Using Interview Data to Identify Evaluation Criteria for Interactive, Analytical Question-Answering Systems

Diane Kelly
School of Information and Library Science, University of North Carolina, 100 Manning Hall, CB #3360, Chapel Hill, NC 27599–3360. E-mail: dianek@email.unc.edu

Nina Wacholder, Robert Rittman, Ying Sun, and Paul Kantor
School of Communication, Information and Library Studies, Rutgers University, 4 Huntington Street, New Brunswick, NJ 08901

Sharon Small and Tomek Strzalkowski
Institute for Informatics, Logics and Security Studies, University of Albany, SUNY, Room 262, Social Science Building, Albany, NY 12222

The purpose of this work is to identify potential evaluation criteria for interactive, analytical question-answering (QA) systems by analyzing evaluative comments made by users of such a system. Qualitative data collected from intelligence analysts during interviews and focus groups were analyzed to identify common themes related to performance, use, and usability. These data were collected as part of an intensive, three-day evaluation workshop of the High-Quality Interactive Question Answering (HITIQA) system. Inductive coding and memoing were used to identify and categorize these data. Results suggest potential evaluation criteria for interactive, analytical QA systems, which can be used to guide the development and design of future systems and evaluations. This work contributes to studies of QA systems, information seeking and use behaviors, and interactive searching.

Introduction

For many years, the Advanced Question Answering for Intelligence (AQUAINT) program of the Advanced Research and Development Activity (ARDA) has encouraged advances in automated question-answering (QA) and the development of innovative technological solutions to problems that face intelligence analysts in their daily work. Ultimately, techniques developed in this program are expected to spread to tools available to the general public (Maybury, 2002). Thus, QA systems under development as part of the AQUAINT program need to be evaluated by target users in realistic settings, with realistic tasks. This requires holistic evaluation methods and metrics that should include (a) system-oriented measures, such as performance; (b) process-oriented measures, such as effort; (c) product-oriented measures, such as report quality; and (d) user-oriented measures, such as comprehension, satisfaction, and preference.

There are, however, few studies to guide the evaluation of analytic QA systems, and there are even fewer studies to guide the evaluation of interactive, analytical question-answering where real users are involved. Most evaluations of QA systems have been conducted as part of the QA Track at TREC. These evaluations are system-oriented rather than user-oriented, and the focus has been on the evaluation of techniques for answer extraction, rather than interaction and use. The TREC focus has been on factual or definitional questions rather than complex, analytical questions (c.f., Voorhees, 2003). Some emerging work in the QA community is starting to investigate how users interact with factual QA systems (Lin et al., 2003) and seek information for more complex QA tasks (Diekema et al., 2004).

A number of QA tasks beyond factual are described in Maybury (2004), including temporal, spatial, and opinion; in this article, we focus on analytical QA. Answers to analytical questions differ from answers to factual questions in that they are multidimensional and typically require bringing together information from multiple sources. The question "How is the al-Qaeda organization funded?" is an example of such a question. Evaluation of support for factual QA usually focuses on whether a system finds the correct answer to a question, rather than investigating the larger contexts in which QA is used and the interactions that occur between an information-seeker, the system, and the answer. While factual QA is likely to be an important part of analytical QA, analytical QA includes other activities, such as

Received October 6, 2005; revised July 18, 2006; accepted July 18, 2006

 ● Published online 21 March 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20575

comparison and synthesis, and demands much richer interactions among the system, the information, and the user. This suggests that perhaps different evaluation measures are appropriate for systems supporting analytical QA than from those supporting factual QA. Diekema et al. (2004) note that "having a relatively successful QA system tailored to the TREC question-answering task does not necessarily ensure success in question-answering applications outside of TREC" (p. 141). The test and evaluation environment of the TREC QA Track has limited generalizability to real-world QA environments, especially for environments where people interact with and use the QA system. Interactive systems are more demanding of user's time, attention, and effort than are simple batch systems. Therefore, it is not enough for such systems to be efficient and effective. Because the process itself is part of the product, user-centered evaluations are critical.

To situate the present work, we note that a system may support either information retrieval or question answering, and may be used in batch mode or interactive mode. In batch mode, a system accepts as input a list of queries and then generates sets of retrieved documents (or answers) for each query. In traditional TREC experiments, including those conducted as part of the TREC QA Track, systems operate in batch mode. In interaction mode, real users enter queries, receive lists of results, and then examine the full-text of those results to determine if the contents are relevant to their information needs. This might include users revising their search queries over many cycles. In a typical interactive QA system, users generate subsidiary questions to investigate distinct components of their information needs and often interact with the system through dialogue. For instance, the HITIQA system may ask clarification or expansion questions before it presents answers (Small et al., 2004). Answers to subsidiary questions may be small units such as passages, sentences, or words or, in some cases, a synthesized unit from multiple documents. Users examine these units to determine whether the system has "answered" the subsidiary question. This represents only one such way a system might be interactive, but the primary distinction between batch-mode and interactive-mode is that a person is in-the-loop, and directs at least part of the process.

While much has been done to establish guidelines and develop methods for evaluating interactive information retrieval systems, it is unclear which, if any, of these methods and objectives generalize to interactive QA system evaluation. In the context of information retrieval, the TREC Interactive Track led research and development of evaluation methods for interactive information retrieval (IR) systems for a number of years (Hersh & Over, 2001a). Typical measures included those related to the system, such as performance and usability, and those elicited from the user such as satisfaction and system preferences. It was also common in the Interactive Track to collect other information about users, such as topic and search familiarity, and to investigate these variables in relation to other system- and user-oriented measures. In TREC-9, the Interactive Track employed factual QA as the assessment task (Hersh & Over, 2001b). Most systems in this Track were document retrieval systems, rather than full-scale QA systems capable of returning specific answers to questions. Generally, systems were modified to return passages or query-biased summaries, or used highlighting to indicate answers within full-text documents. Users did not enter natural language queries in question format; instead, they entered more traditional keyword queries, even though the tasks they were trying to solve were factual questions. While evaluation from the TREC-9 Interactive Track yielded some preliminary insight into how users interact with a system when engaged in QA tasks and how these systems might be evaluated, results from these studies are limited because systems studied were not actual QA systems and search tasks were only factual in nature.

