Do Summaries Help?
A Task-Based Evaluation of Multi-Document Summarization

Kathleen McKeown
Columbia University
kathy@cs.columbia.edu

Rebecca J. Passonneau
Columbia University
becky@cs.columbia.edu

David K. Elson
Columbia University
delson@cs.columbia.edu

Ani Nenkova
Columbia University
ani@cs.columbia.edu

Julia Hirschberg
Columbia University
julia@cs.columbia.edu

ABSTRACT

We describe a task-based evaluation to determine whether multi-document summaries measurably improve user performance when using online news browsing systems for directed research. We evaluated the multi-document summaries generated by Newsblaster, a robust news browsing system that clusters online news articles and summarizes multiple articles on each event. Four groups of subjects were asked to perform the same time-restricted fact-gathering tasks, reading news under different conditions: no summaries at all, single sentence summaries drawn from one of the articles, Newsblaster multi-document summaries, and human summaries. Our results show that, in comparison to source documents only, the quality of reports assembled using Newsblaster summaries was significantly better and user satisfaction was higher with both Newsblaster and human summaries.

Categories and Subject Descriptors

H.4.3 [**Information Systems Applications**]: Communication Applications—*Information browsers*

General Terms

Measurement, Experimentation, Human Factors

Keywords

text summarization, evaluation, user study, news browsing

1. INTRODUCTION

Research on multi-document summarization of news has seen a surge of activity in the past five years, with the development of many multi-document news summarization systems (e.g., [5, 11, 14, 16]) and several that run online on a daily basis [14, 16] generating hundreds of summaries per day. Summarization evaluation methodology has also been actively explored. Since 2001, DUC (Document Understanding Conference), a NIST-run annual evaluation conference, has organized quantitative evaluations of multi-document summarization systems which compare system content against a reference set of model summaries. The DUC corpus of clustered summary/document pairs has spurred research in evaluation methodology on automation [12], metrics [16], and new methods of comparison of multiple models that factor in perceived salience of information [15, 7].

A significant question remains: will the summaries generated by such systems actually help end-users to make better use of the news? Multi-document summaries should enable users to more efficiently find the information they need. To find out whether they do, we performed a task-based evaluation of summaries generated by Newsblaster, a system that provides an interface to browse the news, featuring multi-document summaries of clusters of articles on the same event. We hypothesized that multi-document summaries would enable end users to more effectively complete a fact-gathering task. To this end, we compared the utility of four parallel news browsing systems: one with source documents but no summaries or clusters, one with one-sentence multi-document summaries where the sentence is extracted from one of the articles, one with Newsblaster generated multi-document summaries and one with human written summaries. Both Newsblaster and human summaries were multi-document summaries of the same length (about 200 words); where Newsblaster extracted all of its sentences, however, humans chose content and phrasing without typically using sentence extraction.

Our results show that, in comparison to source documents only, the quality of reports assembled using Newsblaster summaries was significantly better and user satisfaction was higher with both Newsblaster and human summaries. Users of Newsblaster and human summaries drew on summaries significantly more often in assembling their

SIGIR'05, August 15–19, 2005, Salvador, Brazil.

> The conflict between Israel and the Palestinians has been difficult for government negotiators to settle. Most recently, implementation of the "road map for peace," a diplomatic effort sponsored by the United States, Russia, the E.U. and the U.N., has suffered setbacks. However unofficial negotiators have developed a plan known as the Geneva Accord for finding a permanent solution to the conflict.
>
> - Who participated in the negotiations that produced the Geneva Accord?
> - Apart from direct participants, who supported the Geneva Accord preparations and how?
> - What has the response been to the Geneva Accord by the Palestinians and Israelis?

Figure 1: The prompt to one of the four tasks used in the evaluation.

report and were more satisfied, while providing reports of similar quality. More generally, our results demonstrate that full multi-document summarization is a more powerful tool than either documents alone or the one-sentence approach, an approach that is closely related to that used in systems such as Google News. They also provide a frame of reference as human summaries are presumably the best summary that could be provided.

In the following sections, we overview the relevant features of Newsblaster and then discuss the design, execution and results of our evaluation.

1.1 Overview of Newsblaster

Our study drew upon the following key components of Newsblaster:

1. **Article clustering.** Newsblaster clusters the articles it retrieves into *event clusters* about the same real-world topic. On a typical day, Newsblaster will identify about 100 such event clusters, which differ greatly in size. Some event clusters may contain only a handful of related articles, while major domestic or international stories can involve fifty or more.
2. **Event cluster summarization.** Newsblaster generates a concise overview of each event cluster. Its techniques for multi-document summarization include sentence extraction, reformulation and rewriting named entities for clarity.
3. **User interface.** Finally, Newsblaster presents the clusters, summaries, and links to source articles as a user-friendly, publicly accessible Web page. For the task evaluation, we developed a user interface that retained some of these features (e.g., see Figure 4).

We did not evaluate other features of Newsblaster, including search, "update" summaries, and related image collection.

