

REPORT OF THE 3rd AQUAINT DIALOGUE EXPERIMENT

Jean Scholtz and Emile Morse

National Institute of Standards and Technology

jean.scholtz@nist.gov; emile.morse@nist.gov

November, 2003

Executive Summary

The third AQUAINT dialogue experiment was an empirical study to generate baseline data to use for comparisons and to assess progress.

The baseline experiment was conducted using Naval reservists as the subjects and Google as the search engine. NIST worked with researchers from Albany to make the baseline comparable to the experiment that was conducted there this fall. This allowed us to have comparison data for their experiment.

The baseline analysis provides:

- A characterization of the queries analysts used
- A ranking of relevance of the documents returned in response to the queries posed
- A characterization of the reports generated by the analysts

This report summarizes the methodology and materials used in creating the baseline and describes the analysis.

NIST will distribute the baseline materials consisting of:

- A data set to use for the search engine or system
- Scenarios to establish the context of the Q&A work
- Questionnaires for user satisfaction
- An experimental protocol for the conducting the study
- A description of the analysis done by NIST
- Raw data (queries, Camtasia recordings, reports generated, documents returned, relevance rankings) so that researchers can augment the NIST analysis if needed

The NIST results are described in this paper as well. Results include measures that can be used for efficiency and effectiveness metrics.

Contact Jean Scholtz, jean.scholtz@nist.gov, 301-975-2520 or Emile Morse, emile.morse@nist.gov, 301-975-8239, for a copy of baseline materials.

Introduction

The first two dialogue experiments were successful in that teams received feedback about their systems. We also accumulated information about how analysts would interact with dialogue systems. A missing component of our earlier experiments was a comparison. We had input from researchers that this would be useful for publication of research papers. A baseline would also be useful for measuring impact of individual systems and overall progress in the field.

The goal of the third AQUAINT Dialogue experiment was to perform a baseline study that could lead to creation of a benchmark. The design of the baseline study addresses five basic elements of research design: 1. subjects, 2. tasks, 3. data, 4. system configuration including backend and user-interface, and 5. metrics. We will also describe how we used the NIMD Glass Box environment to collect detailed data.

This paper describes the components of the research design of the baseline study and the results of testing with users. While users were used in the testing process, this study is not simply a usability study. The hallmark metrics of usability studies are efficiency, effectiveness, and satisfaction. The metrics that were developed in this benchmark context have gone beyond these user-centered measurements. We believe that the variety of measurements will be attractive to system developers who choose to test their systems.

Methodology

The five basic components of any study design that involves human subjects are: subjects, tasks, data, computer system including the backend and the user interface, and the metrics and measurements. The sixth component that will be described is the NIMD Glass Box environment that was used to collect user interactions.

Subjects

We recruited 5 Navy reservists who perform intelligence analysis as their reserve duty. These people had from 2-20 years experience and their full-time employment ranged from bank teller to attorney to fireman. All were thoroughly professional and they worked in a highly efficient and focused manner.

Tasks

The tasks were developed by the Air Force Research Lab in Rome NY. The scenarios provide context by stating the customer for the report and the deadline. Two of the scenarios were 'short-term' in nature and were due in 5-8 hours. The other two were 'long-term' with due dates in 3-4 weeks.

Data

We used the corpus created by the Albany/Rutgers group [1]. They describe their method for developing the data as: "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. We initially retrieved approximately 3 GBytes; after removal of duplicates and filtering, the final corpus was about 1.2 GBytes." During the 2nd Dialogue Pilot study, we found that a few of the scenarios were not well-supported by the data. For this reason, the CNS collection was expanded upon to ensure that there would be enough material to support a 3-hour test period.

The data that we received was subjected to insertion of minimal HTML markup to allow reasonable viewing in a browser and filtering for remaining duplicates and for documents that had negligible content (<1500 bytes including markup). The final collection was ~135K documents (~1.14 GB). The Google appliance that we had access to indexed ~70K of this set before reaching its preset maximum.