We began the task of exploring evaluation methods and metrics for analytical QA systems in Wacholder (in press), where we reported on the quantitative aspects of a user-centered evaluation of HITIQA. We conducted two three-day evaluation workshops of HITIQA with intelligence analysts serving as study participants. We employed a number of data collection techniques including observation, logging, questionnaires, and interviews. The work reported in this article describes results of an analysis of qualitative data collected through individual and group interviews. The purpose of the current work is to identify potential evaluation criteria for interactive, analytical question-answering systems through the analysis of evaluative comments made by users of the HITIQA system. Our goal in doing so is to suggest possible measures for future evaluations of interactive, analytical QA systems.

Related Work

Automatic QA system evaluation has primarily taken place within the laboratory, with factual or definitional questions and without the added complexity of users. Since 1999, TREC has sponsored a QA Track (Voorhees, 2001), which is gradually becoming one of the longest running TREC tracks. While the TREC model is particularly useful for QA evaluation, especially in early development periods, it is less useful when systems reach a point where they are ready to be used in real-world environments with users. For instance, in the TREC model relevance judgments are fixed and usually binary. Because the TREC QA Track has been concerned primarily with addressing factual or definitional questions rather than analytical questions, this design is not problematic. However, answers to complex analytical questions are multi-dimensional and typically include information from multiple sources.

Users working with QA systems are likely to use systems over an extended period of time, during which time their information needs change as they encounter new information. This type of interaction is more like Bates' (1989) depiction of "berry-picking," where users engage in picking bits of information from various sources they encounter as they seek

DOI: 10.1002/asi

information. As a result of these interactions, users' understandings and thus their information needs and relevance judgments change. When engaged in analytical question-answering, relevance is likely to take on characteristics more like those described in the interactive IR literature than those assumed in TREC-style QA evaluations.

The work of the TREC-9 Interactive Track (Hersh & Over, 2001b) provides some insights into interactive, factual QA, and evaluation. Although this Track was not specifically designed to assess QA systems or to develop evaluation measures for such systems, during TREC-9 the Track employed factual question-answering as the search task. Users completed eight total search tasks. There were two types of tasks. The first type required users to respond with an answer that had from one to four parts (e.g., "What are the names of three U.S. national parks where one can find redwoods?"). The second type required searchers to decide which of two given answers was correct (e.g., "Is Denmark larger or smaller in population than Norway?"). For each task, users were required to answer questions and save supporting documentation for their answers. Evaluation consisted of traditional interactive IR measures such as satisfaction, preference, and effort; effort included measures such as time and number of search iterations. Evaluation of performance was measured by the completeness of the answer (i.e., Does the response contain all, some or none of the items asked for by the question?) and the appropriateness of the cited documentation to support the answers (i.e., Do the documents cited fully support all, some or none of the correct items in the response?). As mentioned earlier, most systems in this Track were document retrieval systems, rather than full-scale QA systems with capabilities of returning answers. Because of this, most systems were modified to return passages or query-biased summaries or used highlighting to indicate answers within full-text documents to support the QA task. Thus, this work is most relevant to the current work with respect to user-system interaction during factual QA and evaluation.

Two groups in particular provided some interesting insights into interactive QA and evaluation at TREC-9 (Belkin et al., 2001; D'Souza et al., 2001). Belkin et al. compared two interfaces designed to support factual QA tasks. The baseline interface presented search results as a ranked list of document titles, with the text of one document displayed in a scrollable window. The experimental interface used a newspaper metaphor consistent with that used by Golovchinsky and Chignell (1997) and presented search results as two rows of three document windows. Each window was scrollable and documents were displayed beginning with the system-computed best passage. Belkin et al. (2001) hypothesized that the baseline interface would best support the first type of task (i.e., list instances) and the experimental interface would best support the second type of task (i.e., comparison). Interfaces were compared using three criteria: performance, effort, and preference. In general, there were no significant differences between interfaces for each measure, except that users engaged in significantly fewer iterations with the experimental interface than with the baseline. Results from most Track participants demonstrated that, overall, users had a difficult time answering many of the questions in the allotted five minutes; Belkin et al.'s (2001) lack of significant results may be related to this. It might also be the case that the traditional interactive IR evaluation measures were not tailored enough to the QA task to detect meaningful differences between systems. With one exception, most participants of the Interactive Track found no significant results.

The exception was D'Souza et al. (2001) who found significant differences in performance between two interfaces. These interfaces were designed to mimic Web search interfaces. Both interfaces provided users with surrogates of the top 100 retrieved documents in five consecutive pages, with each page containing twenty documents surrogates. The baseline interface presented users with traditional Web document surrogates: document titles and short summaries. The experimental interface presented users with alternative document surrogates containing the title and three "best" sentences from the document. Although inferential statistical results for the experiment are not reported in D'Souza et al. (2001), they are later presented in Wu, Fuller and Wilkinson (2001), along with results from an additional experiment with the same interfaces and tasks. Results demonstrated that performance with the experimental interface was significantly better than that of the baseline interface, and that users had fewer iterations, viewed fewer pages of results, and read fewer documents with the experimental interface than with the baseline interface. It was further found that users preferred the experimental interface over the baseline and found it more useful and easier to use. It is unclear if these results are related specifically to the QA task, or if they say something more general about the types of document surrogates users prefer regardless of search task. However, the evidence suggests that the experimental interfaces better supported users in this QA task. Another group (Beaulieu, Fowkes, & Joho, 2001)—who used a traditional IR search interface with a list of titles and a single full-text document window—found that even though their system retrieved documents containing part or all of the answers, in a large number of cases users did not examine these documents, rejected them or did not notice the answers when they did examine the documents. This suggests that the interface is an important component in interactive QA.

There are some emerging studies in the QA literature that explore factual QA with real users in a laboratory setting (Lin et al., 2003) and other more complex tasks in a naturalistic setting (Diekema et al., 2004; Liddy et al., 2004). Lin et al. conducted one of the first evaluations of interactive QA in an exploration of interfaces to support factual QA. In this study, Lin et al. sought to understand the relationship among the amount of text that a QA system returned with its answers, user performance, and preferences. Lin et al. compared four interface conditions: exact answer, answer-in-sentence, answer-in-paragraph, and answer-in-document

DOI: 10.1002/asi

and found that the majority of users preferred the answer-in-paragraph interface. Although there were no significant differences in completion time between interfaces, Lin et al. found that users required fewer iterations to complete the same task (i.e., users asked fewer questions of the system) according to interface: the most questions were asked of the exact-answer interface and the fewest questions were asked of the answer-in-document interface. Users stated that they liked the inclusion of the surrounding contextual information because it allowed them to verify answers and explore further information about the task, instead of explicitly asking more questions. A final noteworthy result was that when responding to multiquestion scenarios, users expressed a desire to ask both specific and general questions. Most current QA systems are not designed to support both types of question asking.