2. RELATED WORK

There has been considerable recent work on multi-document summarization (see [6] for a sample of systems). Ours is distinguished by its use of multiple summarization strategies dependent on input document type, fusion of phrases to form novel sentences, and editing of extracted sentences. Our task-based or extrinsic [17] evaluation contrasts with most recent work on evaluation of summaries, which has focused on quantitative evaluation comparing generated summaries against a set of ideal reference models [6, 12, 15, 7]. There have also been earlier organized and individual task-based evaluations of single document summarization. TIPSTER-III [8] and others [13, 3] used an information retrieval task. Time and accuracy were measured to determine how well a user can judge the relevance of a retrieved document to a query. However, other factors such as summary length, type of query (some make it easy to determine relevance), and document type (when key words accurately characterize the text, these measures don't discriminate well between summaries) have a critical impact on task results [10].

As in our work, a recent evaluation [1] also asks subjects to write reports given a topic. However, they treat the resulting reports as focused summaries (reports are restricted to 50 sentences in length) and they evaluate how well different quantitative metrics compute similarity between the reports. Thus, their work evaluates evaluation metrics.

3. METHODS

We modeled our evaluation on an approach used by the DARPA TIDES program for an Integrated Feasibility Experiment (IFE) (see [4] for a description of the system architecture for this experiment). In the IFE, users are asked to write a report using a news aggregator as a tool. This task also resembles those that intelligence analysts carry out on a day-to-day basis [2] [9].

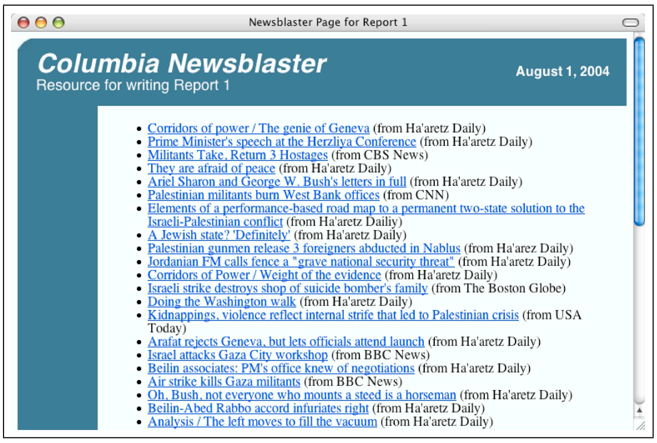
3.1 Evaluation Goals

In designing our user evaluation, we were interested in whether Newsblaster is an effective tool for assisting the processing of large volumes of news. We designed our evaluation to answer the following questions:

- Do summaries help the user find information needed to perform a report writing task?
- Do users use information from the summary in gathering their facts?
- Do summaries increase user satisfaction with the online news system?
- Do users create better fact sets with an online news system which includes multi-document summarization than one does that not?
- In the context of a news browser, what is the comparison of information quality in this task, and user satisfaction, when users have access to Newsblaster summaries versus minimal or human summaries?

3.2 Design

Each subject was asked to perform four 30-minute fact gathering scenarios using a Web interface. Each scenario involved answering three related questions about an issue in the news. These questions were presented to the user as part of a prompt, one of which is shown in Figure 1. The four tasks were, respectively: the Geneva Accord in the Middle East; Hurricane Ivan's effects; conflict in the Iraqi city of Najaf; and attacks by Chechen separatists in Russia. Subjects were given a space to compose their report and a Web page that we constructed as their sole resource. They were told to cut and paste facts from either the summaries or articles on the page, or to paraphrase to write a report. The page contained four document clusters, two of which were centrally related to the topic at hand, and two of which



Newsblaster Page for Report 1

Columbia Newsblaster

Resource for writing Report 1

August 1, 2004

- Corridors of power / The genie of Geneva (from Ha'aretz Daily)
- Prime Minister's speech at the Herzliya Conference (from Ha'aretz Daily)
- Militants Take, Return 3 Hostages (from CBS News)
- They are afraid of peace (from Ha'aretz Daily)
- Ariel Sharon and George W. Bush's letters in full (from Ha'aretz Daily)
- Palestinian militants burn West Bank offices (from CNN)
- Elements of a performance-based road map to a permanent two-state solution to the Israeli-Palestinian conflict (from Ha'aretz Dailiy)
- A Jewish state? 'Definitely' (from Ha'aretz Daily)
- Palestinian gunmen release 3 foreigners abducted in Nablus (from Ha'arez Daily)
- Jordanian FM calls fence a "grave national security threat" (from Ha'aretz Daily)
- Corridors of Power / Weight of the evidence (from Ha'aretz Daily)
- Israeli strike destroys shop of suicide bomber's family (from The Boston Globe)
- Doing the Washington walk (from Ha'aretz Daily)
- Kidnappings, violence reflect internal strife that led to Palestinian crisis (from USA Today)
- Arafat rejects Geneva, but lets officials attend launch (from Ha'aretz Daily)
- Israel attacks Gaza City workshop (from BBC News)
- Beilin associates: PM's office knew of negotiations (from Ha'aretz Daily)
- Air strike kills Gaza militants (from BBC News)
- Oh, Bush, not everyone who mounts a steed is a horseman (from Ha'aretz Daily)
- Beilin-Abed Rabbo accord infuriates right (from Ha'aretz Daily)
- Analysis / The left moves to fill the vacuum (from Ha'aretz Daily)

Figure 2: The evaluation interface screen showing the list of documents that a user sees in the no summary condition.

were peripherally related.[1] Hence there were sixteen clusters in the study overall. We selected the clusters by doing a manual search through Newsblaster clusters to find groups that were either peripherally or closely related. Each cluster contained, on average, ten articles. Subjects thus had to find relevant information within forty articles to answer in-depth analysis questions for each of four scenarios. While all of the articles were related to the scenario topic, only about half of the articles contained answers to the specific questions.