System

The display used by the subjects had 4 major components. [include screenshot]

1. MS Word – used for creating reports by cut-paste and typing.
2. Internet Explorer – used to explore document collection indexed by Google on the NIST intranet.

3. Glass Box Control Panel – This software has been developed under the ARDA NIMD Program by Battelle. The underlying software captures mouse and keyboard events, cut/paste events, http requests (queries, results, other web pages), screen capture via Camtasia, etc). The subject used the Control Panel to begin and end recording of the session. All data is logged to a database which can be queried later to determine ‘interesting’ events.
4. Relevance Assessment Tool – We asked the subjects to rate each document that they viewed. We used a 4-point relevance scale (Key, Important, Marginal, and Irrelevant). We also allowed the subject to indicate whether the current document was a Duplicate. A second scale was used for drilling down on individual documents – we asked the subject to say whether the content of the current document was about ‘Background’, ‘People’, ‘Organizations’ and/or ‘Events’. The second scale was only active when the document judged to be Key, Important or Marginal. [Include screenshot]

Testing Outline

On each day when testing was performed, the subject was instructed to report to the VUG (Visualization and Usability Group) laboratory facility at NIST at 8:30 am. The overall purpose of the experiment was explained to him and instructions for using the 4 main areas of the display were given. The display is described below in the ‘System’ section. The subjects were told that the output of their work would be a Word document; they were instructed not to work to achieve a piece of finished intelligence but rather they should use an approach that felt comfortable to them and might be one or more of the following: an outline of their approach, a matrix that they might usually develop, an indication of their general strategy, snippets of relevant documents that they would like to refer to or include in a final report. The gist was to keep the subjects focused on analysis rather than on report writing. Subjects were encouraged to ask questions and when they were satisfied and comfortable, the first task scenario was presented and the session was started. Subjects were told that they should work until one of three situations occurred: 1. they felt that they had all the information that they would need to write a good report on the topic; 2. they felt that they were making little progress and that working more would not be useful; 3. the 3-hour time limit had occurred. At the end of working on the scenario, the subjects were asked to answer an on-line questionnaire about their experience.

After lunch, the same procedure was repeated with a different scenario. The scenarios used are included in Appendix A.

Metrics and Measures

The following measures were collected:

- Time spent on each scenario

- Query characterizations

- # queries/analysts/scenario

- # distinct queries

- # revised queries

- query length

- Relevant documents

- % queries producing x relevant documents

- Ranking of documents returned

- Depth factor for relevant documents

- Report Characteristics

- Segments of documents used in report

These measures can be combined into metrics of efficiency and effectiveness. For example,

Efficiency: # queries, % revisions, # relevant documents/ query, # relevant documents/ time

Effectiveness: # segments of documents used in report/# relevant documents scanned; average ranking of documents scanned

Baseline Pack

Baseline data consists of:

- Description of data collection (Readme.doc & this paper)
- Scenarios developed by AFRL
- Query Trails
- Reports generated by analysts during the baseline study
- List of document reviewed with analysts' relevance ratings
- List of passages that were cut and pasted, i.e. an indicator of implicit relevance
- Camtasia recording of analyst sessions
- A data set to use for the search engine or system
- Questionnaires for user satisfaction
- An experimental protocol for the conducting the study
- A description of the analysis done by NIST
- Raw data (queries, Camtasia recordings, reports generated, documents returned, relevance rankings) so that researchers can augment the NIST analysis if needed

Results

Session Length

Subjects were instructed to take up to 3 hours to work on their tasking. Table 1 shows the actual time spent by each analyst on each of the two scenarios that they worked on. Immediately after the subject completed the scenario, even before the survey questionnaire was filled out, the test administrator asked the subject why he was stopping. Three of the 10 sessions ran up against the 3 hours time limit. Five of the remaining 7 sessions ended because the analyst felt that they had sufficient information to write a report. The final 2 cases, marked by an asterisk in the Table, were terminated because the analyst felt that he had exhausted the system's ability to provide additional information.

