A Comparison of Boolean and Natural Language Searching for the TREC-6 Interactive Task

William Hersh
Bikram Day

Division of Medical Informatics and Outcomes Research
School of Medicine
Oregon Health Sciences University

Abstract

The TREC-6 interactive task used a multi-site experimental protocol, where each participating site compared an "experimental" system with a common "control" system used at all sites. For the Oregon Health Sciences University site, the "experimental" system was a Boolean interface to the MG system, while the control system was, as for all sites, the natural language ZPRISE system. Performance was measured by aspectual recall and precision. OHSU searchers did well overall, achieving the highest overall aspectual precision. These searchers did obtain below-average aspectual recall overall, although they achieved above-average aspectual recall with the control system, indicating that for the TREC-6 interactive task, a natural language searching system was superior to a Boolean one.

Background

A long-standing research issue of interest to information retrieval (IR) researchers at Oregon Health Sciences University (OHSU) is whether end-user searchers achieve better results with Boolean or natural language searching. Previous research in this area is decidedly mixed. The first study to compare Boolean and natural language searching with real searchers was the CIRT study, which found roughly comparable performance between the two when utilized by search intermediaries (Robertson and Thompson 1990). Turtle found, however, that expert searchers using a large legal database obtained better results with natural language searching (Turtle 1994). We have performed several studies of medical end-user searching comparing Boolean and natural language approaches. Whether using recall-precision metrics in bibliographic (Hersh, Buckley et al. 1994) or full-text databases (Hersh and Hickam 1995), or using task-completion studies in bibliographic (Hersh, Pentecost et al. 1996) or full-text databases (Hersh, Elliot et al. 1995), the results have been comparable for both types of systems.

The TREC experiments have been an important landmark for IR experimentation, providing a common large database and set of queries for the entire research community. We therefore chose to use the TREC-6 interactive task to compare systems that featured Boolean and natural language searching. Since the "control" system for the task was ZPRISE, a natural language system, we used a Boolean system for our "experimental" system. The system we chose to use was a Web-based Boolean interface that has been developed for MG, an experimental retrieval system developed at the Royal Melbourne Institute of Technology (RMIT) (Witten, Moffat et al. 1994). MG has a natural language searching interface as well, which was not used for this experiment. Therefore the results are more applicable to the Boolean searching interface, which we designate in the paper as MG-B, than the MG system as a whole.

Methods

Systems

Both the "experimental" MG-B and "control" ZPRISE systems were compiled and run on a Sun Ultrasparc 140 with 256 megabytes of RAM running the Solaris 2.5.1 operating system. Each system ran as a server, with MG accessed by a CGI form from a Web page and ZPRISE accessed using its Z39.50 client software.

The systems were accessed by Compaq DeskPro 200 MHz Pentium Pro machines running Windows 95. For PRISE, the Xwin32 X-Windows software package (StarNet Inc., Sunnyvale, CA) was used. MG was accessed via Netscape Navigator 3.0 via the Boolean interface available from the MG Web site that we modified slightly to allow logging and to make some minor cosmetic changes. Figure 1 shows the MG-B interface. The interface performs an OR between words on the same line and an AND between lines.

Experimental Design

The details of the overall interactive task are described elsewhere in the proceedings. We describe here the experimental design at OHSU. The searchers were librarians or researchers who work in our IR research laboratory. None of the searchers was familiar with the experimental design coming into the experimental session, and none of them had ever used ZPRISE or MG.

Each searcher was given a series of questionnaires that had been circulated by email from the Rutgers group. A pre-experiment questionnaire asked users about demographic information and past searching experience. A post-search questionnaire was given after each search, while a post-experiment questionnaire was given at the end of the session to obtain feedback about the experiment and searching systems that were used.

Results

Searchers

Table 1 characterizes the searchers based on the information provided on the pre-experiment questionnaire. Even though none of the searchers had used the ZPRISE or MG systems before, all had substantial searching experience. One searcher was actually a librarian, while the other three have done substantial Web searching over the last several years.

