Designing a Realistic Evaluation of an End-to-end Interactive Question Answering System

Nina Wacholder*, Sharon Small#, Bing Bai*, Diane Kelly, Robert Rittman*, Sean Ryan#, Robert Salkin*, Peng Song*, Ying Sun*, Liu Ting#, Paul Kantor*, and Tomek Strzalkowski#**

*Rutgers University #SUNY Albany ** University of North Carolina

Corresponding author: nina@scils.rutgers.edu

Abstract

We report on the development of material for an evaluation exercise designed to assess the overall design and usability of HITIQA, an interactive question-answering system for preparing broad ranging reports on complex issues. The two basic objectives of the evaluation were (1) To perform a realistic assessment of the usefulness and usability of HITIQA as an end-to-end system, from the information seeker's initial questions to completion of a draft report; and (2) To develop metrics to compare the answers obtained by different analysts and evaluate the quality of the support that HITIQA provides. We used qualitative and quantitative tools to obtain data about analyst's comfort with the HITIQA system, especially its novel features such as the ability to answer complex questions and the interactive dialogue. Because of the impracticality of measuring the quality of HITIQA output with the standard metrics of precision and recall, we developed a new task –cross-evaluation--to indirectly measure the quality of the answers obtained using HITIQA; in this black-box assessment, analysts rate the quality of their own and their colleagues' reports.

Overview

We report on the development of material for an evaluation exercise designed to assess the overall design and usability of HITIQA, an interactive question-answering system for preparing broad ranging reports on complex issues. The HITIQA Project is part of the ARDA AQUAINT program that aims to make significant advances in automated question-answering. The objective of HITIQA is to answer analytical, non-factoid questions such as "How is the al Qaeda organization funded?"; the answers to such questions are multi-dimensional, and typically can only be found by bringing together information from multiple sources (Small et al. 2003). HITIQA responds to a user's question either by providing text that answers the question or by engaging the user in an interactive dialogue whose goal is to clarify the types of information that are of interest to the user. Evaluation of the quality of answers of these complex questions is intrinsically difficult, even compared, for example, to the already difficult task of evaluating factoid questions (e.g., Breck et al. 2000; Sparck Jones 2001; Voorhees 2001).

To assess the results of two years of development and to develop metrics to guide future evaluation, we invited the intended users -- intelligence analysts employed by the US government -- to participate in two three-day workshops, held in September and October 2003. The two basic objectives of the workshops were:

1. To perform a realistic assessment of the usefulness and usability of HITIQA as an end-to-end system, from the information seeker's initial questions to completion of a draft report. In particular, we wanted to determine how long it took users to feel confident of their ability to use HITIQA and whether they were comfortable with the interactive dialogues and visual display panel.
2. To develop metrics to compare the answers obtained by different analysts and evaluate the quality of the support that HITIQA provides.

Each of these objectives entails a particular challenge. Performing a realistic assessment of HITIQA is difficult because many of the resources that the analysts use, as well as the reports they produce, are classified and therefore inaccessible to researchers. Assessing the quality of the support that the system provides is tricky because analytical questions rarely have a single right answer. It is not obvious how to define, for example, the precision of the system. We conducted an evaluation that included both qualitative and quantitative tools to assess the usefulness and usability of the system. Because of the impracticality of measuring the quality of HITIQA output with the standard metrics of precision and recall, we also designed a new task, cross-evaluation, to indirectly measure the quality of the answers obtained by the analysts by having them assess each others' reports (Sun et al., under review).

Workshop Task

The analysts' primary task was preparation of reports in response to "scenarios" – complex questions that may entail multiple subquestions. The scenarios were developed in conjunction with several U.S. government offices. One, for example, asked for a report on aspects of al Qaeda such as membership, sources of funding and activities. Another asked for information on the chemical weapon Sarin (Figure 1).