There were four summary condition levels in the experiment:

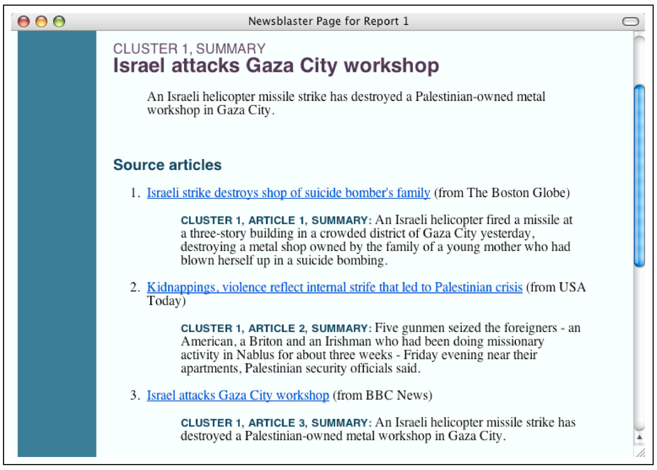
Level 1: Subjects were given no summaries. The Web page presented a list of document headlines (with no grouping by event cluster) that were relevant to the scenario (Figure 2). This included exactly the same documents that appeared in the four event clusters.

Level 2: Subjects were given a one-sentence summary for each source article, plus a one-sentence summary for each entire cluster. The per-document summary was generated by extracting the article's first sentence, an approach that is used as a baseline in evaluations [6]. The cluster summary was the one-sentence summary of the single article closest to cluster centroid. This is similar to approaches used in commercial online news systems such as Google News.

Level 3: Subjects were given a Newsblaster multi-document summary for each cluster.

Level 4: Subjects were given a human multi-document summary for each cluster. We hired summary writers from outside our research group to write summaries. Writers were recruited by posting a notice on a university job and career Web site that solicited applicants with good verbal skills, e.g., English or Journalism majors, students with high verbal scores on their GRE or SAT tests, or other evidence of writing ability.

Subjects had access to source documents in addition to the summaries. Links to the documents were available on



Newsblaster Page for Report 1

CLUSTER 1, SUMMARY

Israel attacks Gaza City workshop

An Israeli helicopter missile strike has destroyed a Palestinian-owned metal workshop in Gaza City.

Source articles

1. Israeli strike destroys shop of suicide bomber's family (from The Boston Globe)

   CLUSTER 1, ARTICLE 1, SUMMARY: An Israeli helicopter fired a missile at a three-story building in a crowded district of Gaza City yesterday, destroying a metal shop owned by the family of a young mother who had blown herself up in a suicide bombing.

2. Kidnappings, violence reflect internal strife that led to Palestinian crisis (from USA Today)

   CLUSTER 1, ARTICLE 2, SUMMARY: Five gunmen seized the foreigners - an American, a Briton and an Irishman who had been doing missionary activity in Nablus for about three weeks - Friday evening near their apartments, Palestinian security officials said.

3. Israel attacks Gaza City workshop (from BBC News)

   CLUSTER 1, ARTICLE 3, SUMMARY: An Israeli helicopter missile strike has destroyed a Palestinian-owned metal workshop in Gaza City.

Figure 3: The evaluation interface screen showing a typical page for the single sentence summary condition. The user sees the summary and a list of articles each with its own summary.

the same page when Summary Level 1 and Summary Level 2 were used (Figure 3) or by clicking on the cluster name when Summary Levels 3 and 4 were used (Figure 4).

Each scenario was followed by a survey that asked subjects to rate different aspects of their experience (e.g., difficulty of the task) along a five point scale, as well as some multiple choice questions. At the end of the experiment, each subject answered additional questions about their overall experience and had the opportunity to give comments.