Diekema et al. (2004) evaluated the Knowledge Acquisition and Access System (KAAS), a QA system designed to help undergraduate students answer questions about aeronautical engineering. This evaluation took place in a real world environment, with undergraduate aeronautical engineering majors as system users. One goal of the study was to understand how users evaluate QA systems; data used for analysis came from system logs, as well as from survey responses to open-ended questions. Based on an analysis of student responses to the open-ended questions, Diekema et al. (2004) identified five major categories of evaluation criteria: system performance, answers, database content, display, and expectations: (a) *system performance* included criteria such as speed and reliability; (b) *answers* included criteria such as accuracy and relevance; (c) *database content* included criteria such as authority, scope, and size; (d) *display* was the most developed category and contained criteria such as querying style, question formulation assistance, and output; and (e) *expectations* appear as a general category related to students' desire to compare everything to Google. As did Lin et al. (2003), Diekema et al. found that answer context was an important consideration in interface design, especially with regard to the particular user group and functions of the system. Diekema et al. noted that users of KAAS are "learners in the field and not well served with exact answer snippets. For their task, they need answer context information to be able to learn from the answer text." Finally, Diekema et al. noted that questions students asked of the system did not resemble questions from more standard QA evaluation test collections. Instead, students used a broad range of question types including complex, multi-faceted questions.

Method

The data used in the study reported in this article were collected during an intensive evaluation of the HITIQA system. This evaluation was conducted with intelligence analysts as system users and took place during two, three-day workshops. The workshops were held at the Institute for Informatics, Logics and Security Studies at the State University of New York Albany where HITIQA was developed. During these workshops, analysts used HITIQA to prepare reports on various scenarios and participated in a number of evaluation exercises including questionnaires and interviews. The data used in this study comes from individual interviews and focus group interviews from the first of these workshops. A summary of the HITIQA system and evaluation workshops is provided below. This includes a description of both interview methods and of the method used to conduct the qualitative analysis reported in this article. A more thorough discussion of the HITIQA system and our evaluation method can be found in Wacholder, et al. (in press).

HITIQA System

HITIQA (Small et al., 2004) is a complex end-to-end QA system specifically designed to let users ask exploratory, analytical questions whose answers cannot be readily judged as right or wrong. The answers to analytical questions are considerably more complex than answers to factual and definitional questions such as those used in the TREC QA track. Answers to factual questions are usually shorter, more straightforward and not matters of opinion. For example, the correct answer to the factual question "Who shot Abraham Lincoln?" should contain the answer "John Wilkes Booth." In contrast, an adequate answer to an analytical question such as "How has Russia reacted to the bombing in Kosovo?" may encompass reactions by a variety of individuals and organizations; a brief one-dimensional answer (e.g., "The government was unhappy") typically does not satisfy the user's information need. An answer of some depth is required; depth depends on the characteristics of the person who asks the question and the context in which the question is asked.

HITIQA is intended to be used with a large text corpus that contains multiple documents that, in turn, contain partial answers related to one or more dimensions of the answer. Users ask open-ended questions to probe available information and express their information needs using any desired syntax; users are able to engage in an interactive quasi-natural language dialogue with the system in order to clarify their information needs and to use a visual interface to explore the available answer space. HITIQA returns answers in the form of paragraph-size units of text that can be saved into a report.

Summary of HITIQA Evaluation

The evaluation was from the perspective of the intended users of the system, intelligence analysts. The evaluation method was guided by the principles of participatory design (Schuler & Namioka, 1993) and was conducted as a formative evaluation (Flagg, 1990). In particular, the principles of participatory design allowed us to involve analysts as active collaborators in the design process rather than as experimental subjects. The formative evaluation approach furthered this collaborative approach because it allowed for a less rigorously structured and controlled evaluation environment.

DOI: 10.1002/asi

The availability of intelligence analysts provided us with an unusual opportunity to learn how useful they felt that HITIQA was, to observe their information seeking and use behaviors, and to observe their reactions to the evaluation method. The evaluation was conducted in two three-day workshops held a month apart.

Participants. Four United States Naval Reservists participated in Workshop I. All four participants were intelligence analysts for the United States government. Their participation was treated as a work assignment. There were three males and one female, with varying degrees of experience.

Scenarios, tasks & corpus. Analysts' primary task at the workshop was to prepare draft reports—similar to the kind that they might submit to their managers at work—in response to scenarios. Scenarios were complex questions that encompassed multiple subquestions. For instance, one scenario asked for a report on various aspects of al-Qaeda such as membership, sources of funding, and recruitment activities. To ensure that the scenarios were like those that analysts investigate at work, several government agencies collaborated in the development of five primary scenarios used in this workshop.

Analysts' primary task in their real jobs is the preparation of reports, usually written, that are submitted to managers for approval and then passed up the management chain. Thus, the task of preparing draft reports was selected as the work task in this study. Analysts were asked to prepare a report as close as possible to what they would prepare in their normal work environments. The combination of work task and scenario are similar to Borlund and Ingwersen's (1997) simulated work tasks in that an attempt was made to model analysts' typical work tasks and provide a context for accomplishing these tasks.

Although scenarios covered a range of topics, all were on the general topic of weapons of mass destruction, the choice of which was, in part, based on the document corpus. This corpus was a custom built corpus that consisted of data from the Center for Non-Proliferation Studies (CNS; 230 MB), which was used as part of the AQUAINT program. This corpus was supplemented with about 1.2 GBs of data collected from the Web on international politics and global security.

Data collection. Workshop I consisted of three main activities: training, preparation of reports based on scenarios, and evaluation. A number of data collection techniques were used during this workshop including observation, logging, questionnaires, and interviews. On the first day of the workshop, analysts were provided with an overview of the workshop and received training on HITIQA. On the second day, analysts used HITIQA to prepare reports on two different scenarios. After the preparation of each report, analysts completed an online evaluation questionnaire. Day 2 concluded with a forty-five minute focus group interview. Analysts spent the final day completing an exit questionnaire, participating in individual interviews, and participating in a final focus group. The focus group and individual interviews are described in more detail below; data from these interviews form the basis of the analysis reported in this article.

