In Proceedings of the Ninth IFIP TC13 International Conference on
Human-Computer Interaction (INTERACT 2003), September 2003,
Zurich, Switzerland

What Makes a Good Answer? The Role of Context in Question Answering

**Jimmy Lin, Dennis Quan, Vineet Sinha, Karun Bakshi,
David Huynh, Boris Katz, and David R. Karger**

MIT AI Laboratory/LCS
Cambridge, MA 02139, USA
{jimmylin,dquan,vineet,karunb,dfhuynh,boris}@ai.mit.edu, karger@lcs.mit.edu

Abstract: Question answering systems have proven to be helpful to users because they can provide succinct answers that do not require users to wade through a large number of documents. However, despite recent advances in the underlying question answering technology, the problem of designing effective interfaces has been largely unexplored. We conducted a user study to investigate this area and discovered that, overall, users prefer paragraph-sized chunks of text over just an exact phrase as the answer to their questions. Furthermore, users generally prefer answers embedded in context, regardless of the perceived reliability of the source documents. When researching a topic, increasing the amount of text returned to users significantly decreases the number of queries that they pose to the system, suggesting that users utilize supporting text to answer related questions. We believe that these results can serve to guide future developments in question answering interfaces.
Keywords: question answering, user study, interface design

1 Introduction

Question answering (QA) has become an important and widely-researched technique for information access because it can deliver users exactly the information they need instead of flooding them with documents that they must wade through. Current state-of-the-art systems are capable of answering more than eighty percent of factoid questions such as "what Spanish explorer discovered the Mississippi" in an unrestricted domain (Voorhees, 2002). Despite significant advances in the underlying technology of question answering systems, the problem of designing effective user interfaces has largely been unexplored.

Developments in question answering have focused on improving system performance against a standard set of questions; the QA track at the TREC conferences (Voorhees, 2001; Voorhees, 2002) is a notable example. However, such batch-run experiments neglect an important aspect of the information access process: human interactions with the actual system. Because system improvements, as measured by batch experiments, may not translate into actual benefits for end users (Hersh et al., 1999), the problem of computer-human interaction should be studied in parallel.

With notable exceptions, e.g., (Katz et al., 2002), current question answering systems are text-based, in that they return to users fragments of text containing the answer to their queries. In this sense, question answering is related to other information access techniques, e.g., document retrieval, in which entire documents are retrieved, and passage retrieval, in which paragraph-sized chunks of text are returned. The unique challenge and advantage of question answering systems is the promise to deliver succinct answers that directly satisfy users' information needs, phrased in natural language. Naturally, this begs the question: What exactly qualifies as a succinct answer? How much text should a question answering system return? These are the questions we seek to explore.

We believe that the most natural response presentation style for question answering systems is *focus-plus-context* (Leung and Apperley, 1994), which is closely related to the *overview-plus-detail* (Green et al., 1997) presentation style. A system should directly answer the user's query and provide additional contextual information. Since most current question answering systems extract answers from textual documents, the text surrounding the answer serves as a natural source of context. Although images, sounds, and even multimedia segments may provide better answers, we focus on textual responses in this paper. Here is a sample interaction with a question answering system (the *focus* in bold within the paragraph *context*):

> **Question:** Who was the first man on the moon?
> **Answer:** Neil Armstrong
>
> Neil Armstrong was the commander of the Apollo 11 mission to the moon in 1969. **On July 20, 1969, Armstrong became the first man to walk on the moon**, and made his famous statement, "That's one small step for a man, one giant leap for mankind."

Given this overall setup, we attempt to address the question of *how much context*, i.e., how much text

