Facilitating Physicians' Access to Information via Tailored Text Summarization

Noemie Elhadad, M.S.*, **Kathleen McKeown, Ph.D.***, **David Kaufman, Ph.D.**†, **Desmond Jordan, M.D.**‡

* Department of Computer Science
Columbia University
New York, NY 10027, USA

†Department of Medical Informatics
Columbia University
New York, NY 10032, USA

‡Departments of Anesthesiology and Medical Informatics
Columbia University
College of Physicians and Surgeons
New York, NY 10032, USA

We have developed a summarization system, TAS (Technical Article Summarizer), which, when provided with a patient record and journal articles returned by a search, automatically generates a summary that is tailored to the patient characteristics. We hypothesize that a personalized summary will allow a physician to more quickly find information relevant to patient care. In this paper, we present a user study in which subjects carried out a task under three different conditions: using search results only, using a generic summary and search results, and using a personalized summary with search results. Our study demonstrates that subjects do a better job on task completion with the personalized summary, and show a higher level of satisfaction, than under other conditions.

INTRODUCTION

When a clinician needs to know what the most recent research results suggest for treating or diagnosing a patient, one option is to conduct a search over a medical library such as PubMed. Traditional search engines, however, return a list of journal articles, not all of which are relevant, and the clinician must search for findings pertaining to her patient, which may be buried in any one of the articles. Previous research showed that search engines by themselves are not enough to meet the needs of physicians [1]: they lack flexibility and adaptability. We hypothesize that if a summary of the journal articles, tailored to the patient characteristics, is presented along with search results, the clinician can more easily find the information sought.

We have developed a summarization system, TAS (Technical Article Summarizer), that, when provided with a patient record and the articles returned by a search, automatically produces a summary of the input articles focusing on the findings relevant to the patient. In this paper, we present a user study that tests whether personalized summaries help clinicians more efficiently access the medical literature. Subjects were given the task of extracting all findings relevant to a patient and were presented with information under three different conditions: a list of articles returned from a search, a generic summary linked to the search results, and a personalized summary linked to the search results. We measured quantitative differences among the conditions (quality of task output), as well as qualitative differences (user satisfaction as revealed through responses to a post-study questionnaire and their video-taped comments when carrying out the task).

BACKGROUND

TAS was developed for physicians and physicians in training who need to access the literature in the context of treating a specific patient. Given the electronic patient record of the patient under care and a list of clinical studies, TAS generates a summary which contains findings that are reported in the input articles and that are relevant to the patient being treated [2]. Thus, the summaries are personalized to the patient's characteristics.

The main contributions of the summarizer are *personalization* and *generation*. The findings from the input articles are first extracted. Findings not pertaining to the given patient are then filtered out, personalizing the content of the summary to the patient. TAS merges the pertinent extracted pieces of information, identifying and highlighting repetitions and contradictions across the input articles. An ordering algorithm we developed places important information near the beginning, using lexical overlap to place related sentences near each other, yielding a coherent summary. Summary sentences are generated by re-using phrases in the input articles, yielding a fluent summary. Figure 1 shows an example of a personalized summary of four clinical studies. The patient in question is a 59 year old man with hypertension and hypercholesterolemia. He recently had atherosclerosis of the saphenous vein. The physician wants to know whether atherosclerosis could have been prevented.

Our study contrasts the use of a personalized and generic summary, with the goal of determining whether the use of summaries improves access to desired information, and if so, whether the improvement comes strictly from the presence of any summary or from personalization. Figure 2 shows portion of a summary of the same input articles, this time using a modified version of TAS which only extracts findings reported in the article, regardless of relevance. It does not filter, merge, order or generate a summary, but simply outputs the extracted sentences in the original order of the input articles. The extracted, generic summary is much longer (20 sentences in the generic summary vs. 4 sentences in the personalized summary) and poten-

Agressive lipid-lowering strategy and moderate low-density LDL-C lowering strategy were associated with atherosclerosis progression [1,2].

Predictors for atherosclerosis progression and graft worsening were stenosis of the graft, prior myocardial infarction, years post CABG, high triglyceride level, small minimum graft diameter, low HDL-C, high LDL-C, high mean arterial pressure, low ejection fraction, male gender, and current smoking [1,1].

There was no association between warfarin and progression of atherosclerosis [2,3].