Focus group interviews. Two focus group interviews (FGI) were conducted. Focus group interviews were selected for the evaluation because they take advantage of the dynamics created by group discussion of a particular topic or issue. In particular, the dynamic of the group is thought to bring out aspects of the topic that may not be anticipated by researchers or emerged in one-on-one interviews (Lederman, 1996). It is further believed that the group context of FGIs provides a synergy that can potentially lead to incremental increases in output.

Focus group interviews were conducted in a private, general purpose meeting room. The first of these occurred at the end of the second day of the workshop; the second interview was conducted at the end of the third day. All four analysts were present for the first FGI, but only three analysts were present for the second. A Focus Group Guide, created by the researchers, identified key questions chosen to stimulate discussion and elicit feedback. Focus groups were lead by one member of the investigative team, while other members of the investigative team took notes. Each interview was tape-recorded and later transcribed for analysis. Each interview lasted approximately forty-five minutes. A recording failure occurred during one of these interviews, so only notes were used during the analysis of data from this interview.

Individual interviews. On the final day of the workshop, each analyst participated in a one-hour individual interview session with a member of the investigative team (the analyst who left early was interviewed at the end of the second day). Interviews can serve a number of purposes, but we were most interested in using them in this study as an exploratory technique to obtain information about analysts' preferences and attitudes about HITIQA. Interviews were appropriate in this context because they provided analysts with an opportunity to reflect orally on their use and perceptions of HITIQA. Interviews also allowed analysts to communicate feelings and attitudes without having to choose from a selection of responses as they would do with a closed questionnaire.

All interviews were conducted following an Interview Schedule which provided instructions to the interviewer for conducting the interview. One purpose of the interview was to obtain feedback from analysts about their use of HITIQA, and in particular, about HITIQA's interface. We wanted to insure as much uniformity in the delivery and course of the interviews as possible because different members of the investigative team were conducting interviews. The Interview Schedule provided a structure for accomplishing this, and it presented a series of interface screen shots that were used to guide the interview. Interviews took place in front of a

DOI: 10.1002/asi

computer that could be used to access HITIQA; analysts were encouraged to use the system at any time during the course of the interview to provide examples or demonstrations, or to elaborate on their feedback. Each interview was tape-recorded and later transcribed for analysis. Interviews were conducted in private offices.

Qualitative Analysis Techniques

Analysis of transcriptions of focus group interviews and individual interviews form the basis for the development of the classification scheme presented in this article. An inductive method was used to identify and categorize comments analysts made during these two types of interviews. The initial step in the qualitative analysis used in this study included (a) reading the interview transcripts, (b) writing notes and memos on the data, (c) developing tentative ideas about categories, (d) classifying utterances into these tentative categories, and (e) iterating the process until all relevant utterances were classified.

In this analysis, there was no definitive unit of analysis; instead, the unit of analysis for this study can be considered as similar to what Ericsson and Simon (1993) describe as verbal statements. Verbal statements are utterances of any length. Verbal statements are appropriate units in this study, because the data analyzed in this study did not exist in complete sentences, paragraphs, or other grammatical structures but were produced as a result of conversation (i.e., interviews). When coding this type of data, the *concept* functions as the organizing principle, and thus the size of units can vary within a given communication. A given code category may be applied to utterances of different lengths, and codes are not mutually exclusive. That is, some utterances can be classified into more than one category. Further, this approach is not exhaustive; utterances unrelated to the objectives of one's study do not have to be classified into any category.

Coding was the main categorization strategy used in this study. Coding was used to fracture data, rearrange it into categories, and facilitate the comparison of data within and between categories (Strauss, 1987). Although categories can be drawn from existing theory and used in a deductive manner during coding, in this study coding categories were developed inductively during analysis (Maxwell, 1996). The process of open coding was used to discover codes and classify utterances. Strauss and Corbin (1990) characterize open coding as, "the part of analysis that pertains specifically to the naming and categorizing of phenomena through close examination of data … during open coding the data are broken down into discrete parts, closely examined, and compared for similarities and differences" (p. 62). Hence, in open coding, codes are suggested by the researcher's examination and questioning of the data. This process is iterative; when new codes are added, previously categorized units are reviewed to see whether utterances need to be reclassified. Coding ceases when saturation has been reached and all relevant utterances have been classified.

In addition to open coding, memoing was used extensively throughout the process to capture analytical thinking about the data and preliminary descriptions of emerging categories. Memo writing was further used to define the properties of each category, specify conditions under which each category developed and changed, and note the consequences of each category and its relationships with other categories (Charmaz, 2002). Following open coding, selective or focused coding occurred, where the researcher used the most frequently appearing initial codes to sort, synthesize, and conceptualize data (Charmaz). Once coding was complete, codes and utterances were extracted from transcripts and memos, and they were placed into an Excel spreadsheet to facilitate organization and sorting.

Findings

Twelve categories emerged initially from the analysis. Two hundred and twelve utterances made by four intelligence analysts during individual interviews and two focus groups were coded into these categories. After coding data iteratively, these twelve categories were reduced to seven. A description and discussion of each category, along with supporting quotes from interview transcripts, are presented below.

Redundancy

One of the first categories to emerge from the data was that of redundancy of the answers provided by the system. Users made numerous comments regarding the duplication of information provided by the system. In most cases, duplication was an issue because users almost always opened the full-text of documents, from which answers were generated, to check to see if there was more relevant information in the surrounding text or to verify the answer. If the system generated more than one answer from a single document, then there was a good chance that users would see the same text from the same document more than once. One user noted, "a lot of these answers and summaries are, from what we know, redundant to the same source" and another noted, "there was no way to distinguish, or to discriminate between documents you've already seen and documents that the system provided again on another query."