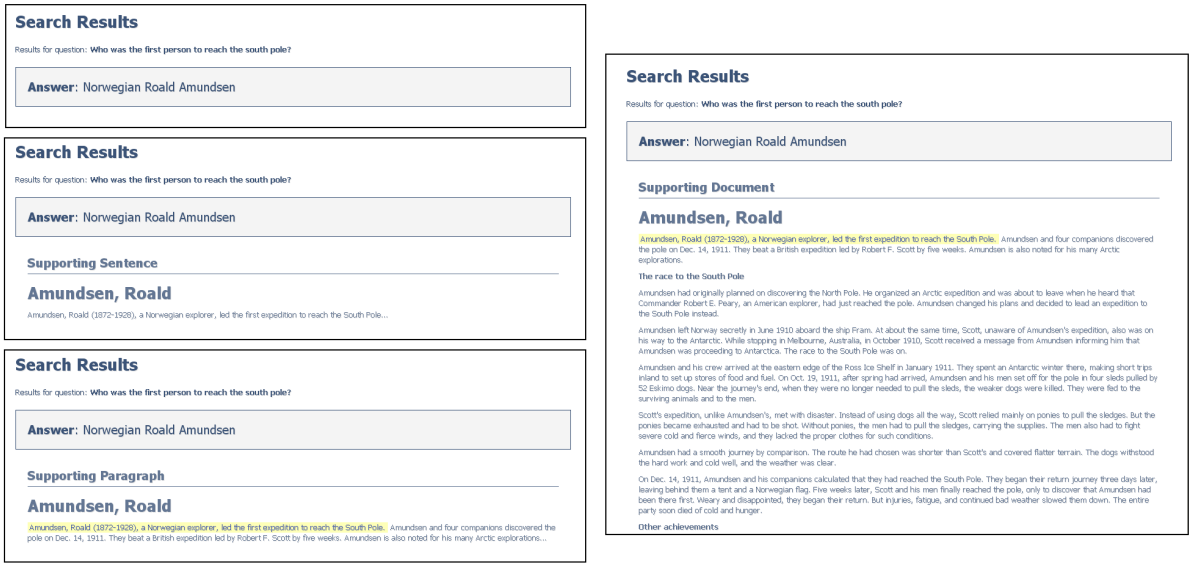

Figure 1: Various interface conditions tested in this experiment. Shown above are the responses to the question "Who was the first man to reach the south pole?": exact answer (top left), answer-in-sentence (middle left), answer-in-paragraph (bottom left), answer-in-document (right).

should a question answering system return to the user. We explored two variables that affect context preferences: source reliability (trustworthiness of the source text), and scenario size (i.e., whether the user asks a single question or a set of related questions).

2 Related Work

Although question answering systems have been studied extensively (Katz, 1997; Brill et al., 2001; Hovy et al., 2002; Katz et al., 2002; Moldovan et al., 2002; Lin et al., 2002), most research has focused on the underlying answer extraction algorithms. To our knowledge, no studies regarding the effects of context have been conducted on QA systems. However, the use of context in traditional information retrieval (IR) systems has been extensively studied. Recently, the effectiveness of spatial and temporal contextual clues (Park and Kim, 2000), category labels (Dumais et al., 2001), and top-ranking related sentences (White et al., 2002) has been explored empirically through user studies. Furthermore, the Interactive Track at TREC has generated interest in information retrieval interfaces; for example, Belkin et al. (2000) compared different views of IR results for a question answering task.

From these studies, it is unclear whether results gathered from studying traditional information retrieval interfaces can be directly applied to question answering.[1] We believe that context has a fundamentally different purpose for question answering and warrants separate research. Because information retrieval systems return a list of documents that the user must then browse through to extract relevant answers, research has been focused on supporting this browsing behavior and reducing cognitive load. For example, structural and temporal contexts help users navigate through a collection of hypertext documents; category labels give users a general idea of the documents' topics. The goal of question answering systems is very different: they seek to directly provide information that satisfies the user's information need, obviating the need for browsing. Thus, we believe that the role of context in QA systems is not to support browsing, but rather to justify the answer and to offer related information.

3 Interface Conditions

Since the amount of context that can be returned to the user is a continuous variable, we had to "discretize" context in order to support our experiments. Under the focus-plus-context framework and taking into account natural language discourse principles, we developed four different interface conditions. In each case, the focus was on the answer to the user question. The context was simply the text surrounding the answer, which varied in length for the different interface conditions. In more detail (see Figure 1):

- **Exact Answer.** Only the *exact* answer is returned to the user, without any additional context. For ex-

[1] See Hearst (1999) for an overview of IR interfaces.

ample, the exact answer to "when was the Battle of Shiloh" would be *April 6-7, 1862*. Exact answers are most often named entities (e.g., dates, locations, names), noun phrases, or verb phrases.