Table 1: Time on task (hour:min)

Analyst ID	Scenario				Mean
	I1	I2	s1	s2	
navy2		2:17	2:06*		2:11
navy3	2:15			2:04	2:09
navy4	3:06		2:57		3:02
navy5		2:07		2:30	2:18
navy6	2:56		1:54*		2:25
Mean	2:46	2:12	2:19	2:17	

Query-related measures

Table 2 shows the number of queries generated by each analyst for each scenario. They ranged from a low of 3 to a high of 24. Interestingly, the number of queries was not related either to the particular analyst or to a particular scenario. The numbers seem to come from a bimodal distribution – 5 values between 21 and 24 and 5 between 3 and 12. We currently have no explanation for this odd distribution. Perhaps when we inspect the individual queries and query trails, we might find explanations. Spelling errors, non-productive searches and refining a query will all

be associated with an inflated count of queries. A second source for trying to explain this observation might be in the survey responses – perhaps an analyst’s prior knowledge might be related to the number of queries. For instance, if an analyst had a great deal of knowledge about hazardous materials, he might know exactly what to ask if given the ‘Sarin’ scenario, leading to low numbers of queries.

Table 2: Number of queries per scenario

Analyst ID	Scenario				Grand Total
	I1	I2	s1	s2	
navy2		24	3		27
navy3	11			21	32
navy4	24		12		36
navy5		21		11	32
navy6	6		22		28
Grand Total	41	45	37	32	155

The data in Table 3 shows the average number of words in queries. These results refute the conventional wisdom which says that people submit queries with 1 or 2 words. Preliminary analysis of the query trails (i.e., all queries for a single scenario session) reveals that analysts often spend time refining their queries by changing the order of terms and choosing synonyms. This phase of query refinement proceeds for up to 3 or 4 minutes and entails up to 5 revisions. During this time the analyst does not follow any links in a page of results. A second strategy that analysts use is to set the context of the query by prefacing each query with words that limit the domain. For instance, when Navy2 performed task L2, he prefaced each of his 24 queries with ‘south african wmd’.

Table 3: Average length of queries (words)

	I1	I2	s1	s2	Mean
navy2		4.75	4.67		4.71
navy3	3.00			2.81	2.91
navy4	2.67		3.92		3.30
navy5		3.57		2.64	3.11
navy6	1.83		4.41		3.12
Mean	2.50	4.16	4.33	2.73	

Table 4 shows a typical query trail. The subject generated 24 queries and inspected 102 documents. The blocks surrounding the sections of the table indicate boundaries between topics. The first block shows that the subject started with an overly broad query – al-qaida. She inspected the first results page and decided to revise her query. She repeated this process for the next few minutes until she hit on ‘al-qaida leaders’. At this point, she spent 45 minutes inspecting 20 documents from looking at 14 pages of results with 10 document summaries per page. The shaded lines indicate queries that produced significant time spent on inspecting the results.

Table 4: Query Trail from Navy4 on Scenario L1

Query	#docs	Time
al-qaida	0	00:16
al-qaida key figures	0	00:28
al-qaida leadership	0	00:25
al-qaida playing cards	0	00:20
al-qaida cards	0	00:08
al-qaida playing cards	1	01:10
al-qaida leaders	20	45:24
al-qaida organizations	17	20:03
+cdi al-qaida organizations	3	04:23
al-qaida guantanamo	0	01:02
al-qaida facilities	0	00:10
al-qaida +facilities	0	00:26
al-qaida +facilities -nuclear	13	25:59
al-qaida financial support	0	00:13
al-qaida financial resources	4	08:06
al-qaida "financial resources"	8	11:39
al-qaida "training programs"	2	02:44
terrorist training programs	0	00:13
terrorist training camps	2	01:54
terrorist training camps al-Qaida	0	00:15
+"terrorist training camps" al-Qaida	16	25:46
weapons al-Qaida	5	12:02
weapons al-Qaida - iraq	0	00:09
weapons al-Qaida -iraq	11	10:39
Total	102	2:53:55

A complete set of query trails can be found in the accompanying DVD.