Figure 1: Text of Sarin scenario

The Department of Homeland Security has requested a complete report on the chemical weapon, sarin. This report is due in 5 hours. In your report, include its potency and potential impact on a community, what countries and organizations have been involved in producing it, where these locations are, the production method and how it has developed, who possesses it now, who distributed it (if through trade, what was traded for it?), potential means of use, how can this be integrated into warheads, any known defenses against it, and who is at the greatest threat. Provide any other information that you see relevant.

The analysts' task was to ”prepare a report like what you would do in your normal work environment”

To obtain an adequate supply of text to support extensive questions and compensate for the absence of classified data, we created a new corpus. Taking as a starting point data from the Center for Non-Proliferation Studies (CNS) collected for the AQUAINT Program, we used Google to mine the web for similar subject matter. The final corpus was about 1G; this proved to be sufficient to support use of HITIQA to 'solve' each of the scenarios.

To obtain valid results, it was important to be sure that the analysts had a good comprehension of how to use HITIQA and of the fundamental ways in which it differs from search engines; it provides answers to specific questions and supports interactive dialogue. The entire first day of the first workshop was devoted to training. Analysts completed a two-part proficiency exam designed to demonstrate their competence and identify areas of confusion. After this, the 'real' evaluation proceeded.

Workshop Tasks

Analysts participated in a battery of tasks designed to provide a realistic assessment of the usefulness and usability of HITIQA as an end-to-end system, from the information seeker's initial questions to completion of a draft report. These include:

- **Session questionnaire:** A set of 16 questions, completed after each scenario, in which the analyst assessed the reality of the scenario, their comfort with the system, and their level of satisfaction with the results. The questionnaire is in Appendix A.
- **Final questionnaire:** A set of 17 questions, completed at the end of each workshop, in which analysts assessed various aspects of the system such as the interactive dialogue and visual interface. Analysts also assessed HITIQA's usefulness in finding information and its readiness for use in their regular work environment. The questionnaire is in Appendix B.
- **Individual interviews:** At the end of the first workshop, researchers with considerable experience at such tasks interviewed each analyst to elicit feedback about the HITIQA interface.
- **Group discussions:** A series of open-ended group discussions (at least one every day of the workshop) in which analysts reported their reactions to using HITIQA and their assessment of its strengths and weaknesses.
- **Cross-evaluation:** A 'black-box' method of evaluation obtained by having analysts rate the quality of their own and their colleagues' reports. If the use of one system produces higher quality reports than the use of another, the first system can be said to be better. (Sun et al. under submission).

Results

In this section, we summarize some of the most important results of these assessment tasks. More details are reported in [Wacholder et al. 2003a], [Wacholder et al. 2003b] and Sun et al. [under review].

Reality of the scenario and task

One of our primary concerns was to design tasks that are similar in scope and difficulty to those that the analysts are used to performing at work and to be sure that they feel comfortable using the system. Five questions in the session evaluation dealt with this issue; for example, one question asks how the scenarios compare in difficulty with the tasks the analysts normally perform at work. The mean score for these five questions was 3.75 on a 5 point scale (five is the best score). The question yielding the lowest score (M=2.88) was “How did the scenario compare in difficulty to tasks that you normally perform at work?”. this slightly above average rating of difficulty of the tasks was quite satisfactory for our purposes. We therefore conclude that the task was realistic and that the results of the evaluation are meaningful for the intended use of HITIQA.

Usability and usefulness of HITIQA

The final evaluation, the individual interviews and the informal discussions were designed to elicit quantitative and qualitative evaluations of HITIQA. In the final evaluation, analysts were asked to rate their agreement with statements such as “ HITIQA is hard to use”; “In general, I like the HITIQA interactive dialogue”; and “Having HITIQA at work would help me find information that I can't currently find”. The mean normalized score for final evaluation was 3.74 on a 5 point scale for Workshop I; this means that the system received many more ratings of 4 and 5 than of 1 and 2. The session evaluation table in Appendix A shows that the users' comfort with the system declined slightly between the two workshops. We have tentatively traced this to the effect of one individual's bad experience (including a persistent misspelling and technical problems) at Workshop II. The scores of the other two analysts went up between Workshops I and II.