- **Answer-in-Sentence.** The exact answer is returned to the user, along with the sentence from which the answer was extracted.
- **Answer-in-Paragraph.** The exact answer is returned to the user, along with the paragraph from which the answer was extracted; the sentence containing the answer is highlighted.
- **Answer-in-Document.** The exact answer is returned to the user, along with the entire document from which the answer was extracted; the sentence containing the answer is highlighted.

4 User Study

We conducted a user study to investigate the effects of two variables on user preferences regarding context (the interface conditions that were described previously): reliability of source and size of scenarios. We hypothesized that trustworthiness of the source would be inversely correlated with the amount of context required for a user to judge a particular answer, i.e., the user would require more context to accept an answer from a less trustworthy source than from a more trustworthy one. We also hypothesized that when users are researching a topic, i.e., asking multiple related questions, context would play an important role—the answer to related questions might be found in the surrounding text. By presenting users with scenarios that either contained a single question or multiple related questions, we explored the relationship between question answering and document browsing.

4.1 Methods

Thirty-two graduate and undergraduate students were asked to participate in this experiment. All participants were between the ages of 20 and 40, and have strong backgrounds in computer science. Although all participants were experienced in searching for information (e.g., on the Web), none had any experience with question answering systems.

The experiment was divided into two parts: the first phase tested the effects of source reliability, and the second tested the effect of scenario size. Before starting the study, users were given a brief introduction to question answering systems. Prior to the first phase and following the second phase, users were asked to complete short surveys. The study concluded with an open-ended interview, in which participants were encouraged to share their general thoughts.

Since the purpose of this study was not to investigate the effectiveness of an actual question answering system, but rather to isolate criteria for effective interfaces, our study worked with a system that could answer every one of the test questions with one hundred percent accuracy. Answers were taken from an electronic version of the WorldBook encyclopedia.

4.1.1 Surveys

After a short introduction to question answering systems, but before starting the first phase, users were asked some general questions about their impressions of the importance of question answering for various tasks (on a five point Likert scale ranging from one, "not very important", to five, "very important"): replacing or augmenting normal Web search engines; accessing personal documents, email, etc.; interacting with the operating system and applications (e.g., a natural-language command interface); researching factual information (e.g., writing a report and finding facts); and troubleshooting (e.g., finding out what's wrong with the computer). In addition, we asked users to rate the importance of source reliability and additional context with respect to overall satisfaction (on the same five point scale). The purpose of this survey was to elicit users' preconceived notions of question answering systems.

After the study (but before the concluding interview), the users were given the same survey, with the addition of a few more questions. In addition to the questions used in the first survey, we asked the users which interface condition they preferred overall. We also asked about the importance of some other factors affecting question answering: system speed, multimedia responses, and presentation of alternative answers. The goal of the exit survey was to determine if users' views on question answering had changed.

4.1.2 Source Reliability

This phase of the study implemented a click-through experiment to determine how much context a user needed in order to accept or reject an answer, depending on the perceived trustworthiness of the source document. Eighteen questions[2] (see Figure 2 for examples) were presented to the user, randomly associated with one of three trust conditions:

- **Trusted:** the answer was obtained from a neutral, generally reputable source, e.g., an encyclopedia.
- **Biased:** the answer was obtained from a source known to be biased in its viewpoints, e.g., the advocacy site of a particular special interest group.

[2]Relatively obscure questions were purposely chosen to reduce the chance that a user would know the answer directly.

Examples of questions used in Phase I:

1. When did the 6-day war begin?
2. Who was the first person to reach the south pole?
3. When was the Rosenberg trial?
4. Where is Devil's Tower?
5. What is the active ingredient in Tylenol?

Examples of questions used in Phase II:

Scenario 1

When was the battle of Shiloh?
What state was the battle of Shiloh in?
Who won the battle of Shiloh?

Scenario 2

Who won the Nobel Peace Prize in 1992?

Scenario 3

What is Marilyn Monroe's real name?
When was Marilyn Monroe born?
Where was Marilyn Monroe born?
When did Marilyn Monroe die?

