Text Summarizer in Use: Lessons Learned from Real World Deployment and Evaluation

Mary Ellen Okurowski
Harold Wilson
Joaquin Urbina
Department of Defense
9800 Savage Rd.
Fort Meade, MD. 20755

Tony Taylor
SRA Corp.
4939 Elkridge Landing
Suite #195
Linthicum, MD. 21090

Ruth Colvin Clark
Clark Training & Consulting
17801 CR 23
Dolores, Colorado 81323

Frank Krapcho
Kathpol Technologies Inc.
6835 Deerpath Suite #102
Elkridge, MD. 21705

1.0 Introduction

Much of the historical and current summarization literature has been technology-centered with the questions posed and answered having implications for technology development. Though commercial summarization products have appeared in the market place and developers continue to explore new summarization areas, few papers have been user-centered, examining summarization technology *in-use*. In this paper, we show how applied work and the knowledge gleaned about technology in-use can temper theoretical considerations and motivate as well as direct development likely to result in higher return on investment.

2.0 Background

The importance of understanding the function a summary serves for users is widely acknowledged, and seminal works defining summary types by functions (Paice, 1990; Sparck-Jones, 1993) are frequently cited by developers. Task orientation defines extrinsic technology assessments, and the research literature on how to assess performance for machine generated summaries in an experimental task scenario has grown (Brandow *et al.,* 1994; Morris *et al.,* 1999; Jing *et al.,* 1998; Merlino and Maybury, 1999; Wasson, 1998; Tombros *et al.,* 1998; Firmin and Chrzanowski, 1999; and Mani *et al.,* 1999). An increasing number of research papers on summarization systems now also describe some type of extrinsic evaluative task (e.g. Salton *et al.,* 1999; Strzalkowski *et al.,* 1998). A number of factors (i.e. characteristics of summaries, documents, users, and tasks) have surfaced which have implications for technology use. More research assessing technology (or any aspect of it) *in-use* on a user's own data even in a development mode along the lines of McKeown *et al.* (1998) is needed. While experimentation designs involving subjects performing short term controlled tasks may yield results of statistical significance, generalizability to the user community is limited.

In addition, the level of user support text summarization systems should provide also continues to be speculative. More interest lies in new areas of inquiry like visualization and browsing techniques (e.g., Boguraev *et al.,* 1998), multi-document summarization (e.g., McKeown and Radev, 1995), multi-media summarization (e.g., Merlino and Maybury, 1999), summarization of documents with graphics (e.g., Futrelle, 1998) and multi-lingual summarization (e.g., Cowie, 1998). But systematic user studies on interface support, applicability of proposed summarization features, or on the real-world use of demonstration and prototype systems or even commercial systems have not materialized.

3.0 Overview

This paper presents a user study of a summarization system and provides insights on a number of technical issues relevant to the summarization R&D community that arise in the context of use, concerning technology performance and user support. We describe initial stages in the insertion of the SRA summarizer in which (1) a large scale beta test was conducted, and (2) analysis of tool usage data, user surveys and observations, and user requirements is leading to system enhancements and more effective summarization technology insertion. In our user study, we begin with a brief description of the task and technology (3.1). We then describe the beta test methodology (3.2) and analysis of tool usage data (3.3). We focus on what we learned in our user-centered approach about how technology performance in a task and user support affect user acceptance (3.4) and what significant technology-related modifications resulted and what studies are in progress to measure tool efficacy, summarization effectiveness, and the impact of training on tool use (3.5). Though work to enhance the text summarization system is underway, we focus in this paper on user-centered issues. Our work is predicated on the belief that there is no substitute for user generated data to guide tool enhancement.

3.1 Task and Technology

The task is indicative. Our users rely on machine generated summaries (single document, either generic or query-based, with user adjustment of compression rates) to judge relevance of full documents to their information need. As an information analyst, our typical user routinely scans summaries to stay current with fields of interest and enhance domain knowledge. This scanning task is one of many jobs an analyst performs to support report writing for customers in other Government agencies. Our goal is to generate summaries that accelerate eliminating or selecting documents without misleading or causing a user to access the original text unnecessarily.