While placing markers next to documents users have opened ("it could still be a link, but it actually give more information about the document … I could just see all four of them are from the same document") or grouping answers with respect to source might address these issues ("I would have liked to have seen the answers grouped by document; or document and then the question that it addresses"), these comments elucidate an important issue with respect to user models of the question-answering interaction. Question-answering systems are designed to return answers to questions, rather than full-text documents, even if a single document provides answers to multiple, different questions. However, full-text was important to our users. Viewing the

DOI: 10.1002/asi

full-text, of course, can lead users to discover additional information about other questions that they had planned to ask in future interactions. If users ask these questions again to find additional information, then it is likely that the system will display this previously viewed text again in the form of a different answer, because the system model is such that users look at answers rather than full-text. Analysts commented, "I would like the system to know what information was already displayed … to know that that document is one that I have already seen" and "If you are in the same session and ask more than one query, then it should know not to go back to that same one [document] because you've already seen that." These user behaviors are consistent with Lin et al.'s (2003) work on factoid QA systems that found that users preferred larger portions of text because they could discover additional pieces of information about their need, and with Belkin et al. (2001) and Wu et al. (2001) who found that less iteration occurred with interfaces designed specifically to support the QA task.

Novelty

Analysts made many comments about the novelty of information provided by the system. Analysts' comments about novelty were of two types. The first type was related to the system providing information that the user has not seen previously. This type of novelty is the opposite of redundancy, and it is certainly a characteristic that almost any good information system should have. The second type of novelty describes the system's ability to provide users with novel ways of thinking about the scenario. Analysts particularly welcomed this type of novelty when they were stuck. For one analyst this was an unexpected consequence of using the system, "But frankly I was surprised at some of the fields that did come up, that the system responded with. I think that they were beneficial or helpful in my particular scenario." Another analyst speculated about how stimulating a system might be, "why didn't you think about this, or why didn't you go through this, you know, different angles and stuff."

A key feature of HITIQA engages analysts in an interactive dialog after their initial questions. The purpose of this dialog is to advise analysts of other themes that were found in the data. Novelty, with respect to stimulating thinking, was a consequence of this interaction as evidenced in the following comments, "it's the greatest thing because then I don't have to ask that question so it asks for me," and "okay, I thought that eliminates me asking the question … and I said oh yeah, that's another question, you know, I was gonna ask it. So yeah, that is cool. I want to capture that. So it provides me this information." Of course, one danger with any type of system-driven dialog is redundancy: "But the second and third time, I'm like thinking to myself, okay it's asking the same question, but is it related to this question or is it just doing the general, pulling everything all over again." Clearly, introducing a feature explicitly to stimulate a user's thinking about a topic can be risky but, if done effectively, can be valuable.

Completeness

This category of responses is related to the system's ability to provide all relevant information. With respect to the task which analysts were engaged, this included finding all relevant information to all facets of the scenario. For analysts, it is extremely important that all potentially relevant information is retrieved by the system. Given the critical nature of their task, missing one key piece of information could result in serious security consequences:

> So, it's an unfortunate thing because intelligence by its nature is a guessing game, frankly. You take what's available and you try to extrapolate. So, I'm not sure the system as it stands now, in my mind, gave me enough information to try to put together an 80% solution. You're never ever, I don't think you're ever going to reach that 100% stage.

> You don't want to feel that there's something huge out there that you missed just because you didn't type in the specific term that the computer likes versus another one.

> You want all the information related to it and the computer asking me supplemental questions [i.e., dialog] and I think more times than not you would want that information, regardless, to look at unless something is just so out in left field.

Other analysts described their desire to see all potentially relevant documents even if this meant extra work for them sorting through results:

> Well the default is, I'd rather have it presented to me and then discard it than to never see it at all.

> I'm more interested in what I'm missing than what I'm initially getting.

> I didn't, but because I went through and literally looked at every one.

> It was like, dump it on me and I'll sort through it.

For analysts, the cost of missing a relevant piece of information is far greater than the cost of spending additional time sifting through documents. Another analyst expressed a desire to understand more about what the system "believed" were the most relevant or important answers, "What matters is how pertinent was that information to the question I had to answer or the problem I had to solve or the issue I had to address, and there is nothing in here that addresses that." Analysts' beliefs about how completely the system answered their questions contribute to their confidence in the system. The issue of confidence is described more fully in a subsequent section.

Specificity, Selectiveness, and Comprehension

Comments related to this category were about the level of specificity with which the system provided information, and to its selectiveness, that is, the level of specificity with which

DOI: 10.1002/asi

analysts could question the system. The system's ability to provide succinct and specific answers to questions pleased one analyst, "On a good note, I asked 'Where does Al Qaida get its funding?' and I was surprised at the information provided. It responded directly to the point, I was able to capture the information in a very short order. That was surprising to me." Although another analyst identified some limitations to providing specific information:

> I think trying to get to the granularity that you're shooting for is admirable, but it gives an operator too many choices, and frankly, it gets to a stage where the fewer choices you have, the better off you are.

Many responses were related to the level of specificity with which analysts could question the system. In particular, analysts indicated that they were interested in asking both general and specific questions during their interactions with the system. One analyst noted, "You start general and then get more specific; you want to be sure and get everything so you start general and then drill down." This comment illustrates both a desire to conduct an exhaustive search, to ensure finding all potentially relevant documents, as well as a desire to question the system in different ways. Another analyst commented, "I felt like I had to ask a general question to try to get some information rather than ask a specific question for a specific answer." This finding is similar to the findings of Lin et al. (2003) and Diekema et al. (2004).

Related to this issue were other comments analysts made about their line of questioning and the system's lack of sensitivity to nuances between questions:

> When I asked it a new question, this is the same thing, but it was a different line of questioning that I was asking for information.

> I have a lot of questions that I was asking it and it seems like it was repeating the same stuff.

> The system would take the keyword from the query and try to develop answers using the keywords, but I'm wondering to what extent does the program grasp the context.

> There's no way to distinguish between things that you've already looked at from another question's perspective.

This idea of a "question's perspective" is an interesting one, as it demonstrates the great challenges that face natural language processing systems: Users can distinguish between questions that have very minor linguistic differences. Although two questions may be very similar in content, small linguistic differences and the context and order in which users ask questions give questions different "perspectives." One analyst noted, "it's hard to put that together in a way the machine can understand, that you know will allow the machine to understand it as well as you would," while another commented that "the system couldn't figure out what I was asking for." The analyst who made this last comment went on to express the desire for more user training when users are expected to enter question queries, not keywords:

> The other thing is I think there's got to be some kind of understanding by the analyst of how we capture questions. What terms can you use? Can you use relationships? Or can you use difference? Or does it make a difference, you know? Does it understand the range of interrogative that could possibly be presented to it? For example, does it understand "launch" or "release"?