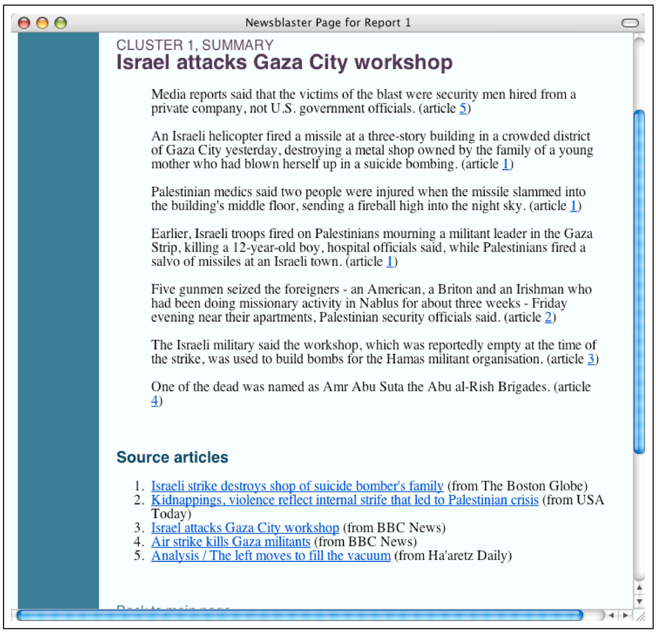
3.3 Study Execution

We recruited 45 subjects for three studies, where subjects wrote reports under the four different summary conditions noted above. Our subjects came from a variety of backgrounds: 73% were university students, of whom 32% were engineering students. The rest were undergraduate liberal arts students, journalism students, or law students. A pre-experiment questionnaire revealed that most used online newspapers as their primary news source, and read the news about an hour per day. All were native speakers of American English. Subjects were paid for their participation. An additional monetary prize was promised for the five writers whose reports scored the highest.

The subjects in the first study below alternated between Summary Level 3 and Level 4 (i.e., Newsblaster and human summaries); we controlled for scenario order and level order. The subjects in the next two studies had a single summary condition, Summary Level 1 (no summaries) or Level 2 (single-sentence summaries), and we controlled for scenario order.[2] Altogether, a total of 138 reports were written. We aimed at 11 subjects per summary level for each scenario (note that in Study A, subjects wrote for only two scenarios and thus we needed to double the number) and more subjects than expected showed up for Study C.

[1] With the exception of one of the scenarios, where one cluster was related and three were peripheral.

[2] The design included two order permutations. While this is not a complete crossed design, we found no effect of level or scenario order on report quality.



CLUSTER 1, SUMMARY

Israel attacks Gaza City workshop

Media reports said that the victims of the blast were security men hired from a private company, not U.S. government officials. (article 5)

An Israeli helicopter fired a missile at a three-story building in a crowded district of Gaza City yesterday, destroying a metal shop owned by the family of a young mother who had blown herself up in a suicide bombing. (article 1)

Palestinian medics said two people were injured when the missile slammed into the building's middle floor, sending a fireball high into the night sky. (article 1)

Earlier, Israeli troops fired on Palestinians mourning a militant leader in the Gaza Strip, killing a 12-year-old boy, hospital officials said, while Palestinians fired a salvo of missiles at an Israeli town. (article 1)

Five gunmen seized the foreigners - an American, a Briton and an Irishman who had been doing missionary activity in Nablus for about three weeks - Friday evening near their apartments, Palestinian security officials said. (article 2)

The Israeli military said the workshop, which was reportedly empty at the time of the strike, was used to build bombs for the Hamas militant organisation. (article 3)

One of the dead was named as Amr Abu Suta the Abu al-Rish Brigades. (article 4)

Source articles

1. Israeli strike destroys shop of suicide bomber's family (from The Boston Globe)
2. Kidnappings, violence reflect internal strife that led to Palestinian crisis (from USA Today)
3. Israel attacks Gaza City workshop (from BBC News)
4. Air strike kills Gaza militants (from BBC News)
5. Analysis / The left moves to fill the vacuum (from Ha'aretz Daily)

Figure 4: The evaluation interface screen showing a multi-document summary generated by Newsblaster. The user clicks on the cluster title to see the list of associated articles.

Study A: 21 subjects wrote reports for two scenarios each in two summary conditions: Level 3 and Level 4. Together, all four scenarios were covered.[3]

Study B: 11 subjects wrote reports for all four scenarios, using Summary Level 2.

Study C: 13 subjects wrote reports for all four scenarios, using Summary Level 1.

3.4 Scoring the Reports

As illustrated in Figure 1, subjects were asked to assemble lists of important facts that addressed a three-part prompt. We scored the quality of the resulting reports on the basis of how well subjects included appropriate content. To do this, we needed a gold standard and a metric for comparing report content against the gold standard. To score the reports, we used the Pyramid method for evaluation [15], which has been demonstrated to be a reliable method for summary evaluation. The method uses multiple models, thus making the report scores less sensitive to the specific model used. The Pyramid method allows an importance weight to be assigned to different information units, or content units. This is important for a subjective task such as report writing, where different facts are more or less important.

As a gold standard, we constructed a pyramid [15] of facts, or content units, for each scenario question for each summary level, using the reports written by the rest of the study participants for the same question. For example, to score a report from the human summary condition, we constructed the pyramid using reports created using all other conditions (i.e., no summaries, minimal summary, Newsblaster summary), plus the reports written by different people also with human summaries. This yielded, on average, 34 reports per pyramid, far greater than the number of summaries (five) needed to yield stable results [15]. Using this method, any fact (whether expressed as a word, a modifier, or a clause) that appears in more than one report is included in the pyramid. Facts that appear in more reports appear higher in the pyramid and are associated with a weight that indicates the number of times they are mentioned. Thus, more important facts have higher weight. If there are n reports, then there will be n levels in the pyramid. The top level will contain those facts that appear in all n articles, the next level facts that appear in $n-1$ articles, and so forth. A report SCU that does not appear in a pyramid has weight 0. Repetitions of the same SCU also have weight 0 and thus, duplication in a report does not increase the score. An ideal report of length x facts will include all facts from the top level, the next level and so forth, until x facts are included.