The system in this user study is a version of the SRA sentence extraction system described in Aone *et al.* (1997, 1998, 1999). Users retrieve documents from a database of multiple text collections of reports and press. Documents are generally written in a journalistic style and average 2,000 characters in length. The number of documents in a batch may vary from a few to hundreds. Batches of retrieved texts may be routinely routed to our summary server or uploaded by the user. The system is web-based and provides the capability to tailor summary output by creating multiple summary set-ups. User options include: number of sentences viewed, summary type applied and sorting, other information viewed (e.g. title, date), and high frequency document terms and named entities viewed. Users can save, print or view full text originals with summaries appended. Viewed originals highlight extracted sentences.

All system use is voluntary. Our users are customers and, if dissatisfied, may elect to scan data without our technology.

3.2 Beta Test Methodology

In the fall of 1998, 90+ users were recruited primarily through an IR system news group and provided access to the SRA system summarizer to replace their full text review process of scanning concatenated files. Procedural (how-to) training was optional, but approximately 70 users opted to receive a one-on-one hands-on demonstration (about forty-five minutes in length) on texts that the new user had retrieved. The beta testing took place over a six month period. With no stipulation on the length of participation, many users simply tried out the system a limited number of times. Initial feedback gave us a clear picture of the likelihood of continued use. Our relatively low retention rate highlighted the fact that the experimental conditions in previous summary experiments may be misleading and masked factors that do not surface until users use a system in a daily work in a real-world setting.

3.3 Analysis of Tool Usage Data

Usage data were collected for all system users and analyzed through web logs. These logs were a record of what users did on their actual work data. For each user, our logs provided a rich source of information: number of summary batches, number of documents in each, whether documents were viewed, and set up features--summary type, summary lines viewed, number of indicator (high frequency signature terms) lines viewed, number of entity (persons, places, organizations) lines viewed, query terms). Table 1 below illustrates the type of representative data collected, questions of interest, and findings.

Table 1: Questions of Interest, Tool Usage Data, Findings

Questions	Data	Finding
Were documents summarized?	number of summary events	Users routinely accessed our system to read machine generated summaries.
Did users actually tailor the system?	number of current set-ups	Most users did not appear to fully exploit the flexibility of the system. The beta test population had a median of only two set-up types active.
Did the users select generic or query-based summaries?	type of summary	Usage data indicated that about half the population selected generic and the other half query-based summaries. (Note: The default set-up was the generic summarization.)
Is there a difference among summary types for the number of sentences viewed?	number of sentences viewed by summary types (generic, query-based, lead)	The hypothesis of equal median number of sentences available for viewing sentences was tested. The number of sentences viewed with generic summary type (3) is significantly different from either query-based (5) or lead (6).
Do users choose to use indicators and entities when tailoring browsing capability?	indicator/entity preferences for non-default set-ups (on or off)	Users tended to retain indicator and entity preferences when tailoring capabilities. (But users generally modified a default set-up in which both preferences have a line viewed.)

Table 1: Questions of Interest, Tool Usage Data, Findings

Questions	Data	Finding
Does training make a difference on system use or user profile type? Users were categorized (advanced, intermediate, novice) on the basis of usage features with Hartigan's K-Means clustering algorithm.	training and tool use data	A chi-squared test for independence between training and use reflected a significant relationship (p value close to 0) i.e., training did impact the user's decision to use the system. However, training did not make a difference across the three user profile types. A Fisher Exact test on a 3x2 contingency table revealed that the relative numbers of trained and untrained users at the three user profile types were the same (p-value= 0.1916) i.e., training and type are independent.

As we began to analyze the data, we realized that we had **only** a record of use, but were not sure of what motivated the use patterns. Therefore, the team supplemented tool usage data with an on-line survey and one-on-one observations to help us understand and analyze the user behavior. These additional data points motivated much of our work described in 3.5. Throughout the six month cycle we also collected and categorized user requirements.

