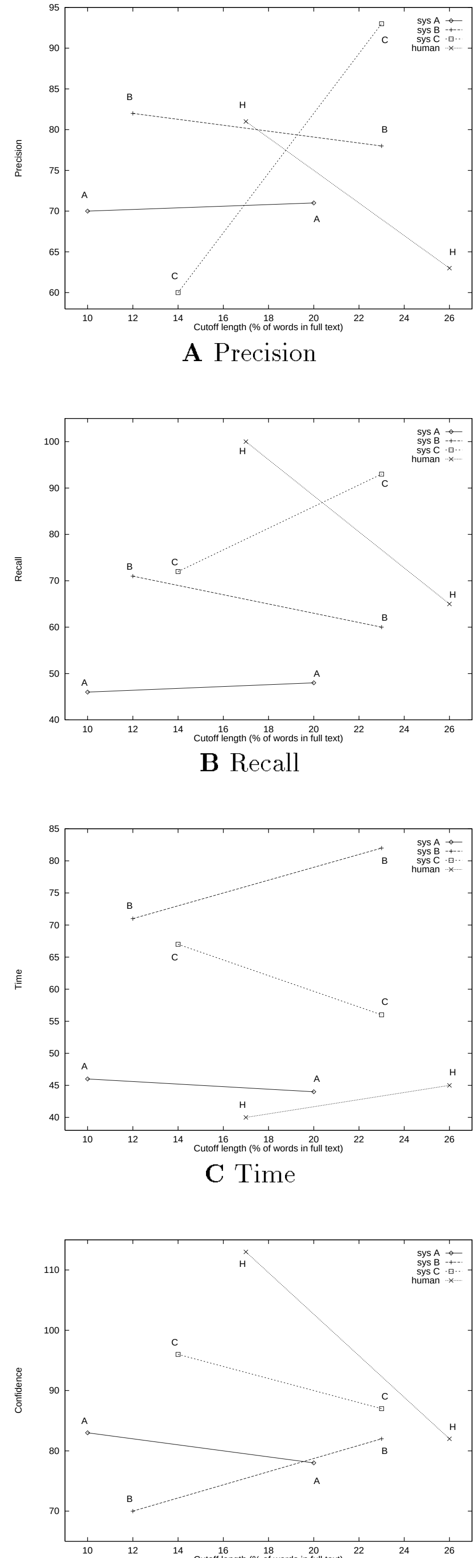
This randomness poses a question to the proposed cutoff length requirement of the ad hoc retrieval task based evaluation method. In the proposed TIPSTER summarization evaluation, a cutoff length will be established; any summary exceeding that margin will be truncated. But our experiments show the values of each system performance measure can either increase or decrease when the length of summary changes in a certain small range. That is, a system can achieve its best performance at different lengths. By setting a cutoff length, we're not evaluating which is the best system for the task, but which is best while all summaries are at a given length. If a different cutoff length is chosen, the results can be different.

To truncate a summary for *every* article at a cutoff length for evaluation is also questionable. The optimal length of a summary highly depends on the user needs, the genre of the article, and even within the same genre, the particular article. This length limit does not give an advantage to systems that are smart enough to identify the dependencies of the text.

According to the above, we suggest eliminating the cutoff length requirement currently proposed for the TIPSTER evaluation. Instead, systems should decide where to stop. Summarization systems that can help more in the task and in less time is most suitable, no matter how long the summaries are.

Query Selection In Table 13 we present average performance and standard deviation of methods according to the query posed.

Table 13 shows that queries can be divided into "easy" and "hard". An example of an "easy" query is the first query; high average precision and recall with a very low standard deviation indicate that all systems do equally well on this query. Therefore, use of this query will not make a clear distinction between summarization systems. Since selection of query is key to task-based evaluation, recognizing "easy" quaries is necessary for a sound experiment. Exactly how to recognize such queries is a topic for future work.



A Precision

B Recall

C Time

D Confidence

Figure 1: Performance / Length Relations

Text Selection We get relatively high performance by using keywords as a document representation (Table 10); precision and recall are 74% and 86% of using full text, while time is only 23%. This indicates that for some texts their relevance to a query is apparent by only examining words used in text. In such cases, randomly chosen sentences have a good chance to provide enough information to make the right decision; in other words, good and bad summarizers will not be separated by the evaluation. This kind of document is not useful in evaluating performance of summarization systems and should be removed. These documents can be identified by low difference of human performance while using full text or only a list of keywords.

Title In some texts, titles are very good indicative summaries themselves and judges can decide the relevance of documents simply by reading titles alone. This implies that providing a summary with title can affect the relevance decision. For this reason, we presented all documents to judges without titles. To verify this hypothesis, we asked a human subject to decide the relevance of documents by title alone. Results of this experiment are presented in Table 11.

These results indicate that for some documents we selected, titles are very informative. We suggest that future experiments should test the informativeness of document titles in an IR-task based evaluation and remove them if they contain too much information for making decisions. Without removing the titles, evaluations will not test the summaries themselves.