We score a report using the Pyramid scoring metric, which computes a ratio of the sum of the weights of report facts to the sum of the weights in an optimal report with the same number of facts. More formally, let T_j refer to the jth tier in a pyramid of facts. If the pyramid has n tiers, then T_n is the top-most tier and T_1, the bottom-most. The optimal score for a report with X facts is:

$$\text{MAX} = \sum_{i=j+1}^{n} i \times |T_i| + j \times (X - \sum_{i=j+1}^{n} |T_i|) \qquad (1)$$

where j is equal to the index of the lowest tier an optimally informative report will draw from. Then the pyramid score W is the ratio of D, the sum of the fact weights in the report, to MAX, the optimal report score.

This method has the following desirable properties:

- It avoids postulating an exhaustive ideal report, which would be impossible to reproduce in the 30 minute time frame of the study. In fact, we first attempted to construct such ideal reports, but each took us two days to construct. Instead, the pyramid predicts which facts are most important to include in the given time limit by comparing the choices of all participants.
- It predicts that there will be multiple reports of the same length that are equally good. For example, if two reports are of the same length and each draw on different facts from the same tier in the pyramid, they will receive equal scores.
- It takes into account the relative importance of facts according to the report writers themselves.

No specific instructions were given to the report writers about how long their reports should be. Consequently, some people wrote much longer answers to some report questions than did others. Figure 5 shows a histogram of lengths of answers to report questions, where length is measured in content units.[4] The wide variation in length of answers to any of the three questions from each of the four prompts was

[3] These subjects also wrote four reports in a session; thus all subjects saw all four scenarios in one of two orders. However, we were experimenting here with two additional control conditions for summary level that are not relevant for comparison with the two other studies reported here.

[4] Length also varied widely in number of characters, or words; as we score on the basis of content units, we compare lengths using this measure. See Section 3.4 or [15] for a description of content units.

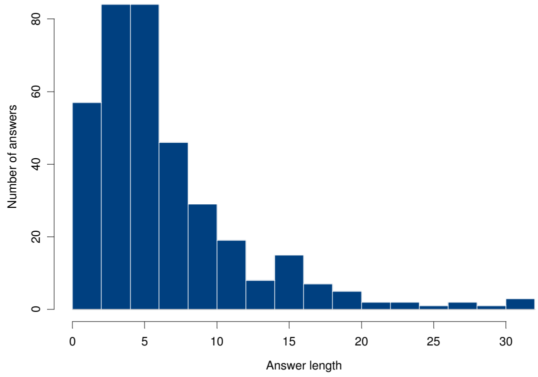


Figure 5: Distribution of the length of the reports in content units across all four conditions

unexpected, as it did not show up in our pilot study with far fewer subjects, on one scenario. It has been observed that report length has a significant effect on evaluation results [10]. To avoid the distortion that would arise from treating reports of such widely disparate length equivalently, we restricted the length of reports to be no longer than one standard deviation above the mean, removing outliers. To do this, we truncated all question answers to a length of eight content units, which was the third quartile of lengths of all answers.

3.5 Method of Analysis

We used analysis of variance (ANOVA) to study the impact of the type of summary level on report quality. The dependent variable was the score for each report and summary type was used as a factor with four levels: machine multi-document summary, human multi-document summary, no summary at all and minimal summary.

In addition to the main factor of interest (summary level), we included other factors in the model to estimate their contribution to the report quality. These factors were report writer, report topic, and question.

4. RESULTS

We measured effectiveness of the summaries in a fact-gathering task in three ways:

1. By scoring the reports and comparing scores across summary conditions;
2. By comparing user satisfaction per summary condition, as reported by the per-scenario surveys; and
3. By comparing whether subjects preferred to draw report content from summaries or from documents, measured by counting the citations they inserted following each extracted fact.

4.1 Content Score Comparison

The results of scoring reports for content are shown by summary condition in Table 1. The quality of reports tends to improve when subjects carry out the task with better quality summaries; the pyramid scores are lowest when the subjects use documents only and highest when the subjects use the human summaries. Differences between the scores are not significant (p=0.3760 from ANOVA analysis), but when we drop the scenario subjects had most difficulty with, differences are significant as noted below.

We suspected that we would see differences when we looked at different scenarios and, as we shall see, the ANOVA does show that scenario is a significant factor. When scoring reports and in informal discussion with subjects, we observed that some scenarios were more difficult than others; the documents in the clusters for these scenarios did not contain as much information for the answers. For example, in the Geneva Accord scenario, one subject wrote "The [user study] page brought up a large amount of useless articles and information, and especially on the last article [Geneva], only a few of the articles had any relevance [sic] at all." The Hurricane Ivan scenario, on the other hand, seemed one for which subjects could provide responses to questions. In retrospect, this may have been due to the fact that the event clusters for Geneva contained more editorials with less "hard" news, while the clusters for Hurricane Ivan contained more breaking news reports.