3.4 Insights on Text Summarization

3.4.1 Technology Performance

Insight 1: *For user acceptance, technology performance must go beyond a good summary. It requires an understanding of the users' work practices.*

We learned that many factors in the task environment affect technology performance and user acceptance. Underpinning much work in summarization is the view that summaries are time savers. Mani *et al.* (1999) report that summaries at a low compression rate reduced decision making time by 40% (categorization) and 50% (adhoc) with relevance asessments almost as accurate as the full text. Although evaluators acknowledge the role of data presentation (e.g., Firmin and Chrzanowski, 1999; Merlino and Maybury, 1999), most studies use summary system output as the metric for evaluation. The question routinely posed seems to be "Do summaries save the user time without loss in accuracy?" However, we confirmed observations on the integration of summarization and retrieval technologies of McKeown *et al.* (1998) and learned that users are not likely to consider using summaries as a time saver unless the summaries are efficiently accessed. For our users a tight coupling of retrieval and summarization is pre-requisite. Batches automatically routed to the summary server available for user review were preferred over those requiring the user to upload files for summarization. Users pointed out that the uploading took more time then they were willing to spend.

User needs and their work practices often constrain how technology is applied. For example, McKeown *et al.* (1998) focused on the needs of physicians who want to examine only data for patients with similar characteristics to their own patients, and Wasson (1998) focused on the needs of news information customers who want to retrieve documents likely to be on-topic. We too discovered that the user needs affect their interest in summarization technology, but from a more general perspective. Text REtrieval Conferences (e.g., Harman, 1996) have baselined system performance in terms of two types of tasks--routing or ad-hoc. In our environment the ad-hoc users were less likely to want a summary. They simply wanted an answer to a question and did not want to review summaries. If too many documents were retrieved, they would simply craft a more effective query.

Measuring the efficiency gains with a real population was quite problematic for technology in-use. We faced a number of challenges. Note that in experimental conditions, subjects perform on full and reduced versions. One challenge was to baseline non-intrusively the current (non-summary) full text review process. A second was to measure both accuracy and efficiency gains for users performing on the job. These challenges were further exacerbated by the fact that users in an indicative task primarily use a summary to eliminate most documents. They have developed effective skimming and scanning techniques and are already quite efficient at this task.

In short, our experience showed that technologists deploying single document summarization capability are likely be constrained by the following factors:

- the ease of technology use
- the type of user information need
- how effective the user performs the task without the technology.

3.4.2 User Support

Insight 2: *Users require more than just a good summary. They require the right level of technology support.*

Although the bulk of the research work still continues to focus on summarization algorithms, we now appreciate the importance of user support to text summarization use. The SRA software was quite robust and fast. The task of judging relevance with a summary (even a machine generated one) instead of the full text version does not require a user to acquire a fundamentally different work practice. Yet our system was not apparently sufficiently supporting tool navigation. One of the reasons was that our on-line help was not developed from a user perspective and was rarely accessed. Another was that browse and view features did not maximize performance. For example, the interface employed a scroll bar for viewing summaries rather than more effective Next Or Previous buttons. Users frequently asked the same questions, but we were answering them individually. Terminology clear to the technologists was not understood by users. We also noticed that though there were requirements for improvement of summarization quality, many requirements were associated with these user support issues.

One of the more unexpected findings was the under-utilization of tailoring features. The system offered the user many ways to tailor summaries to their individual needs, yet most users simply relied on default set-ups. Observations revealed little understanding of the configurable features and how these features corresponded to user needs to say nothing of how the algorithm worked. Some users did not understand the difference between the two summary types or sorting effects with query-based summary selection. Non-traditional summary types--indicators and named entities--did not appear to help render a relevance judgment. We came to understand that just because technologists sees the value to these features does not mean that a user will or that the features, in fact, have utility.

3.5 Technology-related Modifications

3.5.1 User-centered Changes to Technology Work Practices

On technology performance, we learned that

- seamless integration with an IR system was preferred
- users with static queries were more likely customers for a summary service
- gains in efficiency are hard to measure for a task already efficiently performed in a real-world situations.