Scenario 4

What is the capital of Burkina Faso?

Figure 2: Sample questions used in the user study.

- **Unknown:** the answer was obtained from a source whose authority had not been established, e.g., a personal homepage.

Because the focus of this study was the *perceived* reliability of the source and not the actual source itself, the source citation was not given. Instead, each answer source was labeled with one of the trust conditions described above, e.g.,

> **Question:** Where did the Ukulele originate?
> **Answer:** Portugal
> **Source:** a trusted source.

Furthermore, the actual answer context did not change; only our labeling of it did.

At the start of each question, only the exact answer was presented (along with an indication of the source reliability). The user had four choices: to accept (believe) the answer as given and move onto the next question, to reject (not believe) the answer as given and move onto the next question, to request more information, or to request less information. If the user requested more information, the next interface condition was given, i.e., the first click on "more information" gave the answer-in-sentence interface condition, the second time gave the answer-in-paragraph interface condition, and the third time gave the answer-in-document interface condition. When the entire document was presented, the user had to either choose to accept or reject the answer. The "less information" option was basically an undo operation for reverting back to the previous interface condition.

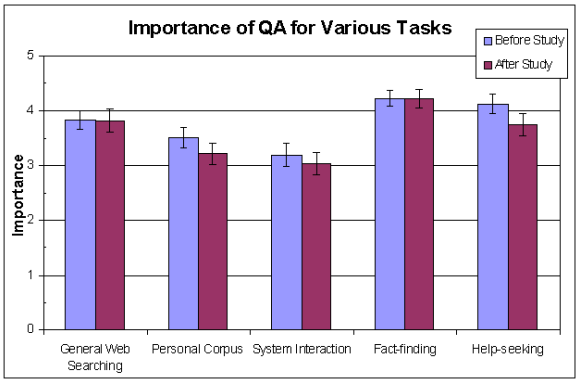

Figure 3: Importance of QA for various tasks ($\pm$ standard error about mean).

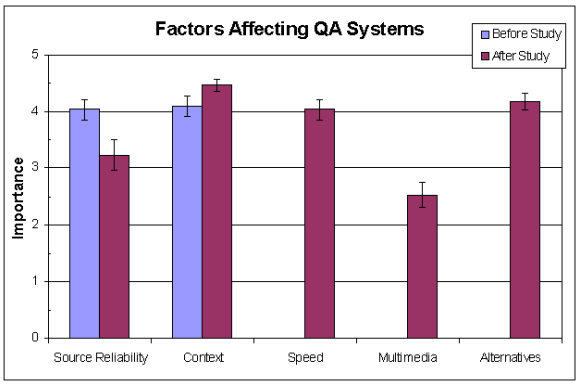

Figure 4: Importance of various factors affecting QA systems ($\pm$ standard error about mean).

For this phase of the user study, the major goal was to determine how much context the user needed in order to accept or reject an answer, i.e., how much of the source document did the user require to make a judgment regarding the validity of the system response.

4.1.3 Scenario Size

In the second phase of the study, participants were asked to directly interact with our sample question answering system. The goal was to complete a series of "scenarios" as quickly as possible. A scenario consisted of either a single question or a set of closely-related questions on the same topic (see Figure 2 for examples). In this phase of the user study, a total of eight scenarios were used: four with a single question, two with three questions, one with four questions, and one with five questions. Each scenario was randomly associated with a fixed interface condition (unlike the

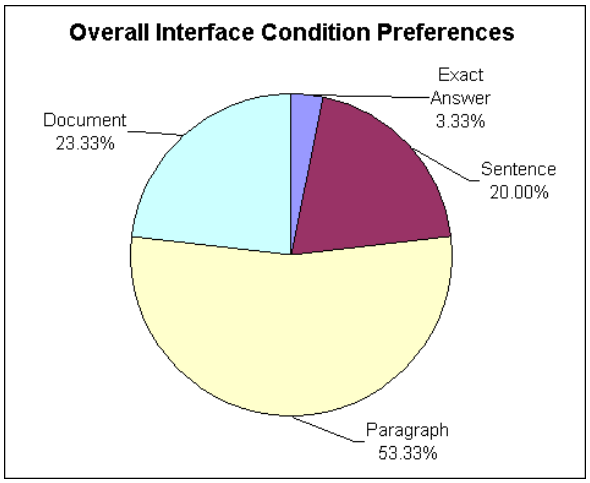

Figure 5: Users' overall preference of the various interface conditions.