Given the problematic feedback on scenario 1 and the difference in types of documents in the clusters, we concluded that there was a design problem for this scenario. We removed the Geneva Accord scenario scores from the mix and recomputed averages of pyramid scores per summary condition as shown in Table 2. These results show that report quality is lowest with documents only, improves with minimal one-sentence summaries and improves again with Newsblaster summaries. The full ANOVA tables for all three scenarios apart from Geneva are shown in Table 3; the ANOVA table shows that summary level is a marginally significant factor in the results.

Our primary interest in the experiment was to measure the impact of the different multi-document summaries, determining exactly which summary levels made a difference. So, given the ANOVA model, we compared the report scores under each multi-document summary condition to those written under different summary conditions. 95% simultaneous confidence intervals for the comparisons were computed by the Bonferroni method. The only difference that was significant at the 0.05 level was that between Newsblaster summaries and no summaries at all. Thus, we conclude that report quality with Newsblaster summaries is significantly better than reports produced with documents only. The differences between Newsblaster and minimal or human summaries are not significant, although results with human summaries are slightly below Newsblaster summaries.

The ANOVA shows that scenario, question and subject are also significant factors in the result. Furthermore, there are significant interactions between summary level and scenario, between summary level and question, and between scenario and question.

4.2 User Satisfaction

Six of the questions in our exit survey required responses along a quantitative continuum. Each of the responses was assigned a score from 1 to 5 and a natural-language equivalent, with low scores corresponding to deep dissatisfaction and high scores expressing full satisfaction. For each question, Figure 6 shows the questions and the responses for each summary level at the extremes of the possible responses. It also shows the averages of the subjects' responses at the

Predictor	Df	Sum of Sq	Mean Sq	F Value	P-value
summary	3	0.304763	0.1015878	2.50084	0.0605509
scenario	2	1.188178	0.5940891	14.62501	0.0000012
question	2	1.151827	0.5759136	14.17757	0.0000017
user	42	4.290857	0.1021633	2.51501	0.0000098
summary:scenario	3	0.255512	0.0851706	2.09669	0.1018044
summary:question	6	0.739997	0.1233329	3.03615	0.0072526
scenario:question	4	0.313579	0.0783948	1.92989	0.1067646
summary:scenario:question	12	0.334852	0.0279043	0.68694	0.7631112

Table 3: ANOVA analysis of question score depending on summary level, scenario, question and user

Summary Level	Pyramid Score
Level 1 (documents only)	0.3927
Level 2 (one sentence summary)	0.3976
Level 3 (Newsblaster summary)	0.4377
Level 4 (Human summary)	0.4390

Table 1: Mean Pyramid Scores on Reports, all Scenarios included.

Summary Level	Pyramid Score
Level 1 (documents only)	0.3354
Level 2 (one sentence summary)	0.3757
Level 3 (Newsblaster summary)	0.4269
Level 4 (Human summary)	0.4027

Table 2: Mean Pyramid Scores on Reports, Scenario 1 (Geneva Accords) excluded.

Question	Level 2	Level 3	Level 4
1. Which was most helpful?			
source articles helped most	64%	48%	29%
equally helpful	32%	29%	29%
summaries helped most	5%	24%	43%
2. How did you budget your time?			
Most searching, some writing	55%	48%	67%
Half searching, half writing	39%	29%	19%
Mostly writing, some searching	7%	24%	14%

Figure 7: Multiple choice survey questions.

bottom of the table. This numeric representation of user satisfaction increases monotonically from Level 1 to Level 4.

Subjects were asked to compare their experience in the study with the experience they would expect to have on the same task using a Web search. Subjects were more likely to think the system they used was more effective than a Web search when they used Newsblaster than when they used either documents only (p=0.0798[5]) or single-sentence summaries (p=0.0101). Users were more likely to feel that they had read more than they needed to with documents only and with single-sentence summaries than with either Newsblaster summaries or human summaries. The difference for this question is marginally significant between subjects with human summaries and subjects with no summaries (p=0.0924)

Questions 3 and 4 show that subjects found it easier to assemble their facts with summaries than with documents only to complete the task and that they were more likely to feel they had enough time with summaries than with documents only. A pairwise χ^2 test shows the difference is marginally significant for question 3 between human summaries and no summaries (p=0.0838), is significant between one sentence summaries and no summaries (p=0.0401), although not quite significant between Newsblaster summaries and no summaries (p=0.1682). The difference for question 4 is significant, or marginally so, between each condition with a summary and no summaries (p=0.0636, Newsblaster/no summary; p=0.0126 human/no summary; p=0.0001 one sentence/no summary). There is no significant difference between different summary levels for either question 3 or 4. There were no significant differences between responses for the different summary levels for either question 5 or 6.