In response, we have established a summary service in which retrieval results are directly routed to our summary server and await the user. We plan to integrate the summarization tool into the IR system. (Uploading batches and then submission to the server is still an option.) We also abandoned the naive idea that data overload equates to summarization requirements and realized that the technology does not apply to all users. We have more effectively selected users by profiling characteristics of active users (e.g. daily document viewing work practice, document volume, static query use, etc.) and have prioritized deployment to that population which could most benefit from it.

In order to demonstrate tool summarization efficiency, we needed to baseline full-text review. We considered, but rejected a number of options--user self-report and timing, observations, and even the creation of a viewing tool to monitor and document full text review. Instead, we baselined full text scanning through information retrieval logs for a subgroup of users by tracking per document viewing time for a month period. These users submit the same queries daily and view their documents through the IR system browser. For the heaviest system users, 75% of the documents were viewed in under 20 seconds per document, but note that users vary widely with a tendency to spend a much longer browse time on a relatively small number of documents. We then identified a subgroup of these users and attempted to deploy the summarizer to this baseline group to compare scanning time required over a similar time frame. We are currently analyzing these data.

System in a work environment is considered a good indicator of tool utility, but we wanted some gauge of summary quality and also anticipated user concerns about an emerging technology like automatic text summarization. We compromised and selected a method to measure the effectiveness of our summaries that serves a dual purpose--our users gain confidence in the utility of the summaries and we can collect and measure the effectiveness of the generic summaries for some of our users on their data.

We initially piloted and now have incorporated a data collection procedure into our software. In our on-line training, we guide users to explore tool capabilities through a series of experiments or tasks. In the first of these tasks, a user is asked to submit a batch for summarization, then for each of five to seven user-selected summaries to record answers to the question:

"Is this document likely to be relevant to me?"(based on the summary)
____yes ____no

Then, the user was directed to open the original documents for each of the summaries and record answers to the question:

"Is the document relevant to me? "
(after reading the original text)
___yes ___no

In a prototype collection effort, we asked users to review the first ten documents, but in follow-on interviews the users recommended review of fewer documents. We understand the limits this places on interpreting our data. Also, the on-line training is optional so we are not able to collect these data for all our users uniformly.

Most of the users tested exhibited both high recall and precision, with six users judging relevance correctly for all documents (in Table 2 below). The False Negative error was high for only one user, while the majority of the users exhibited no False Negative errors, a worse error to commit than wasting time viewing irrelevant data, False Positive. Across all the users, 79% of all relevant documents and 81% of the irrelevant documents were accurately categorized by examination of the summary.

Table 2: Relevance Classes by User

User	True Positive	False Positive	True Negative	False Negative
1	5	0	0	0
2	5	0	0	0
3	4	0	0	1
4	1	4	0	0
5	5	0	0	1
1	4	0	0	2
7	7	0	0	0
8	4	0	3	0
9	5	0	0	2
10	0	0	7	0
11	2	0	3	0
12	1	0	2	2
13	0	1	6	0
14	1	0	1	4

3.5.2 User-centered Changes in User Support

On user support , we learned that

- our system did not effectively support user tool navigation
- our users did not fully exploit system tailorable features

In response, we addressed user support needs from three different angles, each of which we discuss below: incorporation of Electronic Performance Support Systems, design and implementation of procedural on-line training and guided discovery training, and user analysis of summary quality.

Electronic Performance Support Systems (EPSS) is a widely acknowledged strategy for on the job performance support. Defined as "an optimized body of co-ordinated on-line methods and resources that enable and maintain a person's or an organization's performance," EPSS interventions range from simple help systems to intelligent wizard-types of support. (Villachica and Stone, 1999; Gery 1991). We elected to incorporate EPSS rather than classroom instruction. Based on an analysis of tool usage data, user requirements, and user observations, experts in interface design and technology performance support prototyped an EPSS enhanced interface. Active system users reviewed these changes before implementation. The on-line perfomance support available at all times includes system feature procedures, a term glossary, FAQ, and a new interface design.