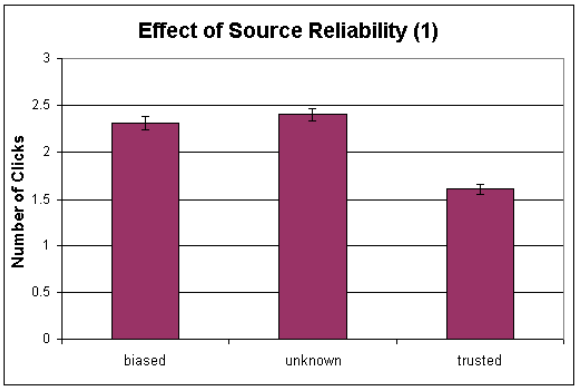

Figure 6: Effect of source reliability on the amount of context required to judge an answer ($\pm$ standard error about mean).

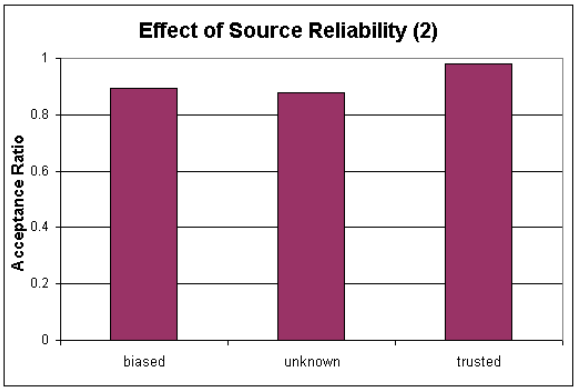

Figure 7: Percentage of answers that were accepted, under various conditions.

previous phase, users could not request more context). A scenario was considered complete when users had entered an answer for every question (in a text box beneath each question) and clicked the "Next" button.

The goal of this phase was to measure the time and the number of queries required to complete each scenario. Users were told that they could interact with the question answering system in any way that they wanted, e.g., by typing as many questions as necessary, by reading as much contextual information as desired.

4.2 Results

Users' opinions on the importance of question answering for various tasks are shown in Figure 3. Of the choices presented, they believed that question answering was most important for fact-finding/research and least important for interacting with a system or application. T-tests showed that our study did not alter users' opinions in a statistically significant way.

The users' views on factors that would affect their overall satisfaction with a question answering system are shown in Figure 4. Initially, they believed that source reliability and context (how much text is in the response) were equally important. However, opinions changed dramatically after the user study. The importance of source reliability showed a statistically significant drop, $t(28) = 3.72$, $p < 0.01$. Conversely, the importance of context rose a statistically significant amount, $t(28) = -2.57$, $p < 0.05$.

Figure 5 shows users' overall interface condition preferences. We discovered that users liked the answer-in-paragraph interface condition the best and the exact answer interface condition the least. The majority of users remarked that paragraphs formed a "good size chunk of information"; the exact answer was too little, and the entire document was often too much. They also noted that "the sentence doesn't give you much over just the exact answer," i.e., displaying the sentence containing the answer often does not provide the user any useful amount of additional information. For example, a sentence answering a question about a particular person's birthday may simply be "he was born on March 14, 1879." In particular, pronouns posed a big problem, since sentences with pronouns taken out of context often cannot be meaningfully interpreted. However, coreference resolution technology could be integrated into QA systems to address this issue.