Responses to the multiple choice questions are shown in Table 7. Responses to question 1 again show that users were more satisfied with human level summaries than Newsblaster summaries and with Newsblaster summaries than with one sentence summaries. More than four times the proportion of subjects replied that summaries were more useful than source articles with Newsblaster summaries than with one sentence summaries. Responses to question 2 show that subjects spent the least time searching when given Newsblaster summaries, but unintuitively, the most time when given human summaries.

In the space for open comments, many subjects commented on the need for a method of searching the interface for events about particular keywords. An efficient searchable interface over summaries is being developed as part of Newsblaster, but was not evaluated in this study.

4.3 Citation patterns

The above results were echoed in the citation habits of subjects. When subjects wrote a report, they were asked to cite the location where they found a fact that they extracted for their report. We compared the number of times a subject extracted facts from source articles with extractions from summaries. The citations in Level 2 (one sentence summary) reports credited summaries 8% of the time.

[5] For the user study questions, significance is determined using a pairwise χ^2 test.

Question	Level 1	Level 2	Level 3	Level 4
1. Was the system better or worse than a general Web search would have been? 1. A Web search would have been a lot more effective 5. Newsblaster was a lot more effective than a Web search	3.37	3.39	4.14	4.10
2. What best describes your experience reading source articles? 1. I read a LOT more than I needed to 5. I only read those articles I needed to read	2.83	2.70	3.10	3.10
3. How difficult do you think it was to write the report? 1. Very difficult 5. Very easy	2.27	3.07	2.95	3.00
4. Do you feel like you had enough time to write the reports? 1. I needed more time 5. I had more than enough time	2.43	3.91	3.38	3.57
5. What best describes your experience using article summaries? 1. They had nothing useful to say about the topic 5. They told me everything I needed to know for the reports	n/a	3.16	3.29	3.43
6. Did you feel that the automatic summaries saved you time, wasted time, or had no impact on your time budget? 1. Summaries wasted time 5. Summaries saved me time	n/a	4.09	3.95	4.14
Average	2.75	3.39	3.47	3.56

Figure 6: The survey questions asked after each task, and the average responses per summary level.

For Levels 3 (Newsblaster) and 4 (human), the proportion was 17% and 27% respectively. This means that report writers were much more likely to reuse text from Newsblaster multi-document summaries than from the minimal summaries (p=0.0057 on one-sided t-test); there was a tendency to include more content when presented with human summaries than with Newsblaster summaries (p=0.1257 on one-sided t-test).

5. DISCUSSION

We began by asking whether multi-document summaries help users to find the information they need. Our user study shows that when a news browsing system contains such summaries, there is a significant increase over the no-summary condition in the quality of information that they include in their report. Users feel that they are able to find substantially more of the information that is relevant. This result demonstrates that summaries do help subjects do a better job of using news to assemble facts on given topics.

When we developed Newsblaster, we speculated that summaries could help users find necessary information in either of the following ways:

1. they may find the information they need in the summaries themselves, thus requiring them to read less of the full articles, or
2. the summaries may link them directly to the relevant articles and positions within the articles where the relevant information occurs.

Our user study confirms these beliefs and shows that as the quality of the summary increases (from Level 2 to Level 4), the greater the effect. The increase in citations shows that as quality of the summary increases, users significantly more often find the information they need in the summary without a significant decrease in report quality. At the same time, they report that they read fewer articles when they have either a Level 3 or a Level 4 summary. This confirms our belief that better multi-document summarization saves reading time and facilitates finding the relevant documents. This is further reinforced by the fact that almost five times the proportion of subjects using Level 3 summaries than those using Level 2 summaries reported that summaries were more helpful for the task than source articles. That number almost doubles again for subjects with Level 4 summaries.

There are some issues we need to address in future studies. First, we expected to find a significant increase in report quality as summary quality increased. We only found a significant increase in quality between reports written with documents only and reports written with summaries. The lack of significant increase between Summary Level 2 and Summary Level 3 could be due to a number of factors. There were two problems in presentation of information for these two levels. First, the interface for Summary Level 2 identified individual articles with a title and a one sentence summary; we modeled this design after commercial online news providers. The interface for Summary Level 3 only had titles for each article. In order to pinpoint the effect of different quality summaries on report quality, we need to run a follow-on study which compares how subjects do with a single sentence multi-document summary paired with a list of article titles only and how they do with a Newsblaster summary paired with a list that contains both titles and one-sentence, single document summaries. Second, the interface for Summary Level 2 shows the list of individual articles on the same Web page as the multi-document summary for the cluster. In contrast, the interface for Summary Level 3 shows the multi-document summary and cluster title on the same page and requires the subject to click on cluster title to see the list of individual articles. The different number of clicks required in the interface may have affected time-to-task completion as well as search strategy.

Another problem that we noted was that reports written by subjects were of widely varying length. Reports varied from a minimum of 102 words to a maximum of 1525 words. We adjusted for this in the current study by truncating reports. Lengthy reports not only had more material, but tended to have more duplication of facts, which clearly makes for less effective reports. The impact of truncating reports requires follow-up study. We plan to correct for these two problems with more specific directions about the length and nature of report required. We also will experiment with modifications of the task so that subjects will write coherent reports, rather than cut and paste sentences from documents. We hypothesize that this will require more synthesis of material, and lead to more consistency in length.