Document Importance Measures

Subjects were asked to rate documents that they viewed. The relevance categories were defined as follows:

- Key document – Reference documents or such an important document that your report would be inferior if the information were not included or unique information that is not likely to be duplicated in another source.
- Important – Very useful information.
- Marginal – Relevant document but contains mainly peripheral information or information that is easily found in other sources.
- Not Relevant – Document has no bearing on the current topic.

Subjects were provided with hardcopy of these definitions to refer to during their sessions. The information was also available via a hyperlink in the Relevance Assessment window. Subjects expressed confidence in their ability to apply the categories to the documents being viewed.

Ten scenarios produced 604 viewed documents. Less than 20% of the documents viewed were judged by the subjects to be Key or Important and 59 were not rated. Analysis of the logs indicates that many of the documents that were not rated were used as a source of cut-and-pasted material; unrated documents were often viewed for a significant period of time. Both these facts lend support to the idea that the analysts simply forgot to rate the page.

The values for duplicates in the Table are the number of documents that the subject reported as duplicates. We have not yet analyzed the data to confirm that the documents were actually identical copies of previously seen

material, that is, now many duplicates were false positives. We also have not yet determined how many failures to detect duplicates there were. Both of these calculations are possible due to the detailed Glass Box logs and will be measured.

Table 5: Relevance ratings for each scenario

Scenario	Key	Important	Marginal	Irrelevant	unrated	duplicates	Total
l1	16	54	69	37	12	5	193
l2	5	8	22	53	15	28	131
s1	6	24	86	72	17	22	227
s2	4	11	9	11	15	3	53
Total	31	97	186	173	59	58	604

In addition, when documents were determined to be key, important or marginal, subjects were encouraged to check boxes in the Relevance Assessment panel to indicate whether the content of the document was related to Background, People, Organizations, and/or Events. The results of this analysis are not presented here but are available in the spreadsheet in the DVD of this report.

While the above Relevance Ratings are a direct measure of ‘interestingness’ of documents, there are other indirect ways on detecting whether documents are interesting. For example, documents that are printed, saved, bookmarked, read for a long time, or selected for inclusion in another document are likely to be more interesting than those that receive less attention. Our subjects did not have access to a printer and they did not save or bookmark documents. Saving and bookmarking are useful techniques if an analyst wants to use the documents at a later time, but our analysts had a single 3-hour block in which to perform their task. On the other hand, all the analysts made extensive use of cut-and-paste operations. As shown in Table 6 there were 372 instances of material being cut from web pages.

Table 6: Number of cut-and-paste operations

Analyst ID	l1	l2	s1	s2	Total
navy2		39	28		67
navy3	18			14	32
navy4	55		59		114
navy5		16		24	40
navy6	102		17		119
Total	175	55	104	38	372

Figure 1 shows the relationship between the numbers of cuts that were done vs. the relevance rating of the documents from which the cuts were made. As expected, cutting happened most often from documents that were judged key or important. A significantly lower fraction of marginal documents were used and only one irrelevant document was cut from. It is interesting to note the relatively high fraction of unrated documents that were the source of copied material. This observation lends support to the idea that the analysts simply forgot to rate the documents and, by inference, those document which were not rated but were cut from are likely to belong in the key or important category.