Predictors of late MACE were unstable angina and CHF [4].

Figure 1. *A personalized, generated summary of four clinical studies. The numbers in brackets are pointers to specific sentences in the input articles.*

tially harder to read. Extracted sentences appear out of context, and the absence of an appropriate ordering strategy yields a less coherent text. Sentence extraction, however, is the primary approach used in text summarization today, and thus, represents a realistic baseline against which to compare personalization.

METHODS

Our study tests two hypotheses: (1) summaries help users access relevant information; and (2) personalized, generated summaries are better than generic, extractive summaries in accessing relevant information.

Study design

The study was designed as a task-based evaluation.[1] Each subject was presented with three independent clinical scenarios, each reviewed and validated by one of the authors, an experienced cardiac anesthesiologist: (1) a female patient with atrial fibrillation, (2) a male patient with atherosclerosis of the saphenous vein, and (3) a male patient who must undergo aortic valve replacement. A scenario consists of the latest discharge report of a patient record, a clinical question about the patient and a set of seven to eight preselected input articles. The articles were found by querying PubMed for clinical trials for queries related to the scenario. In order to ensure that the articles both reflected typical search engine results and constituted good input for the scenario, from among the search results, we selected the five articles judged by our expert to be most relevant to the scenario and two to three less relevant articles.

Each subject was asked to read the patient record and then select from among all the presented articles the findings relevant to the patient and question. To make the task more realistic, we asked subjects to complete each scenario in at most 15 minutes. Under this time

[1]All material presented to the subjects, including guidelines, interface, and questionnaire, is available at `http://www.cs.columbia.edu/~noemie/tas_eval`

In the order of their importance they were: maximum stenosis of the graft at baseline angiography; years post-SVG placement; the moderate low-density lipoprotein cholesterol (LDL-C) lowering strategy; prior myocardial infarction; high triglyceride level; small minimum graft diameter; low high-density lipoprotein cholesterol (HDL-C); high LDL-C; high mean arterial pressure; low ejection fraction; male gender; and current smoking [1]. [...]

There was a tendency toward less atherosclerosis on baseline angiography in the aggressively treated group, but the angiographic end points evaluating change from baseline take account of baseline status [2]. Warfarin had no beneficial effect on the progression of atherosclerosis in the LMCA [2].

Univariate predictors were restenotic lesion (odds ratio (OR): 2.47, confidence interval (CI): 1.13 to 3.85, P = 0.0003), unstable angina (OR: 1.99, CI: 1.27 to 2.91, P = 0.04) and congestive heart failure (CHF) (OR: 1.97, CI: 1.14 to 3.24, P = 0.02) for in-hospital MACE, and peripheral vascular disease (PVD) (OR: 2.18, CI: 1.34 to 3.44, P = 0.002), intra-aortic balloon pump placement (OR: 2.08, CI: 1.13 to 3.85, P = 0.02) and previous MI (OR: 1.97, CI: 1.14 to 3.25, P = 0.007) for late MACE [3]. Independent multivariate predictors for late MACE were restenotic lesion (relative risk (RR) 1.33, P = 0.02), PVD (RR: 1.31, P = 0.01), CHF (RR: 1.42, P = 0.01) and multiple stents (RR: 1.47, P = 0.004) [3]. [...]

There was no significant difference in angiographic outcome between the warfarin and placebo groups [4]. No significant differences in angiographic outcomes were observed between the warfarin and placebo groups [4]. [...]

Figure 2. *Portion of a generic, extractive summary for the same four clinical studies as in Figure 1.*

constraint, subjects did not have time to read each input article in its entirety and were forced to figure out a searching and reading strategy to help them identify relevant information efficiently. We implemented an interface that allowed subjects simply to click on a sentence in any input article to select it. The selected sentences are automatically displayed in a separate window, aggregating all the selections from the different articles, and are also highlighted inside each article. Articles are color-coded, so that the subject knows at any time which sentence came from which article. A screen shot is shown in Figure 3.

To verify our hypotheses, each subject was tested under three conditions:

Search A list of articles was provided to the subject, in the same format as the results of a search engine. For each article, title, publishing journal and year of publication were displayed.

Generic A generic, extractive summary (e.g., Figure 2) was provided to the subject. The references in the summary were pointers to specific sentences in the input articles that were extracted and included in the summary. The summary was followed by a list of articles in the same format as in the Search condition.