Comments made by the analysts in the group discussion and in the individual interviews confirmed that analysts liked the interactive dialogue and were very pleased with the results. For example, one analyst said “I learned more about Sarin gas in 30 minutes than I probably would have at work in a half a day.” As desired, the analysts also made many suggestions for improving the interface and the interoperation of the visual and text display. For a research system undergoing its first rigorous evaluation, these results are very satisfactory – they support the value of the design of the HITIQA system, including the interactive mode and the visual display and encourage us to move forward with this approach.

Table 1: Results of cross-evaluation of reports

Criteria	**Means Workshop I**	**Means Workshop II**	**Differences**	Significance
Covers the important ground	3.17	4.05	.88	.01**
Avoids irrelevant materials	2.67	3.62	.95	.04*
Is well organized	2.58	3.86	1.27	.00**
Reads clearly and easily	2.75	3.76	1.01	.02*
Overall rating	2.83	3.90	1.07	.01**

** Statistically significant at .99 level. *Statistically significant at .95 level.

Quality of reports

Because HITIQA supports the finding of answers to analytical questions in a highly interactive fashion, it is impractical to assess the quality of output by standard measures of precision and recall. We therefore introduced the cross-evaluation task, in which analysts evaluate the product that they and their colleagues have produced – the final report. **Table 1** shows the results of the cross-evaluation of reports produced at both workshops. Although we have no earlier results to compare, we nevertheless are able to look at the change in quality of reports over the two workshops.

We see that the mean scores of the reports prepared at Workshop II are better than those prepared at Workshop I. A one-way analysis of variance (one-way ANOVA) was conducted to determine whether the quality of the Workshop I reports is significantly different from those of Workshop II. The last column in Table 1 shows that the difference between Workshop I and Workshop II was significant for all five criteria. Informal discussion with the analysts confirmed that they felt more satisfied with the reports prepared at Workshop II than at Workshop I. Some of this improvement can be attributed to increased familiarity with HITIQA and to improvement of the system, but the analysts themselves felt that the very fact that they knew that their reports would be 'graded' by their peers caused them to concentrate more on producing high quality reports. The fact that the results of the cross-evaluation went up for reports produced at Workshop II suggests that the bad experience of one analyst did not interfere with the overall quality of the reports. While we did not have two well-defined systems to compare (because of the difference in experience and training), conceptually we can think of the two workshops as two different systems These results suggest that cross-evaluation is a promising method for evaluation of information access systems and can readily be used for cross-system comparison.

Conclusion

We conclude that our efforts to conduct a realistic assessment of HITIQA and to develop a measure for evaluating the results of a complex question-answering system succeeded. We plan to conduct additional evaluations and to develop additional metrics for efficiently assessing report quality.

Acknowledgements

This paper is based on work supported by the Advanced Research and Development Activity (ARDA)'s Advanced Question Answering for Intelligence (AQUAINT) Program under contract 2002-H790400-000. Special thanks to the analysts and also to Heather McCallum-Bayliss and John Rogers for helping to arrange the workshops.

References

Breck, Eric, John Burger, Lisa Ferro, Lynette Hirshman, David House, Marc Light and Inderjeet Manerji (2000) How to evaluate your question answering system every day ... and still get real work done. *Proceedings of LREC 2000*, Athens, Greece.

Jones, Karen Sparck (2001). Automatic language and information processing: rethinking evaluation. *Natural Language Engineering* 7(1):29-46.

Small, Sharon, Liu Ting, Nobuyuki Shimuzu and Tomek Strzalkowski (2003) HITIQA, An interactive question answering system: A preliminary report. *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering*. June, 2003.