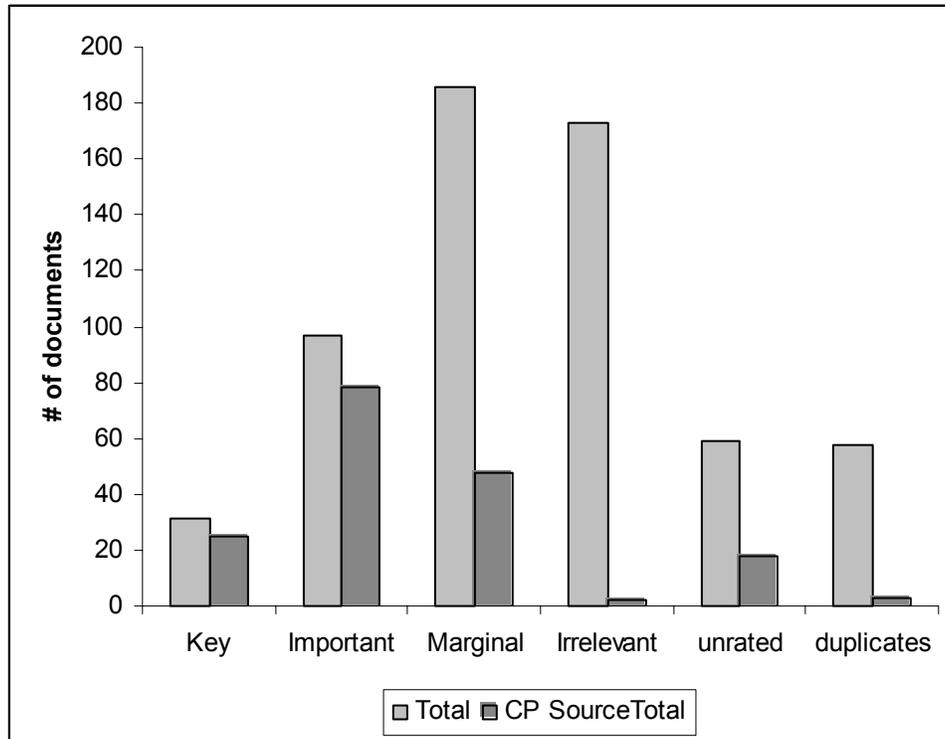


Figure 1: Distribution of Relevance Ratings and Copy-Paste Actions

Search Depth

Inspection of the Glass Box data showed that the analysts requested many pages of search results. Although conventional wisdom says that people only look at 1 or 2 results pages (i.e. 10-20 document summaries), the analysts in this study were very thorough in their explorations. Table 7 shows the frequency that each analyst requested deeper results pages. The column labels refer to the position in the ranked list of results. For instance, navy2 requested the fourth results page, containing documents 31-40, nine times.

Table 7: Number of times deeper results pages were requested

ID	Next n results										Total
	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100	101-200	
n2	10	10	9	8	6	5	2	2	3		55
n3	6	1	1							4	12
n4	16	10	6	6	5	5	4	4	4	26	86
n5	13	4	1	1	1						20
n6	14	8	6	5	5	4	2	2	2	9	57
Total	59	33	23	20	17	14	8	8	9	39	230

Of the five analysts, three explored past the first four pages of results numerous times. Interestingly, three of the analysts (even one who normally used only the first 20 results) looked at the last page of results frequently. Presumably this is to check to see what is at the edges of relevance to the query.

Survey Results

Subjects were asked to fill out a survey form after completing analysis of each scenario. The questions were selected from the set developed during the Albany Workshops. Of the original 16 questions, we eliminated two

questions that were specific to evaluation of visual interfaces, since the baseline interface was not intrinsically visual. Responses were provided on a 5-point scale. Table 5 shows a synopsis of the question followed by the definition of the end-points of the scale for the particular questions. Finally, the average response value is shown.