Personalized A personalized, generated summary (e.g., Figure 1) was provided to the subject. The

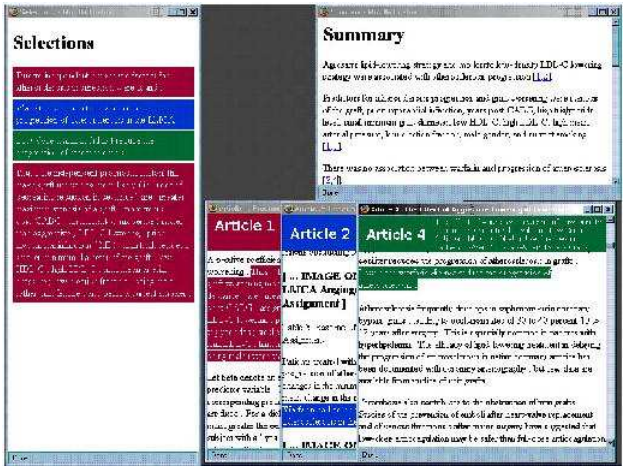

Figure 3. *The user interface for the study.*

references in the summary were pointers to specific sentences in the input articles from which information in the summary was drawn. References have the same function as in the Generic condition. The summary was followed by a list of articles in the same format as in the Search condition.

Each scenario was presented to each subject under one of the three conditions. Given three scenarios and three conditions, this meant that three subjects were required to yield a single data point for each scenario under each condition. The order in which conditions were presented was systematically varied. Thus the first subject saw Scenario 1 first under the Search condition, while the second subject saw Scenario 1 first under the Generic condition, etc.

Each scenario was followed by a questionnaire (see Figure 4). The questions were answered on a five-point scale, whith a high grade indicating a positive answer.[2] While the first four questions are about the scenario difficulty and the interface in general, Q5 through Q8 are about how the summary facilitated the task. Q9 was a free-text question for the subject to report any comment on the summary. Therefore, when tested under the Search condition, the subjects were asked to answer only questions Q1 to Q4.

Twelve subjects were recruited from the Cardiac Intensive Care Unit at New York Presbyterian Hospital, ranging from fourth year medical student to attending physician for 15 years. Subjects were financially compensated.

Analysis

Our analysis features objective measurement of how well the subjects did on the task in terms of quality of results, as well as a subjective evaluation which measures subject satisfaction with the system.

[2]For Q5, a high grade indicated that the sentences in the articles were accessed mostly through the references inside the summary

Objective evaluation To quantify each subject's performance on the task, before the study began, we collected a gold-standard set of findings for each scenario. Our medical expert read each of the articles in the three scenarios on paper and highlighted the sentences which conveyed findings pertinent to the patient, following the same guidelines as the ones given to the subjects. We imposed no time limit on the collection of the gold-standard, and it took approximately four hours to complete.

During the study, the interface stored which sentences were selected by the subject, as well as timing information. We scored a subject's selection in terms of its precision (the number of subject-selected sentences present in the gold standard divided by total number of subject sentences) and recall (the number of subject selected sentences present in the gold standard divided by total number of gold standard sentences) compared to the gold-standard. We combined precision and recall into the F_2 measure, which equally weights precision and recall using the harmonic sum of the two numbers. We used the ANOVA test to determine the impact of the testing conditions on the performance of the subjects, as measured by F_2. We also looked at the effect of scenario and subject to verify their possible impact.

Subjective Evaluation We relied on both the post-scenario questionnaire filled out by the subjects and the video transcript of the subjects to assess user satisfaction with interacting with the interface under the different conditions. We used the paired student's t-test to analyze the questionnaire answers. For the video, we asked subjects to think aloud and performed cognitive analysis on their behaviors [3]. Data used in the analysis included a) an annotated and time stamped transcript from the audio tape, b) video of the participant and video captures of the screen, and c) selected findings. The transcript was coded for actions, goals, inferences about the articles or scenarios, comments about the system, and expressions of uncertainty.

RESULTS

We report on the results of six subjects who completed the study out of the twelve recruited.[3]

Task performance analysis

All subjects but one spent all 15 minutes under every condition. We understand this result as a confirmation that the alloted time was a constraint, making the task more difficult to perform accurately without the benefit of a reading strategy. On average, the generic summaries contained 36.6 sentences and the personalized summaries contained 8.3 sentences.