4.2.1 Source Reliability

The effect that source reliability had on the amount of context required to judge an answer is shown in Figure 6 and Figure 8. The bar graph shows the average number of times the user clicked on "More Information" before he or she made a judgment to either ac-

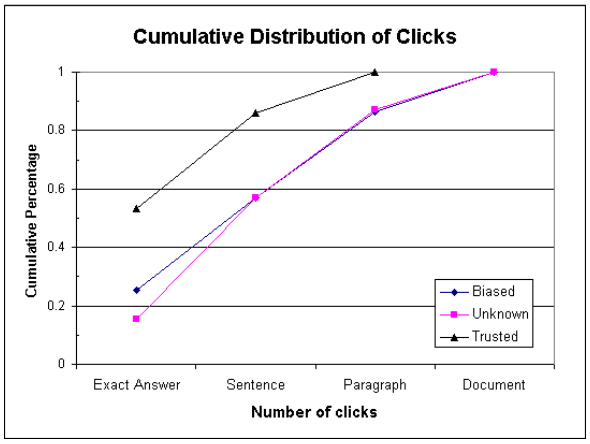

Figure 8: Cumulative percentage of clicks made before making a judgment.

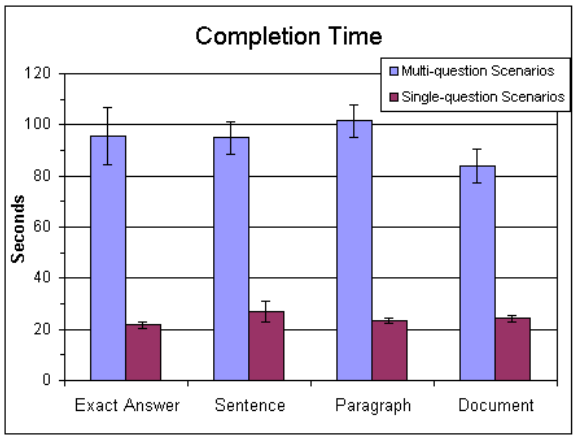

Figure 9: Completion time for scenarios ($\pm$ standard error about mean).

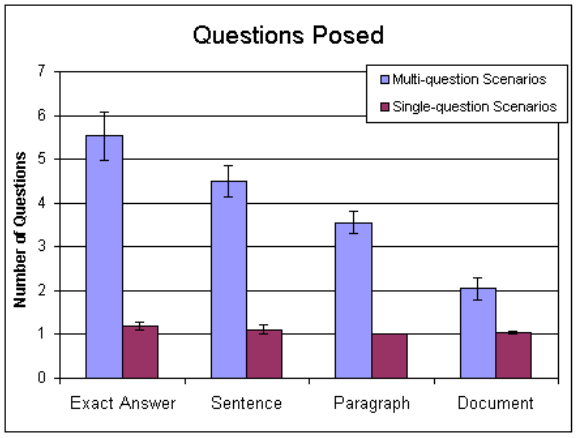

Figure 10: Number of questions posed for each scenario ($\pm$ standard error about mean).

cept or reject the answer; the line graph shows the cumulative distribution. For both trusted and unknown sources, users needed at least a paragraph, on average, to form a judgment on the answer; for trusted sources, users needed less than a paragraph. ANOVA revealed that the overall difference in clicks was statistically significant, $F(2, 555) = 45.4$, $p \ll 0.01$, but that the difference between biased and unknown conditions was not, $t(370) = -0.927$, *ns*.

The users' final judgments of the answers are shown in Figure 7. For trusted sources, users accepted nearly 98% of the answers, while this percentage was near ninety percent for the other two conditions. ANOVA revealed that the overall differences were statistically significant, $F(2, 555) = 7.42$, $p \ll 0.01$, but the tiny difference between the biased and unknown conditions was not, $t(370) = 0.48$, *ns*.

Our interviews confirmed that source reliability was less important than the users had initially thought; this was reflected in surveys. Users were surprised that they "didn't care all that much" about what source the answer came from. They were compelled to read at least some portion of the text before making a judgment, regardless of source reliability. The instructions given for this phase clearly stated that although the source reliability varied, the question answering system itself was generally reliable, i.e., one could count on the system to correctly extract whatever was in the document. Nevertheless, some users remarked that they just "didn't trust the computer" and "want[ed] to at least check it [the source] out."

4.2.2 Scenario Size

Results from this phase were grouped into single-question scenarios and multi-question scenarios; completion times are shown in Figure 9, and the number of questions posed by the users is shown in Figure 10. For multi-question scenarios, the answer-in-document interface condition resulted in a lower average completion time; however, this difference was not statistically significant, $F(3, 108) = 0.863$, *ns*. The small variations in completion time for single-question scenarios were not statistically significant and proved to be a good control.