Table 5: Mean Values of 5-Point Likert Survey Responses

Q#	Question	1	5	Mean
1	How realistic was the scenario? In other words, did it resemble tasks you could imagine performing at work?	not realistic	realistic	4.50
2	How did the scenario compare in difficulty to tasks that you normally perform at work?	less difficult	more difficult	3.38
3	How confident were you of your ability to use the system to accomplish the assigned task?	less confident	more confident	3.75
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?	less comfortable	more comfortable	3.25
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?	not satisfied	very satisfied	3.38
6	In general, did the display of answers help you to navigate the answers in order to see what information was available?	not at all	very much	3.17
7	In general, did the answers that the system provided make sense in relation to the questions that you asked?	not at all	very much	4.38
8	In general, was it hard to formulate questions about this scenario that resulted in useful responses from the system?	easy	difficult	2.29
9	In general, were the answers that the system provided helpful in meeting the goals set forth in the scenario?	frustrating	helpful	3.88
12	How would you assess the length of time that it took to perform this task?	much longer	reasonable	4.00
13a	Improve your final report?	not at all	a lot	3.75
13b	Answer specific questions that you currently have trouble answering?	not at all	a lot	3.13
13c	Increase the speed with which you find information?	not at all	a lot	3.63
13d	Find information that you have trouble locating?	not at all	a lot	2.63

Average values that are particularly high (≤ 4.0) or low (> 2.5) include the answers to Questions 1, 7, 8, and 12. Question 1 asked about realism of the scenarios; analysts reported that the tasks were highly realistic. Question 8 asked about the ease of formulating questions based on the scenarios; the analysts replied that it was very easy to map the information into queries. Both these observations allow us to conclude that the quality of the scenarios was very high. Analysts reported in answering Question 12 that the amount of time for performing the analysis was reasonable. This is an interesting observation and shows the types of expectations that intelligence analysts bring to the testing situation. They were not intimidated by the short turn-around times specified in the scenarios nor were they deterred by the requirement to work diligently for a 3-hour period. A highly positive rating in Question

7 indicates that the use of Google as a baseline system was successful in fulfilling the analyst's expectations of a generic system.

References

[1] Wacholder, Nina, Paul Kantor, Sharon Small, Tomek Strzalkowski, Diane Kelly, Robert Rittman, Sean Ryan and Robert Salkin. (2003). Evaluation of the HITIQA Analysts' Workshops. Report to ARDA.

Appendix A: Scenarios used in the baseline generation

S1: Nuclear Arms Relationship: Russia and Iraq

The Department of Defense has demanded a report on how Russia has influenced the nuclear arms program in Iraq. The department needs the summary by COB today. List the extent of the nuclear program in each country including funding, capabilities, quantity, etc. Your report should also include key figures in both the Russia and Iraq nuclear programs, any travels that these key figures have made to other countries in regards to a nuclear program, any weapons that have been used in the past by either country, any purchases or trades that have been made relevant to weapons of mass destruction (possibly oil trade, etc.), any ingredients and chemicals that have been used, any potential weapons that could be under development, other countries that are involved or have close ties to Russia or Iraq, possible locations of development sites, and possible companies or organizations that these countries work with for their nuclear arms program. Add any other information relating to the Russian and Iraqi Nuclear Arms Programs.

S2: Chemical Weapon: sarin

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.

L1: The al-Qaida Terrorist Group

As an employee of the Central Intelligence Agency, your profession entails knowledge of the al-Qaida terrorist group. Your division chief has ordered a detailed report on the al-Qaida Terrorist Group due in three weeks. Provide as much information as possible on this militant organization. Eventually, this report should present information regarding the most essential concerns, including who are the key figures involved with al-Qaida along with other organizations, countries, and members that are affiliated, any trades that al-Qaida has made with organizations or countries, what facilities they possess, where they receive their financial support, what capabilities they have (CBW program, other weapons, etc.) and how have they acquired them, what is their possible future activity, how their training program operates, who their new members are. Also, include any other relevant information to your report as you see fit.

L2: South Africa's WMD Program

You have been given 30 days to develop a comprehensive report on the South African chemical, biological and nuclear warfare program, for your division chief who is to present to the Secretary of Defense. Your report should include several key elements of the South African WMD Program, including what people, organizations, and countries are involved, what chemicals have been purchased and/or used, where the chemicals have been purchased from and from whom, how their WMD program was financed, where these development locations are, any proposed activity (use, distribution, etc.), any money transactions that have been made between these suspects and other organizations, and any other contacts, or travels that have been made by any of the primary figures involved. Supply any further information that can support your documentation.