In Proceedings of the 2003 Conference on Human Factors in Computing Systems (CHI 2003), April 2003, Fort Lauderdale, Florida

The Role of Context in Question Answering Systems

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

MIT AI Laboratory/LCS, 200 Technology Square, Cambridge, MA 02139 USA

{jimmylin,dquan,vineet,karunb,dfhuynh,boris,karger}@ai.mit.edu

ABSTRACT

Despite recent advances in natural language question answering technology, the problem of designing effective user interfaces has been largely unexplored. We conducted a user study to investigate the problem and discovered that overall, users prefer a paragraph-sized chunk 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 users research 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 user interfaces.

Keywords

Question answering, information retrieval, user interface, natural language

INTRODUCTION

Question answering has become an important and widely-researched technique for information access because it can provide users with exactly the information that they need instead of flooding them with documents that they must wade through. Current state-of-the-art systems can answer nearly 85% of factoid questions such as "what Spanish explorer discovered the Mississippi" in an unrestricted domain [1]. Despite significant improvements in the underlying technology, the problem of designing effective interfaces has largely been unexplored.

We believe that the most natural response presentation style for question answering systems is *focus-plus-context* [2], which is closely related to the *overview-plus-detail* [3] presentation style. Since most current question answering systems extract answers from textual documents, the text surrounding the answer serves as a natural source of *context*—a key concept of the presentation styles noted above. We performed a user study to explore the question of how much text a question answering system should return to the user, as well as the effects of source reliability (trustworthiness of the source from which the text was extracted) and scenario size (i.e., whether the user was asking a single question or a set of related questions).

RELATED WORK

The use of context in information retrieval systems has been extensively studied. Recently, the effectiveness of spatial and temporal contextual clues [4], category labels [5], and top-ranking related sentences [6] has been explored empirically through user studies in a Web environment. Furthermore, the Interactive Track at TREC has generated interest in interface issues associated with information retrieval systems. [7] compared a single-document and multi-document view of IR results for a question answering task. The study was inconclusive but hinted that presenting one document at a time was just as effective as displaying multiple documents, and required less cognitive effort.

However, it is unclear whether these results can be directly applied to question answering systems, as the role of context in question answering systems is not to support browsing, but rather to justify the answer and to offer related information. To our knowledge, no studies regarding the effects of context have been conducted specifically on question answering systems.

INTERFACE CONDITIONS

Under the focus-plus-context framework and taking into account natural language discourse principles, we developed four different interface conditions. The context was simply the text surrounding the answer, which varied in length for different interface conditions.

- *Exact Answer*. Only the *exact* answer is returned. For example, 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.
- *Answer-in-Sentence*. The exact answer is returned with the sentence from which the answer was extracted.
- *Answer-in-Paragraph*. The exact answer is returned with the paragraph from which the answer was extracted; the sentence containing the answer is highlighted.
- *Answer-in-Document*. The exact answer is returned with the full document from which the answer was extracted; the sentence containing the answer is highlighted.

EXPERIMENTAL METHOD

Thirty-two graduate and undergraduate computer science students (20–40 years old) were asked to participate in our computer-based experiment. Although all participants were experienced in searching for information (e.g., on the Web), none had any experience with question answering systems. 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 100% accuracy. Answers were taken from an electronic version of the WorldBook encyclopedia.

CHI 2003, April 5–10, 2003, Ft. Lauderdale, Florida, USA.
ACM 1-58113-630-7/03/0004.

Source Reliability

The first 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 relatively obscure question/answer pairs were presented to the user, randomly labeled as either 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), or unknown (the answer was obtained from a source whose authority had not been established, e.g., a personal homepage).

The major goal of this phase 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 the user required to make a judgment regarding the validity of the system response. Because the focus of this phase 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. Furthermore, the actual answer context did not change; only our labeling of it did.

At the start of each trial, 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 gave the answer-in-paragraph interface condition, and the third gave the answer-in-document interface condition. When the entire document was presented, the user had to choose either to accept or reject.

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. 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 with the previous phase, users could not request more context.) A scenario was considered complete when the user had entered an answer for every 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, i.e., by typing as many questions as necessary, by reading as much or as little contextual information as desired, etc.

RESULTS

Surveys showed that users liked the answer-in-paragraph interface condition best (a "good size chunk of information") and the exact answer interface condition the least. They also noted that "the sentence doesn't give you much over just the exact answer." 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 question answering systems to address this issue.

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 analysis revealed that the overall difference in the number of clicks was statistically significant, $F(2,555)=45.4$, $p\approx0.01$, but the difference between biased and unknown conditions was not, $t(370)=-0.927$, *ns*.

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 for our experiments. 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, $F(3,108)=15.45$, $p\approx0.01$. With the document interface condition, users asked less than half as many questions on average as with the exact answer interface condition.

ACKNOWLEDGMENTS

This work was supported by DARPA under contract number F30602-00-1-0545; additional funding was provided by the MIT-NTT collaboration, MIT Project Oxygen, a Packard Foundation fellowship, and IBM. We wish to thank the participants of our study for their time and to thank Mark Ackerman and Greg Marton for their comments.

REFERENCES

1. Moldovan, D., Harabagiu, S., Girju, R., Morarescu, P., Lacatusu, F., et al. Tools for question answering. In *Proceedings of TREC 2002*.
2. Leung, Y. and Apperley, M. A review and taxonomy of distortion-oriented presentation techniques. *ACM Transactions on Computer-Human Interaction*, 1(2):126–160, 1994.
3. Green, S., Marchionini, G., Plaisant, C., and Shneiderman, B. 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, 1997.
4. Park, J. and Kim, J. Effects of contextual navigation aids on browsing diverse web systems. In *Proceedings of CHI 2000*.
5. Dumais, S., Cutrell, E., and Chen, H. Optimizing search by showing results in context. In *Proceedings of CHI 2001*.
6. White, R., Ruthven, I., and Jose, J. Finding relevant documents using top ranking sentences: An evaluation of two alternative schemes. In *Proceedings of SIGIR 2002*.
7. Belkin, N., Keller, A., Kelly, D., Carballo, J., Sikora, C., and Sun, Y. Support for question-answering in interactive information retrieval: Rutgers' TREC-9 interactive track experience. In *Proceedings of TREC-9*, 2000.