Because the analyst engages the system in dialogue, it is natural for the analyst to wonder whether the system really comprehends her intentions.

Confidence in System

Confidence is related to analysts' beliefs regarding the appropriateness and comprehensiveness of answers that the system returns, and to their general belief that the system has returned all relevant information. Confidence is related in many ways to the categories described previously, and was one of the most prevalent themes in the data. Analysts gained confidence by evaluating answers that the system provided. Analysts noted:

> Confidence decreases when I see strange answers or can't find terms that explain why something was retrieved.

> So that goes to the confidence level of the user because it almost might have been buried, if I hadn't seen that, I would have thought well, gee, what other goldmine out there is it not grabbing at because I didn't happen to ask one other little specific aspect of something.

> That's why I couldn't rely on say this information here. I still have to go back and validate it.

As is evident in these responses, confidence was also related to analysts' understanding of the system and their models of how the system worked based on the query they entered and the answers they received. Comments that illustrate this finding were, "Confidence decreases when users see strange answers or can't find terms that explain why something was retrieved," and that users "could be more confident in what they're getting particularly if they see changes when they do things." Another analyst noted:

> One other thing I can think of is that yesterday when I had this system and I asked it one question and it came back with answers that were not right and then I reworded or rephrased the question and eliminate the word "exports" and I think I put in the word "receives" and then it seemed to understand that better and actually gave me an answer back.

Other comments demonstrated it was unclear to analysts how the system worked, "I couldn't always see distinctions between answers with minor differences in how questions were

DOI: 10.1002/asi

asked," and "I was surprised at the sensitivity of the system to certain ordering of terms."

Establishing trust and building users' confidence in the system's ability to return the "correct" answer will be a significant challenge for any interactive QA system. For tasks where users are essentially constructing "answers" from distributed sources, as they were in this study, this is even more of a challenge, because it is more difficult to quickly establish and test the system's effectiveness by, for instance, asking factual questions about which one already knows the answer. Supporting analysts' ability to understand the relationship between their questions and the answers provided is one way to accomplish this, "it seems to me that the question I was asking was not related to the type of information it retrieved, but I clicked on it anyway to find out what was in there and some of it was relevant, some of it was not." Analysts identified several things that would help them gain confidence in the system's abilities:

> If I were to use this system and go back again and do the same scenario that I did this morning, I would expect to get the same documentation yet in my experience I would go back and use this again and get different source documents that would really affect my confidence in the system ... consistency and things like that. I think one thing with confidence is having people go back to do the same thing again and again.
>
> The only thing that I can think for myself that would boost my confidence level in this process, would be the ability to go through and run similar scenarios, different process, and see which one gives me the best output, and one of the questions in the survey was that in my workplace would this be helpful, or would this improve my productivity, or would this provide a better quality product.

Two analysts noted contrasts between human confidence criteria and system confidence criteria:

> Yes, these criteria for confidence are totally different from mine. See that's what I mean. There are two wholly different standards at play.
>
> One of the things that I noticed is that a lot of really good information is in some of those lower confidence level frames.

In many cases, analysts' confidence in the system was related to the discrepancy between their perceptions of the relevance of the information provided by the system and the system's indication of confidence.

For one analyst in this study, verifying answers in the full-text was one way to build confidence in the system: "Analysts' confidence level will decrease with the less text you look at. I want to be able to look at the sources." This comment also supports the discussion developed in the section on redundancy with regard to analysts' desires to view full-text, even when they use a question-answering system.

Coherence and Context

Several comments made by analysts illustrated a desire to understand the origin of answers supplied by the system and the larger context in which they were situated. In particular, analysts often found answers incoherent because they lacked context, and their relationship to one another was not obvious. As mentioned previously, this, in part, led analysts to open and read the full-text of most documents, "analysts want access to full text because they want to see context." Most remarks analysts made about context related to the answer summaries the system provided.

> A good number of them I saw, I guess they're based on keywords, were completely out of context with the question being asked.
>
> I didn't find the bold face summaries useful because they seemed to be completely out of context.
>
> I did not find those helpful in sorting through good information or useful information ... the summaries again, were either so small or disjointed, I didn't find those useful. About 50% of them I found useless glancing over them whether or not they had any information pertaining to my query at all. If it contains some information, I definitely had to open it up, open up the full document to get a feel for it. And at that point I came in to the you know, the problem having to do with the multiple document repeated over and over again.

Another analyst expressed a desire to see the relationship between various answer summaries provided by the system, "how does that relate to say, you know, what we got here in paragraph one, two, three, four, five paragraphs under what they call two answers."

Again, these findings are consistent with both Lin et al. (2003) and Diekema et al. (2004). In particular, Lin et al. found that answer-in-paragraph was the most effective unit in their tests of factual QA. Wu et al. (2001) also found an advantage to providing query-biased sentences as document surrogates, over a traditional search results surrogate for factual questions, and in Beaulieu et al. (2001), it appeared that users missed answers even though the relevant full-text document was retrieved because the display was not designed to support the factual QA task. Clearly, a good QA system should provide interface features to assist users with understanding the context of retrieved answers, as well as the relationship between retrieved answers. A good QA system should further work to resolve ambiguous question terms and track question-asking over time so that it understands the context in which a particular question is being asked and how this context has evolved.

Integration & Control

In addition to the QA system tracking what a user does, analysts also expressed an interest in being able to keep track of their activities and integrate various micro- and

DOI: 10.1002/asi

macro-level activities and functions. With respect to search, does the system allow analysts to move freely from one view or function to the next? Does the system allow analysts to switch seamlessly from one activity to the next? Can analysts keep track of their activities? With respect to task integration, are the various functionalities of the system effectively integrated to allow analysts to complete their goals (i.e., create report)? With respect to control, we noticed that analysts were not always willing to let the system act as an independent agent; thus, this dimension is concerned with the degree to which an analyst can be sure that she is controlling it, and that it is doing what she intends.

Analysts often encountered information, while they researched one question, which either gave them ideas for other questions or was related to another question that they had planned to ask in the future (this is very closely related to the novelty category discussed previously).