With incorporation of the EPSS, we also addressed the under-utilization of the configurable features. Although simple technologies with few options such as traditional telephones do not require conceptual system understanding for effective use, more complex systems with multiple options are often underutilized when supported with procedural training alone. We decided to incorporate both procedural training in a "Getting Started" tutorial and conceptual training in "The Lab." In "Getting Started", users learn basic system actions (e.g., creating set-ups, submitting batches for summarization, viewing summaries). "The Lab", on the other hand, supports guided discovery training in which users explore the system through a series of experiments in which they use their own data against various tool options and record their observations. Given our own experience with under-utilization and research reporting difficulties with unguided exploratory learning (Hsu *et* *al.,* 1993; Tuovinen and Sweller, 1999), we built on the work of de Mul and Van Oostendorf (1996) and Van Oostendorf and de Mul (1999) and their finding that task-oriented exploratory support leads to more effective learning of computer systems. We created a series of experiments that the user conducts to discover how the summarization technology can best meet their needs. For example, users are directed to change summary length and to determine for themselves how the variation affects their ability to judge relevance using their data.

In February, we conducted a study of two groups, one with the EPSS and "Getting Starting" Tutorial and a second with the same level of support and additionally "The Lab". Earlier work by Kieras and Bovair (1984) compared straight procedural training with conceptual training and showed that the conceptually trained users made more efficient use of system features. The goal of our study was to determine just what level of training support the summarization technology requires for effective use. Through surveys, we planned to collect attitudes toward the tool and training and through web logs, tool usage data and option trials. We also planned to assess the users' understanding of the features and benefits of the tool. We are currently analyzing these data.

In addition to the EPSS and the on-line training, we developed a method for taking into account user assessment of our summary quality in a systematic way. User feedback on summarization quality during the beta test was far too general and uneven. We recruited two users to join our technology team and become informed rather than the typical naive users. They designed an analysis tool through which they database problematic machine generated summaries and assign them to error-type categories. Though we expected users to address issues like summary coherence, they have identified categories like the following:

- sentence identification errors
- formatting errors
- sentence extraction due to the "rare" word phenomena
- sentence extraction in "long" documents
- failure to identify abstracts when available

We expect that this approach can complement a technology-driven one by helping us prioritize changes we need based on methodical data collection and analysis.

4.0 Summary

Our experience with text summarization technology *in-use* has been quite sobering. In this paper, we have shown how beta testing an emerging technology has helped us to understand that for technology to enhance job performance many factors besides the algorithm need to be addressed.

5.0 References

Aone, C., Gorlinsky, J. and Okurowski, M.E. 1997. Trainable, scalable summarization using robust NLP. In *Intelligent Scalable Text Summarization.* Madrid, Spain: Association of Computational Linguistics, pages 66-73.

Aone, C., Gorlinsky, J. and Okurowski, M.E. 1998. Trainable scalable summarization using robust NLP and machine learning. In *Coling-ACL 98.* Montreal, Quebec, Canada, pages 62-66.

Aone, C., Gorlinsky, J., Larsen, B. and Okurowski, M.E. 1999. A trainable summarizer with knowledge acquired from robust NLP techniques. In Mani, I. and Maybury, M. (eds.), *Advances in Automatic Text Summarization.* pages 71-80, Cambridge, Massachusetts: MIT Press.

Boguraev, B., Kennedy, C., Bellamey, R., Brawer, S., Wong, Y.Y. and Swartz, J. 1998. Dynamic presentation of document content for rapid on-line skimming. *Intelligent Text Summarization.* (Papers from the 1998 AAAI Spring Symposium Technical Report SS-98-06), pages 109-118.

Brandow, R., Mitze, K. and Rau, L. 1994. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management,* 31(5):675-685.

Cowie, J., Mahesh, K., Nirenburg, S. and Zajac, R., 1998. MINDS--Multi-lingual INteractive document summarization. *Intelligent Text Summarization.* (Papers from the 1998 AAAI Spring Symposium Technical Report SS-98-06), pages 122-123.

de Mul, S. and van Oostendorp, H. 1996. Learning user interfaces by exploration. *Acta Psychologica,* 91:325-344.

Firmin, T. and Chrzanowski, M. 1999. An evaluation of automatic text summarization. In Mani, I. and Maybury, M.(eds.), *Advances in Automatic Text Summarization.* pages 325-336, Cambridge, Massachusetts: MIT Press.