Table 1 shows the mean F_2 measure by condition across the three scenarios. When presented with the

[3]Five of them failed to complete the study. We suspect this was due to diffi cult conditions in the ICU that day.

Table 1. *Mean subject performance per condition.*

Condition	F_2 Mean
Generic	13.9
Search	16
Personalized	27.7

personalized, generated summaries, the subjects performed the best. Their performance decreased when presented with the search interface without any summary. They performed the worst when presented with generic, extractive summaries. A paired student's t-test shows that performance with personalized summaries is significantly better than with generic summaries (p=0.07), and overall, the ANOVA analysis identified the condition to be a factor in variance (p=0.13). Interestingly, although the overall number of selected findings was similar for the Generic and Personalized conditions (mean of 13.5 vs. 14, not statistically significant), the selections made under the Personalized condition were far more accurate.

The two other factors entered in the model, scenario and user, did not contribute to the variance (p value of 0.37 and 0.32 respectively). In other words, the differences in scenarios and the individual users did not influence the performance.

Questionnaire analysis

The average answers to the questionnaire are given for each condition in Figure 4. The subjects had a positive reaction to the interface, independently of the condition; the answers for Q1 to Q4 were not significantly different across conditions. The subjects, however, showed a strong preference for the Personalized over the Generic summaries (Q5 to Q8) (p=0.001).

Cognitively-based video analysis

We illustrate the cognitive analysis by comparing a subject on two scenarios. In Case 1, the subject was presented with a generic summary and in Case 3, she was given a personalized summary. Each case was completed in less than 15 minutes. Figure 5 shows excerpts from Case 1.

Early on in the process, she finds the summary to be largely uninformative and unrelated to her goal of determining the best treatment options. After a few minutes, the clinician abandons the text summary and goes straight to the article index. Her approach is not systematic. This is in marked contrast to her performance when using the personalized summary as indicated in the excerpt shown in Figure 6.

When using the personalized summary, the clinician develops an effective strategy and uses it to select a total of 24 sentences as opposed to just 8 in the first case. The strategy is characterized by the following action pattern: a) review pertinent paragraph in summary, b) click on indexed article (from summary), c) select sentences to add to selection list, d) repeat c until all relevant non-redundant sentences from the article have been added, e) shift focus back to summary, f) select new article or same article with new entry point (e.g., the prior passage in summary may have indexed the results and the subsequent one the discussion). The personalized summary allowed her to easily identify candidate articles and go directly to the relevant (summary indexed) passages or sections within the article. The result was a greater number of selected sentences with a higher number of correct selections. In addition, she expresses greater satisfaction with results.

00:25 Action: Open and Reviews Summary
Comment: Some of it doesn't seem as relevant to the actual treatment options, What's the best treatment? Yeah. So, some of this stuff seems more descriptive.
*****
06:53 Action: Clicks on Article 3 from reference list.
07:05 Action: Selects sentence to be added to list.
07:38 Action: Clicks on article 5 from reference list
Comment: I'm kind of getting bogged down in the summary, in terms of this woman's presenting, she's now in afi b, so we're looking for more of a cardioversion treatment. So going right to the list, some of these articles are talking more about maintenance.

Figure 5. *Excerpts from a subject looking at a generic, extracted summary.*

00:54 Action: Opens Summary, immediately goes to list of index articles.
2:33 Clicks/Selects Article 1 from summary
Comment: From here I'll give the summary a try and see if I can fi gure out how to use it.
2:45 Selects sentence from article 1
3:11 Selects sentence from article 1
4:37 Selects article 2 from summary
4:56 Selects sentence from article 2
Comment: Moving on to article two. Nicely brings you, actually, to the main results sentence which is a great summary.
11:32 Selects Article 5 from summary
11:46 Selects sentence from article 5
11:52 Selects sentence from article 5
Comment: All seem to be relevant and in support of the fi ndings.

Figure 6. *Excerpts from the same subject looking at a personalized, generated summary.*

DISCUSSION

We found that personalized summaries allowed users to complete the task more successfully and with greater satisfaction than under other conditions. Although most users preferred to read the summary first, they liked the feature of the summaries that link to specific sentences inside the articles.