Although for multi-question scenarios, different interface conditions did not have a statistically significant impact on completion time, the effect on the number of questions needed to complete each scenario was very significant, $F(3, 108) = 15.45$, $p \ll 0.01$. With the answer in document interface condition, users asked, on average, less than half as many questions as they did with the exact answer interface condition. As ex-

pected, ANOVA did not reveal statistical significance in the slight difference between the number of questions asked in single-question scenarios,[3] demonstrating the validity of our control.

5 Discussion

The setup of the source reliability experiment was a compromise between investigating complex dependent variables and maintaining the feasibility of the study. In retrospect, we may have made some oversimplifying assumptions. Our original intent was to abstract away the subtle judgments in reliability and simply assert these judgments for the user, i.e., "pretend that this answer came from a source that you considered trusted, biased, or unknown." To our knowledge, users understood that sources were not maliciously disseminating false information. Unfortunately, some users interpreted the source reliability as an external judgment made by a computer system, despite our clarifications in the instructions. This was echoed in comments such as, "I don't trust the computer saying that the source is trusted." Users suggested that they be given the actual citations so that they could evaluate source reliability themselves. However, this would have added an additional layer of complexity to the study. Nevertheless, we believe that these considerations do not negate our results; simply relabeling sources differently produced a statistically significant effect on our user population. Further work is necessary in order to sort out the complex factors at play here.

Although different interface conditions had no significant impact on the completion time of multi-question scenarios, users required fewer interactions to complete the same task, i.e., fewer questions. We believe that this is a significant result, and it highlights the role that context plays in question answering. If users are given additional surrounding text, they will indeed read it. Context helps users to confirm the answer and to respond to additional related questions.

A potential objection to the validity of our results is that we, in effect, conducted our study using a "canned system" that always returned the correct answer. However, state-of-the-art question answering is not very far from being able to achieve just that; currently, the best systems can successfully answer over eighty percent of the types of factoid questions studied here (Voorhees, 2002). Our focus is on studying QA interfaces, with the goal of providing longer-term guidance on the development of future systems.

We hope that our results will be useful in the design of future evaluations. The trend in the TREC QA tracks is towards returning more and more exact answers: In TREC-8 and TREC-9, participants could either return 250-byte paragraph-length or 50-byte sentence-length answers. In TREC-2001, answers were restricted to 50 bytes. In TREC-2002, only exact answers were accepted. While forcing question answering systems to return exact answers might be the correct technological push, i.e., exact answers force systems to develop more sophisticated natural language processing techniques, our studies show that users prefer paragraph-sized chunks over the exact answer only. Although identifying exact answers does not prevent systems from displaying more text as the final response, actual user preferences should still be kept in mind when deploying QA systems and designing evaluations.

Our study revealed several features that users considered important in question answering systems. An often requested feature was the ability of the question answering system to resolve pronouns and ellipses in questions, e.g., being able to follow up a question like "when was the Battle of Shiloh" with "and where was it?" Most current question answering systems treat each question independently and thus are unable to maintain a state-preserving, prolonged interaction.

In addition, we discovered that effective question answering systems must not only be able to extract short answers to specific questions, but must also respond to general questions in a meaningful way. When faced with a multi-question scenario, many users would first attempt to ask a general question, e.g., "who was Marilyn Monroe" in hopes of obtaining a general article about her.[4] Since our question answering system was not designed to handle such queries, the users received no response. When interviewed about why they asked general questions, users responded that they had "hoped to get all the important information all at once." In the case of Marilyn Monroe, they expected to get a biography of her, which might include her birthdate, real name, films she starred in, etc. The ability to answer general questions is an aspect of question answering that warrants future research.

6 Conclusions

In many ways, question answering represents the next step in information access technology. By promising to deliver *answers*, not just *documents*, question answering systems can more effectively fulfill users' information needs. However, as a relatively new field,

[3] The number is not exactly one because people made typing mistakes, experimented with the system, etc.