Futrelle, R. 1998. Summarization of documents that include graphics. *Intelligent Text Summarization.* (Papers from the 1998 AAAI Spring Symposium Technical Report SS-98-06), pages 92-101.

Gery, G. 1991. *Electronic performance support systems: How and why to remake the workplace through the stratgic application of technology.* Tolland, MA: Gery Performance Press.

Harman, D.K. 1996. *The Fourth Text REtrieval Conference (TREC-4).* National Institute of Standards and Technology Special Publication, pages 500-236.

Hsu, J-F., Chapelle, C. and Thompson, A., 1993. Exploratory learning environments: What are they and do students explore? *Journal Educational Computing Research,* 9(1): 1-15.

Jing, H., McKeown, K., Barzilay, R. and Elhadad, M. 1998. Summarization evaluation methods: Experiments and methods. *Intelligent Text Summarization.* (Papers from the 1998 AAAI Spring Symposium Technical Report SS-98-06), pages 51-59.

Kieras, D.E. and Bovair, S. 1984. The role of a mental model in learning to operate a device. *Cognitive Science,* (8), 1-17.

Mani, I., House, D., Klein, G., Hirschman, L., Firmin, T. and Sundheim, B. 1999. The TIPSTER SUMMAC Text Summarization Evaluation. In *Proceedings of EACL99 Ninth Conference of the European Chapter of the Association for Computational Linguistics.* pages 77-83.

McKeown, K. and Radev, D. 1995. Generating summaries of multiple news articles. In *Proceedings of the 18th Annual International SIGIR Conference on Research and Development in Information Retrieval.* pages 74-78.

McKeown, K. Jordan, D. and Hatzivassiloglou, V. 1998. Generating patient specific summaries on online literature. *Intelligent Text Summarization.* (Papers from the 1998 AAAI Spring Symposium Technical Report SS-98-06), pages 34-43.

Merlino, A. and Maybury, M. 1999. An empirical study of the optimal presentation of multi-media summaries of broadcast news. In Mani, I. and Maybury, M. (eds.), *Advances in Automatic Text*

Summarization. pages 391-401, Cambridge, Massachusetts: MIT Press.

Morris, A., Kasper, G., and Adams, D. 1999. The effects and limitations of automated text condensing on reading comprehension performance. In Mani, I. and Maybury, M. (eds.), *Advances in Automatic Text Summarization.* pages 305-323, Cambridge, Massachusetts: MIT Press.

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

Sparck-Jones, K. 1993. What might be in a summary? In *Information Retrieval 93: Von der Modellierung zur Anwendung,* pages 9-26.

Salton, G., Singhal, A., Mitra, M. and Buckely, C. 1999. Automatic text structuring and summarization. In Mani, I. and Maybury, M. (eds.), *Advances in Automatic Text Summarization.* pages 342-355, Cambridge, Massachusetts: MIT Press.

Strzalkowski, T., Wang, J. and Wise, B., 1998. A robust practical text summarization. *Intelligent Text Summarization.* (Papers from the 1998 AAAI Spring Symposium Technical Report SS-98-06), pages 26-33.

Tombros, A., and Sanderson, M. 1998. Advantages of query-based summaries in information retrieval. In *Proceedings of the 21st ACM SIGIR Conference (SIGIR98).* pages 2-10.

Tuovinen, J. and Sweller, J. 1999. A comparison of cognitive load associated with discovery learning and worked examples. *Journal of Educational Psychology,* 9(2):334-341.

Van Oostendorp, H. and de Mul, S. 1999. Learning by exploring: Thinking aloud while exploring an information system. *Instruction Science,* 27:269-284.

Villachica, S.W. and Stone, D. 1999. Performance support systems. In Stolovitch, H.D. and Keeps, K.J., (eds.), *Handbook of Human Performance Technology.* San Francisco: Jossey-Bass Pfeiffer.

Wasson, M. 1998. Using leading text for news summaries: Evaluation results and implications for commercial summarization. In *Coling-ACL 98.* Montreal, Quebec, Canada. pages 1364-1368.