> That sounds like a pretty interesting piece of information on the Security Council and on India's nuclear program so maybe I could form an intel question and then use it later on for another query,

> I wanted more information about that by just, you know, a quick glance, but I'm like do you want me to write that down? I'll ask it another question, but I will still keep this one as a variable.

> Does it pertain to the query I had said I had posed to it or can I use it for another category?

However, analysts noted that the system did not provide adequate support for such discoveries and instead questioning was seen as a discrete activity.

> If I went back and I thought the system satisfied or just gave me information about this particular query then I would save it. And then I would go back and ask him another question. So it would kind of, I would be doing a different process. But if I stayed within say the dialogue box, kind of be like a chat in a sense like everyone was talking about, and then I could just continue on with the conversation versus going out, logging out each time in a sense and asking another question. And then that way you probably keep a running history of everything that you were doing rather than just going back in.

> I guess rather than starting another query because then you lose your train of thought because say for example, I go in here, I look at this document here and I notice that something that sticks out.

Analysts also expressed a desire for features that would allow them to organize their work.

> That's where I tried to keep on track and know which ones [documents] that I actually looked at but I still duplicated it anyway.

> So by setting up a file, what I can do later is probably transfer the information that I got under the original question into the category that I set up. So that was the only way that I thought I could do that. That's an extra step for me you know.

> At this point, I almost wish that there was some way I could drag the groups, and be able to group them, if I was trying to develop a trend from the type of documentation that this system is providing me.

Although the system provided a history function that allowed users to review and save the dialog and answers for each query and session, analysts still expressed a desire to view the history of their work in different way.

> It would be nice if in the process of doing an ongoing query we could see previous queries. What was the question I asked last, the question I asked three iterations ago so that we can tailor all our questions with the ability to save a line of thought as a file.

> Once I went down a thread through the query, I really wish I could have gone back, saved that thread, gone back and modified my search to see if I could pull up different documents but still retain the history function ... you need to be able to see where you've gone and compare the different threads you've taken as far as adjusting your queries.

Limitations of Study

There are several limitations of this study that must be acknowledged before conclusions can be stated. In general, interview techniques are less reliable than controlled, laboratory experiments. If another set of people conducted these interviews, it is possible that another set of data would have been obtained. We took numerous steps to mitigate this, especially because analysts had different interviewers for individual interviews. Our efforts included training all interviewers and using a detailed interviewer schedule. However, it still may be the case that relationships between analysts and their interviewer impacted the quality and depth of responses. In making comparisons across interview transcripts, we noticed no large differences.

There are several potential threats to the validity of the data, most of which relate to analysts feeling comfortable enough to speak their opinions during interviews and avoid censoring themselves. For instance, analysts may have wanted to appear a certain way to researchers, spare the feelings of the researchers, or been embarrassed about their opinions. These issues can be more acute in focus groups because other participants are present. In general, our experiences with analysts indicated that these were not concerns. Analysts spoke candidly with us both on record and off record about the system. A further danger of focus groups is conformity of responses. To address this threat, the focus group moderator was trained to recognize dominate participants and encourage participation from all analysts. A final threat to validity, which we were unable to address, comes from our reliance on handwritten notes for one of the focus groups because of a recording failure.

DOI: 10.1002/asi

This is a study of one system at two points in time, and specific issues identified may be quite different from those which would emerge from a study of another system. While participants were representative of the real users of such systems, the actual number was quite low, which limits our ability to generalize. These participants also volunteered for this assignment, so it is likely that they possessed a special interest in information systems. Furthermore, the work of the analytic community is complex and diverse, and interviews of a different set of users might have revealed a different set of issues. The specific tasks were designed to reflect the real work of such analysts, and according to the analysts, they generally met that criterion. However, a different set of tasks or corpus might have revealed substantially different results.

Conclusions

With these limitations in mind, we nonetheless feel that it is valuable to examine the key issues that have emerged, explore their impact on the development of the system, and look for some organizing principles that might have greater generality than the specific findings here. Examining specific issues that were revealed by our analysis of interview data we find a rather daunting collection, which corroborate and extend the work of Diekema et al. (2004): Redundancy; Novelty; Completeness; Selectiveness, Specificity and Comprehension; Confidence; Coherence and Context; and Integration and Control. In many cases, identical or similar criteria were found, such as completeness, relevance, redundancy, novelty, system reliability, and output organization. New concepts also emerged and our understanding of the results of the current study can be furthered by grouping the concepts into the following structure.

Characteristics of the Information Provided

- *Novelty & Redundancy*: Does the information add to or duplicate the user's knowledge?
- *Completeness*: Is the answer complete?
- *Selectiveness & Specificity*: Does the information match the user's need? Is the information at an appropriate level of detail?
- *Coherence & Context*: Does the information make sense? Can the information be integrated into what the user already knows?

Characteristics of the Interaction

- *Integration & Control*: Does the system allow the user to move seamlessly between activities and lines-of-questioning? Does the system keep track of what the user has experienced and encountered?
- *Comprehension*: Does the system understand what the user wants?
- *Confidence*: Does the user have confidence that the
  - answer is complete?
  - system has understood the question?

We believe that these two broad categories will apply to the study of many other types of systems. To some degree they reflect the major categories observed in the study of library systems (Saracevic & Kantor, 1997), but without the category called "affective." Instead, there is sharpened focus in this work on the confidence that the analyst can have as a result of using the system. This is specific to the context studied here, but may well generalize to any situation in which the results of using the information system will form the basis for decisions that can have serious consequences.

This study provides a plan for gathering and analyzing qualitative information, in a situation involving a real system, real users, and realistic tasks. This has been of great value in the evolution of the system being studied. For example, the issue of redundancy was quickly mapped to a specific issue in system design, and software was added that flags documents that have been previously presented to the analyst so that they are not presented again. Furthermore, a second release of the HITIQA system addressed the problem of integration by providing tools for organizing materials that are intended to increase the analyst's sense of control over the system, and thus their confidence in the final results produced. These are examples of how something which may seem obvious to developers, after the fact, is elicited by systematic interviews and analysis. Without the systematic approach taken here, comments by one analyst or another might not be assembled into a sign that change in the system is needed. We believe that the approach described here will transfer readily to the evaluation of other types of interactive information system. In particular, if it can be integrated with a rapid prototyping approach to system development, it can substantially reduce the cost and time required for developing user-centered, interactive systems.