Contrary to our expectations, we found that generic

Question	Personalized	Generic	Search
Q1. Did you feel like you had enough time to identify all the relevant findings?	2.7	3.2	3.3
Q2. At the end of the task, do you think you have a reasonable answer?	3.2	3.3	3.5
Q3. Did you feel that the interface was helpful in supporting your task?	4.5	4	4.5
Q4. If you had the opportunity, would you use this interface again?	4	3.8	4.2
Q5. How did you access the articles?	3.2	1.5	NA
Q6. Did you read the summary?	3.5	2	NA
Q7. Did you feel that the summary saved you time?	4	2.8	NA
Q8. Did you feel that the summary content was relevant to the given question, patient under care and studies?	4	2.8	NA

Figure 4. *Questionnaire given to the subjects after each completed scenario and the average responses per condition.*

summarization did not improve access to information. Subjects did not like the generic summaries because they were lengthy and often incoherent; since they were generated using extraction of sentences (without modification), they contained dangling references. Much of the information that they contained was not relevant to the patient and as a result, subjects gave up reading the summary and focused on the list of articles at the bottom only.

On the other hand, users are used to seeing search results containing only article title and information about where published, so they did not complain about the interface when only search results were shown. However, when the articles were long, they did not like having to scroll through the whole article to get to the relevant information.

There were some issues that we will investigate in more detail in follow-up studies. There was low agreement among the subjects' selection. Our analysis shows that this is due in part to repetitions across articles. When information was repeated, some subjects selected all instances and others only one, despite our guidelines asking them to select all repetitions.

Finally, subjects could not get over the fact that input articles were pre-selected for them. We chose to do this in order to isolate the effect of summarization on the task. We did not want to evaluate subjects' search strategies. Although we stressed this point in the guidelines, this remained a source of confusion.

RELATED WORK

While there are many summarization systems in different domains, there is little work done in personalized summarization, and there is also little work on summarization of a collection of technical articles. Thus, it is not surprising that most evaluations have focused on generic summarization systems.

How best to evaluate a summarization system is still an open research question. One approach has been to measure how well generated summaries match summaries written by humans [4, 5, 6]. Several task-based evaluations have been conducted for summarization. Evaluations of single document summarization systems use human relevance judgments based on the summary [7], while others have looked at the effect of summarization on comprehension of texts [8].

CONCLUSION

Our user study shows that personalized summaries allow physicians to find information related to patient care in the medical literature more efficiently than do either search engines or generic summarization. Given the task of finding all results pertinent to a patient under care, subjects were able to find more relevant, and more accurate, results with the personalized summary than under other conditions. Answers to a questionnaire, as well as an analysis of video tapes showing subjects using the system, reveal that subjects were also more satisfied with the personalized summaries. Finally, the video analysis shows that the personalized summaries allowed subjects to access relevant findings efficiently either directly in the summary or through links from the summary directly to the point in the article where findings were presented.

References

1. W. Hersh, D. Hickam. How well do physicians use electronic information retrieval systems? A Framework for investigation and systematic review. JAMA 280(15), 1347-1352. 1998.
2. N. Elhadad, M.Y. Kan, J. Klavans, and K. McKeown. Customization in a unified framework for summarizing medical literature. To appear in *Journal of Artificial Intelligence in Medicine*, 2005.
3. D. Kaufman, V. Patel, C. Hilliman, P. Morin, J. Pevzner, Weinstock, R. Goland, S. Shea, and J. Starren. Usability in the real world: Assessing medical information technologies in patients' homes. Journal of Biomedical Informatics, 36, 45-60. 2003.
4. Proceedings of the second, third, and fourth document understanding conference. 2002, 2003, 2004.
5. C.Y. Lin and E. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In Proceedings of HLT-NAACL. 2003.
6. A. Nenkova and B. Passonneau. Evaluating content selection in summarization: The pyramid method. In Proceedings of HLT-NAACL. 2004.
7. A. Kushniruk, M.Y. Kan, K. McKeown et al. Usability evaluation of an experimental text summarization system and three search engines: Implications for the reengineering of health care interfaces. In Proceedings of AMIA. 2002.
8. A. Morris, G. Kasper, and D. Adams. The effects and limitations of automated text condensing on reading comprehension performance. In Information Systems Research, 3(1), 17-35. 1992.