[4] On average, approximately half of one question could be classified as one of these general queries. These queries do not qualitatively affect our results, since users were equally likely to ask general questions with each different interface condition.

question answering research has focused primarily on the underlying technology instead of computer-human interaction issues. Recent advances in answer extraction technology should be followed up with similar advances in interface design. Our research has revealed an interesting schism between the technological drive and actual user preferences. Although question answering systems are evolving towards providing exact answers *only*, our studies have shown that users actually prefer paragraph-level chunks of text (with appropriate answer highlighting). In order to design effective question answering systems in the future, we believe that user considerations should be treated on an equal footing with the underlying technology.

7 Acknowledgements

This paper is an extended version of an interactive poster published in CHI 2003 with the title "The Role of Context in Question Answering Systems." This work was supported by DARPA under contract number F30602-00-1-0545; additional funding was provided by the MIT-NTT collaboration, the MIT Oxygen project, a Packard Foundation fellowship, and IBM. We wish thank Mark Ackerman, Susan Dumais, and Greg Marton for helpful comments on earlier drafts of this paper and all the users who participated in our study.

8 References

Nicholas J. Belkin, Amy Keller, Diane Kelly, Jose Perez Carballo, C. Sikora, and Ying Sun. 2000. Support for question-answering in interactive information retrieval: Rutgers' TREC-9 interactive track experience. In *Proceedings of the Ninth Text REtrieval Conference (TREC-9)*.

Eric Brill, Jimmy Lin, Michele Banko, Susan Dumais, and Andrew Ng. 2001. Data-intensive question answering. In *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)*.

Susan Dumais, Edward Cutrell, and Hao Chen. 2001. Optimizing search by showing results in context. In *Proceedings of SIGCHI 2001 Conference on Human Factors in Computing Systems (CHI 2001)*.

Stephan Green, Gary Marchionini, Catherine Plaisant, and Ben Shneiderman. 1997. Previews and overviews in digital libraries: Designing surrogates to support visual information seeking. Technical Report CS-TR-3838, Department of Computer Science, University of Maryland.

Marti Hearst. 1999. User interfaces and visualization. In Ricardo Baeza-Yates and Berthier Ribeiro-Neto, editors, *Modern Information Retrieval*. Addison-Wesley Longman Publishing Company.

William Hersh, Andrew Turpin, Susan Price, Dale Kraemer, Benjamin Chan, Lynetta Sacherek, and Daniel Olson. 1999. Do batch and user evaluations give the same results? An analysis from the TREC-8 Interactive Track. In *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*.

Eduard Hovy, Ulf Hermjakob, Chin-Yew Lin, and Deepak Ravichandran. 2002. Using knowledge to facilitate factoid answer pinpointing. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002)*.

Boris Katz, Jimmy Lin, and Sue Felshin. 2002. The START multimedia information system: Current technology and future directions. In *Proceedings of the International Workshop on Multimedia Information Systems (MIS 2002)*.

Boris Katz. 1997. Annotating the World Wide Web using natural language. In *Proceedings of the 5th RIAO Conference on Computer Assisted Information Searching on the Internet (RIAO '97)*.

Ying K. Leung and Mark D. Apperley. 1994. A review and taxonomy of distortion-oriented presentation techniques. *ACM Transactions on Computer-Human Interaction*, 1(2):126–160.

Jimmy Lin, Aaron Fernandes, Boris Katz, Gregory Marton, and Stefanie Tellex. 2002. Extracting answers from the Web using knowledge annotation and knowledge mining techniques. In *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*.

Dan Moldovan, Marius Paşca, Sanda Harabagiu, and Mihai Surdeanu. 2002. Performance issues and error analysis in an open-domain question answering system. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)*.

Joonah Park and Jinwoo Kim. 2000. Effects of contextual navigation aids on browsing diverse Web systems. In *Proceedings of SIGCHI 2000 Conference on Human Factors in Computing Systems (CHI 2000)*.

Ellen M. Voorhees. 2001. Overview of the TREC 2001 question answering track. In *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)*.

Ellen M. Voorhees. 2002. Overview of the TREC 2002 question answering track. In *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*.

Ryen W. White, Ian Ruthven, and Joemon M. Jose. 2002. Finding relevant documents using top ranking sentences: An evaluation of two alternative schemes. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-2002)*.