Conclusions

We carried out two large experiments on summarization evaluation to study the 'ideal" summary based method and the task-based method. While our work draws no conclusion about the different cases in which these two methods are applicable, it appears that both can be strengthened by setting parameters appropriately as we suggest. These two methods serve different purposes and thus, users should choose according to their own needs.

Our work does make very specific observations about the effects of different parameters on the outcome of evaluation. For the "ideal" summary based evaluation, our results indicate that reliability of evaluation decreases with summary length. Furthermore, they indicate that precision and recall are not the best measures for computing summary quality. This is due to the fact that a small change in the summary output (*e.g.*, replacing 1 sentence with an equally good equivalent which happens not to match majority opinion) can dramatically affect a system's score. We propose either explicitly accounting for such cases by marking the data or using an enhanced recall and precision measure which uses fractional counts (Hatzivassiloglou & McKeown 1993).

For the task-based evaluation, our results impact summary length, query used, and type of document. Using a uniform summary length across all systems and all texts appears to be detrimental. Our experiments show that there is no correlation between length and improvement in task; rather, it may be more beneficial to allow systems to set their own lengths. Future experimental design should furthermore consider the type of query used, avoiding queries where it is easy to determine relevance of the results and avoiding documents where keywords accurately characterize the text. By avoiding these cases, the evaluation will more accurately measure the impact of the summaries themselves on the IR task.

acknowledgement

This material is based upon work supported by the National Science Foundation under Grant No. IRI 96-19124, IRI 96-18797 and by a grant from Columbia University's Strategic Initiative Fund and by a grant from the US-Israel Binational Scientific Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

We would like to thank Eli Barzilay for his help in data organization. We would also like to thank all volunteers in Columbia, Ben-Gurion and Cornell Universities for participation in our experiments.

References

Brandow, R.; Mitze, K.; and Rau, L. F. 1995. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management* 31(5):675–685.

Buckley, C. 1993. The importance of proper weighting methods. In *Proceedings of ARPA Human Language Technology Workshop*.

Edmunson, H. P. 1969. New methods in automatic abstracting. *Journal of the ACM* 16(2):264–285.

Hand, T. F. 1997. A proposal for task-based evaluation of text summarization systems. In *ACL/EACL-97 summarization workshop*, 31–36.

Harman, D. 1994. In *An Overview of The Third Text Retreival Conference*. Gaithesburg, MD: National Institute of Standards and Technology.

Hatzivassiloglou, V., and McKeown, K. 1993. Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, 172–182.

Hearst, M. 1994. Multi-paragraph segmentation of expository text. In *Proceedings of the 32th AnnXBual Meeting of the Association for Computational Linguistics (ACL-94)*, 9–16.

Johnson, F. C.; Paice, C. D.; Black, W. J.; and Neal, A. P. 1993. The application of linguistic processing to automatic abstract generation. *Journal of Document and Text Management* 1(3):215–241.

Johnson, R. E. 1970. Recall of prose as a function of structural importance of linguistic units. *Journal of Verbal Learning and Verbal Behaviour* 9:12–20.

Jones, K. S., and Galliers, J. R. 1996. *Evaluating natural language processing systems: an analysis and review*. New York: Springer.

Jones, K. S. 1994. Towards better nlp system evaluation. In *Proceedings of the Human Language Technology Workshop*, 102–107. San Francisco: ARPA.

Kupiec, J.; Pederson, J.; and Chen, F. 1995. A trainable document summarizer. In *SIGIR '95*, 68–73.

Mani, I., and Bloedorn, E. 1997. Multi-document summarization by graph search and matching. In *Proceedings of AAAI-97*, 622–628.

Marcu, D. 1997. From discourse structures to text summaries. In *ACL/EACL-97 summarization workshop*, 82–88.

Miike, S.; Itoh, E.; Ono, K.; and Sumita, K. 1994. A full-text retrieval system with a dynamic abstract generation function. In *SIGIR '94*, 152–161.

Morris, A. H.; Kasper, G. M.; and Adams, D. A. 1992. The effects and limitations of automated text condensing on reading comprehension. *Information Systems Research* 3(1):17–35.

Ono, K.; Sumita, K.; and Miike, S. 1994. Abstract generation based on rhetorical structure extraction. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING-94)*, volume 1, 344–384.

Paice, C. D. 1990. Constructing literature abstracts by computer: techniques and prospects. *Information Processing and Management* 26(1):171–186.

Passonneau, R. J., and Litman, D. J. 1993. Intention-based segmentation: human reliability and correlation with linguistic cues. In *Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics (ACL-93)*, 148–155.

Rath, G. J.; Resnick, A.; and Savage, R. 1961. The formation of abstracts by the selection of sentences: Part 1: sentence selection by man and machines. *American Documentation* 12(2):139–141.

Salton, G.; Singhal, A.; Mitra, M.; and Buckley, C. 1997. Automatic text structuring and summarization. *Information Processing and Management* 33(2):193–208.

William, G.; Church, K. W.; and Yarowsky, D. 1992. Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics (ACL-92)*, 249–256.