The complexity of the findings reported here provides an indication of the challenges involved in assessing the next generation of interactive information access systems, where systems' abilities to respond to questions in a more human-like way will increase the difficulty of evaluating systems. We hope that the method and findings from our study of HITIQA will help lay the groundwork for future research in this area.

Acknowledgements

This article is based on work supported in part by the Advanced Research and Development Activity (ARDA)'s Advanced Question Answering for Intelligence (AQUAINT) Program under a contract to the University at Albany, SUNY.

References

Bates, M.J. (1989). The design of browsing and berrypicking techniques for the on-line search interface. Online Review, 13(5), 407–431.

Beaulieu, M., Fowkes, H., & Joho, H. (2001). Sheffield interactive experiment at TREC-9. In D. Harman & E.M. Voorhees (Eds.), TREC-9, Proceedings of the Ninth Text Retrieval Conference. Washington, D.C.: GPO, 645–653.

Belkin, N.J., Keller, A.M., Kelly, D., Perez-Carballo, J., Sikora, C., & Sun, Y. (2001). Support for question-answering in interactive information seeking: The Rutgers TREC-9 interactive track experience. In D. Harman & E.M. Voorhees (Eds.), TREC-9, Proceedings of the Ninth Text Retrieval Conference. Washington, D.C.: GPO, 463–474.

DOI: 10.1002/asi

Borlund, P., & Ingwersen, P. (1997). The development of a method for the evaluation of interactive information retrieval systems. Journal of Documentation, 53(3), 225–250.

Charmaz, K. (2002). Qualitative interviewing and grounded theory analysis. In J.F. Gubrium & J.A. Holstein (Eds.), Handbook of interview research: Context and method. CA: Sage Publications, 675–694.

Diekema, A.R., Yilmazel, O., Chen, J., Harwell, S., He, L., & Liddy, E.D. (2004). Finding answers to complex questions. In M.T. Maybury's (Ed.), New directions in question answering. Cambridge, MA: MIT Press, 141–152.

D'Souza, D., Fuller, M., Thom, J., Vines, P., Zobel, J., de Kretser, O., et al. (2001). Melbourne TREC-9 experiments. In D. Harman & E. M. Voorhees (Eds.), TREC-9, Proceedings of the Ninth Text Retrieval Conference (pp. 437–451). Washington, DC: GPO.

Ericsson, K.A., & Simon, H.A. (1993). Protocol analysis: Verbal reports as data (revised edition). Cambridge, MA: The MIT Press.

Flagg, B.N. (1990). Formative evaluation for educational technologies. Hillsdale, NJ: Lawrence Erlbaum Associates.

Golovchinsky, G., & Chignell, M.H. (1997). The newspaper as an information exploration metaphor. Information Processing & Management, 33(5), 663–683.

Hersh, W., & Over, P. (2001a). Introduction to a special issue on interactivity at the Text Retrieval Conference (TREC). Information Processing & Management 37(3), 365–367.

Hersh, W., & Over, P. (2001b). TREC-9 interactive track report. In D. Harman & E.M. Voorhees (Eds.), TREC-9, Proceedings of the Ninth Text Retrieval Conference (pp. 41–50). Washington, D.C.: GPO.

Lederman, L.C. (1996). Asking questions and listening to answers: A guide to using individual, focus group and debriefing interviews. Dubuque, IA: Kendall/Hunt Publishing.

Liddy, E.D., Diekema, A.R., & Yilmazel, O. (2004). Context-based question-answering evaluation. In K. Jarvelin, J. Allan, & P. Bruza (Eds.), Proceedings of the 27th Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR '04) (pp. 508–509).

Lin, J., Quan, D., Sinha, V., Bakshi, K., Huynh, D., Katz, B., et al. (2003). What makes a good answer? The role of context in question answering. In M. Rauterberg, M. Menozzi, & J. Wesson (Eds.), Proceedings of the Ninth IFIP TC13 International Conference on Human-Computer Interaction (INTERACT 2003), Zurich, Switzerland.

Maxwell, J.A. (1996). Qualitative research design: An interactive approach. CA: Sage Publications.

Maybury, M.T. (2002). Toward a question answering roadmap (MITRE Technical Paper). Retrieved June 26, 2006, from http://www.mitre.org/work/tech_papers/tech_papers_02/maybury_toward/maybury_toward_qa.pdf

Maybury, M.T. (2004). Question answering: An introduction. In M.T. Maybury's (Ed.), New directions in question answering (pp. 3–14). Cambridge, MA: MIT Press.

Saracevic, T., & Kantor, P. (1997). Studying the value of library and information services. II. Methodology and Taxonomy. Journal of the American Society for Information Science, 48(6), 543–563.

Schuler, D., & Namioka, A. (1993). Participatory design: Principles and practices. Hillside, NJ: Lawrence Erlbaum Associates.

Small, S., Strzalkowski, T., Liu, T., Ryan, S., Salkin, R., Shimizu, N., et al. (2004). HITIQA: Towards analytical question answering. In A. Gelbukh (Ed.), Proceedings of the 20th International Conference on Computational Linguistics, Geneva, Switzerland.

Strauss, A.L. (1987). Qualitative analysis for social scientists. Cambridge, UK: University of Cambridge Press.

Strauss, A., & Corbin, J. (1990). Basics of qualitative research: Grounded theory procedures and techniques. CA: Sage Publications.

Voorhees, E.M. (2001). The TREC Question Answering Track. Natural Language Engineering, 7(4), 361–378.

Voorhees, E.M. (2003). Evaluating the evaluation: A case study using the TREC 2002 Question Answering Task. In W. Daelemans & M. Osborne (Eds.), Proceedings of the Seventh Human Language Technology Conference on the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003), Edmonton, Canada (pp. 181–188).

Wacholder, N., Kelly, D., Rittman, R., Sun, Y., & Kantor, P. (in press). A model for realistic evaluation of an end-to-end question answering system. Journal of the American Society for Information Science and Technology.

Wu, M., Fuller, M., & Wilkinson, R. (2001). Searcher performance in question answering. In W.B. Croft, D.J. Harper, & J. Zobel (Eds.), Proceedings of the 24th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR '01), New Orleans, LA (pp. 375–381).

DOI: 10.1002/asi