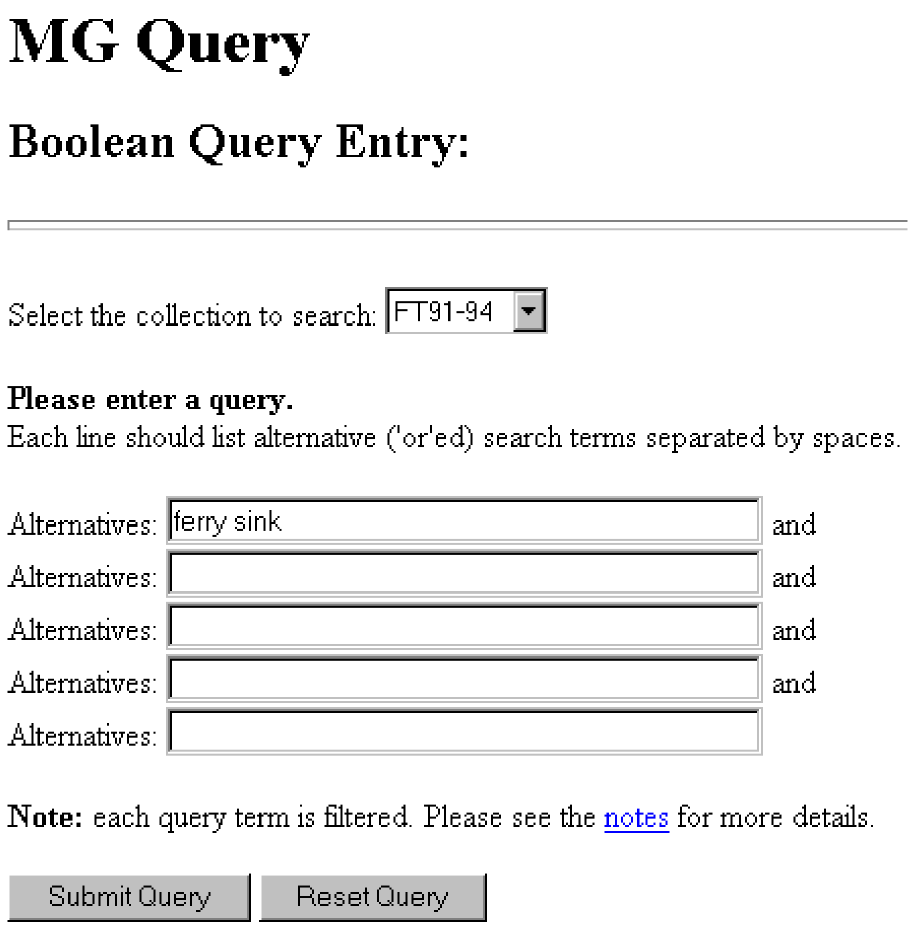

Figure 1 – MG Boolean Web interface.

Searcher	**LD**	**LS**	**KS**	**SM**
Highest degree	Masters	Masters	Bachelors	Masters
Age	41-50	41-50	21-30	31-40
Gender	M	F	M	F
Years searching	8	15	4	5
Experience searching (1-5)				
Library catalogs	5	5	4	3
CD-ROMs	4	5	2	1
Commercial on-line	2	4	1	3
Web browsers	5	5	5	5
Full-text databases	5	4	1	2
Ranked retrieval systems	5	5	1	4
Rel. Feedback sysems	3	3	5	1
Mouse-based interface	5	5	5	5
Occupation	Programmer	Librarian	Student	Programmer

Table 1 – OHSU searcher characteristics. Experience was rated on a 1 to 5 scale, with 1 indicating no experience and 5 a great deal.

Aspectual recall and precision

Table 2 shows the aspectual recall and precision for all of the groups participating in the task, both combined and broken down by experimental and control systems. The OHSU group obtained the highest aspectual precision of any group (0.85, mean 0.73, range 0.63-0.85). OHSU's overall aspectual recall was on the lower end (0.43, mean 0.45, range 0.32-0.52), although this was mainly due to the poor performance of our experimental system. The OHSU searchers were among the top performers in aspectual recall with the control ZPRISE system (0.49, mean 0.45, range 0.38-0.51).