In order to have a realistic task-based evaluation, we developed complex prompts across a range of topics. As a consequence, we could simultaneously investigate a wide range of factors. Given that scenario and question had significant effects on report quality, we need to understand more clearly how the four scenarios contrast, and how question difficulty compares within and across prompts. It is also possible that variables we did not explicitly test for, such as cluster size, article length, semantic coherence within clusters, or semantic distance between clusters, influenced the outcome.

6. CONCLUSIONS

We have shown that it is feasible to conduct a task-based, or extrinsic, evaluation of summarization that yields significant conclusions. Our answer to the question, *Do Summaries Help?*, is clearly yes. Our results show that subjects produce better quality reports using a news interface with Newsblaster summaries than with no summaries. Also, as summary quality increases from none at all to human, user satisfaction increases. In particular, full multi-document summaries, of which Newsblaster and human summaries are representative, help users perform better at fact-gathering than they do with no summaries. Users are also more satisfied with multi-document summaries than with minimal one-sentence summaries such as those used by commercial online news systems. These results affirm the benefit of research in multi-document summarization.

However, we have also demonstrated that many factors influence the degree to which summaries help. A complete answer to the question is clearly complex, and a single study can only give partial insight. A secondary contribution of our experiments is the identification of additional possible effects on task completion, such as specific interface design, report length, and scenario design, none of which were predicted by our pilot. These insights provide a road-map for follow-on studies that can even more finely pinpoint the effect of multi-document summaries on task performance.

7. ACKNOWLEDGEMENTS

We would like to thank Andrew Rosenberg for his help with the statistical analysis in an early stage of the evaluation. This work was supported in part by the Defense Advanced Research Projects Agency under TIDES grant NUU01-00-1-8919.

8. REFERENCES

[1] E. Amigo, J. Gonzalo, V. Peinado, A. Penas, and F. Verdejo. An empirical study of information synthesis tasks. In *Proceedings of ACL-04*, Barcelona, Spain, 2004.

[2] J. W. Bodnar. *Warning Analysis for the Information Age: Rethinking the Intelligence Process.* Center for Strategic Intelligence Research, Joint Military Intelligence College, Washington, D.C., 2003.

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

[4] S. Colbath and F. Kubala. Tap-xl: An automated analyst's assistant. In *Proceedings of HLT-NAACL 2003)*, Edmunton, Alberta, Ca., 2003.

[5] H. Daume, A. Echihabi, D. Marcu, D. S. Munteanu, and R. Soricut. Gleans: A generator of logical extracts and abstracts for nice sumamries. In *Proceedings of the Second Document Understanding Workshop (DUC-2002)*, Philadelphia, Pa., 2002.

[6] Proceeding of the second, third and forth document understanding conference, 2002,2003,2004.

[7] H. Halteren and S. Teufel. Examining the consensus between human summaries: initial experiments with factoid analysis. In *HLT-NAACL DUC Workshop*, 2003.

[8] T. Hand. A proposal for task-based evaluation of text summarization systems. In *Proceedings of ACL/EACL-97 Summarization Workshop*, pages 31–36, Madrid, Spain, 1997.

[9] F. J. Hughes and D. A. Schum. Evidence marshaling and argument construction: Case study no. 4, the sign of the crescent (analysis), January 2003. Manuscript developed for exclusive use by the Joint Military Intelligence College; not for distribution.

[10] H. Jing, R. Barzilay, K. McKeown, and M. Elhadad. Summarization evaluation methods: Experiments and analysis. In *AAAI Symposium on Intelligent Summarization*, 1998.

[11] C.-Y. Lin and E. Hovy. From single to multi-document summarization: A prototype system and its evaluation. In *Proceedings of the ACL*, pages 457–464, 2002.

[12] C.-Y. Lin and E. Hovy. Automatic evaluation of summaries using n-gram co-occurance statistics. In *Proceedings of HLT-NAACL 2003*, 2003.

[13] I. Mani and E. Bloedorn. Multi-document summarization by graph search and matching. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-97)*, pages 622–628, Providence, Rhode Island, 1997. AAAI.

[14] K. R. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, J. L. Klavans, A. Nenkova, C. Sable, B. Schiffman, and S. Sigelman. Tracking and summarizing news on a daily basis with columbia's newsblaster. In *Proceedings of 2002 Human Language Technology Conference (HLT)*, San Diego, CA, 2002.

[15] A. Nenkova and R. Passonneau. Evaluating content selection in summarization: The pyramid method. In *Proceedings of HLT/NAACL 2004*, 2004.

[16] D. R. Radev, S. Blair-Goldensohn, Z. Zhang, and R. Sundara Raghavan. Newsinessence: A system for domain-independent, real-time news clustering and multi-document summarization. In *Human Language Technology Conference (Demo Session)*, San Diego, CA, 2001.

[17] K. Sparck-Jones and J. R. Galliers. *Evaluating Natural Language Processing Systems: An Analysis and Review.* Springer, Berlin, 1995. Lecture Notes in Artificial Intelligence 1083.