Sun, Ying et al. (Under review) Cross Evaluation – a pilot application of a new evaluation mechanism. Submitted to ASIST 2004.

Wacholder, Nina et al. (2003) Evaluation of the HITIQA Analyst's Workshops. Report of ARDA AQUAINT 24 month meeting.

Voorhees, Ellen (2001) The TREC-9 Question Answering Track. *Natural Language Engineering*.

Appendix A: Session Evaluations

	Normalized Mean (1=bad; 5=good)	
Question	**Workshop I***	**Workshop II****
1	4.50	4.00
2	2.88	3.33
3	3.88	3.78
4	3.88	3.78
5	3.63	3.78
6	3.38	3.33
7	3.75	3.44
8	3.71	3.44
9	4.13	3.33
10	3.14	3.11
11	4.00	3.44
12	3.25	3.44
13a	4.13	3.56
13b	3.38	3.56
13c	4.13	3.56
13d	3.63	3.33
	3.71	3.51

*4 analysts **3 analysts

1. How realistic was the scenario? In other words, did it resemble tasks you could imagine performing at work?
2. How did the scenario compare in difficulty to tasks that you normally perform at work?
3. How confident were you of your ability to use HITIQA to accomplish the assigned task?
4. Given that you were performing this task outside of your standard work environment, without many of your standard resources, were you comfortable with the process of preparing your report?
5. Given that you were performing this task outside of your standard work environment, with access to a restricted set of documents, were you satisfied with the quality of the report/answers that you were able to find for this scenario?
6. In general, did the display of answers through the Answer Panel help you to navigate the answers in order to see what information was available?
7. In general, did the answers that the system provided make sense in relation to the questions that you asked?
8. In general, was it hard to formulate questions about this scenario that resulted in useful responses from the system?
9. In general, were the answers that the system provided helpful in meeting the goals set forth in the scenario?
10. In general, did the visual interface usefully represent the content of the answers that the system had found for you?
11. For this scenario, did the visual interface help you to find more precise answers than you would have found without it?
12. How would you assess the length of time that it took to perform this task?
13. If you had to perform a task like the one described in the scenario at work, do you think that having access to the HITIQA system would help…
    (a.) Improve your final report?
    (b.) Answer specific questions that you currently have trouble answering
    (c.) Increase the speed with which you find information?
    (d.) Find information

Appendix B: Final Evaluations

	Normalized Mean (1=bad; 5=good)	
Question	**Workshop I***	**Workshop II****
1	3.50	4.00
2	3.50	4.00
3	4.00	4.33
4	4.25	4.33
5	4.50	4.33
6	3.25	4.33
7	3.50	3.00
8	3.25	3.33
9	3.50	4.00
10	4.00	4.00
11	4.25	3.67
12	4.50	3.67
13	3.33	3.33
14	4.33	3.33
15	2.75	3.00
16	4.25	4.00
17	2.67	3.33
	3.73	3.76

*4 analysts **3 analysts

1. I feel that I have become pretty proficient at using the HITIQA system.
2. The training on the first day gave me the skills needed to use the system successfully.
3. The training materials are hard to understand.
4. The training materials contain most of the information I needed to learn to use HITIQA.
5. My skill at using HITIQA improved over the course of the workshop.
6. HITIQA is hard to use.
7. I couldn't find enough documents with relevant information.
8. In general, I like the HITIQA interactive dialogue.
9. In general, I like the HITIQA visual interface.
10. In general, I like using HITIQA.
11. HITIQA slows down my process of finding information.
12. HITIQA helps me find important information.
13. Having HITIQA at work would help me find information that I can't currently find.
14. Having HITIQA at work would help me find information faster than I can currently find it.
15. HITIQA is not ready yet to be used in my regular work environment.
16. HITIQA would be a useful addition to the tools that I already have at work.
17. HITIQA would let me stop using some of the tools that I currently use at work.