Precision

Site	**Both**	**Experimental**	**Control**
BrklyINT	0.80	0.79	0.80
city	0.76	0.71	0.81
IBM	0.63	0.62	0.64
INQ4iai	0.64	0.63	0.64
INQ4iaip	0.76	0.75	0.77
INQ4int	0.67	0.67	NA
NMSU	0.83	0.82	0.83
OHSU	0.85	0.90	0.81
rmit	0.79	0.80	0.78
rutint1	0.67	0.67	NA
rutint2	0.66	0.66	NA
unc6ia	0.67	0.60	0.75
unc6ip	0.78	0.79	0.77
MEAN	**0.73**	**0.72**	**0.76**

Recall

Site	**Both**	**Experimental**	**Control**
BrklyINT	0.53	0.57	0.49
city	0.39	0.40	0.38
IBM	0.32	0.26	0.38
INQ4iai	0.41	0.36	0.45
INQ4iaip	0.47	0.50	0.44
INQ4int	0.48	0.48	NA
NMSU	0.46	0.47	0.45
OHSU	0.43	0.37	0.49
rmit	0.48	0.47	0.50
rutint1	0.46	0.46	NA
rutint2	0.53	0.53	NA
unc6ia	0.48	0.44	0.51
unc6ip	0.46	0.47	0.46
MEAN	**0.45**	**0.45**	**0.45**

Table 2 – Aspectual precision and recall values for both systems, the experimental system, and the control system for each site.

Table 3 shows the aspectual recall and precision by topic. OHSU searchers tended to follow the average trends.

Table 4 shows the results of individual OHSU searchers. An analysis of variance was done for recall and precision with a three-factor model based on topic, searcher, and system. There were no statistically significant differences in any factor for precision, but for recall there was definite statistical significance between topics ($p < .0001$), near significance between systems ($p = .07$), and no significance between searchers.

Table 5 shows the results of the post-experiment questionnaire. All of the searchers understood the task. They also had a tendency to believe the systems were different. ZPRISE was generally ranked as easier to use than MG-B.

The appendix of this paper contains a detailed narrative of one searcher (KS) for one topic (326i), as required for the interactive site reports. We include this narrative because it highlights some of the issues that arise in interactive searching experiments.

Precision

Topic	**All**	**E**	**C**	**OHSU-All**	**OHSU E**	**OHSU C**
303i	0.71	0.70	0.73	0.85	0.88	0.83
307i	0.79	0.79	0.79	0.85	0.70	1.00
322i	0.45	0.37	0.56	0.85	1.00	0.71
326i	0.92	0.93	0.90	0.92	1.00	0.83
339i	0.79	0.79	0.79	0.80	0.80	0.80
347i	0.74	0.71	0.77	0.83	1.00	0.67

Recall

Topic	**All**	**E**	**C**	**OHSU-All**	**OHSU E**	**OHSU C**
303i	0.89	0.86	0.93	1.00	1.00	1.00
307i	0.30	0.33	0.26	0.29	0.30	0.28
322i	0.11	0.10	0.13	0.19	0.11	0.28
326i	0.44	0.50	0.37	0.39	0.39	0.39
339i	0.72	0.69	0.75	0.55	0.30	0.80
347i	0.21	0.18	0.25	0.15	0.13	0.17

Table 3 – Aspectual precision and recall by query for all sites and OHSU.

Topic	Searcher	System	Recall	Precision
303i	KS	ZP	1.00	1.00
303i	LD	ZP	1.00	0.67
303i	LS	MG-B	1.00	0.75
303i	SM	MG-B	1.00	1.00
307i	KS	MG-B	0.39	0.69
307i	LD	MG-B	0.22	0.71
307i	LS	ZP	0.35	1.00
307i	SM	ZP	0.22	1.00
322i	KS	MG-B	0.11	1.00
322i	LD	MG-B	0.11	1.00
322i	LS	ZP	0.22	0.67
322i	SM	ZP	0.33	0.75
326i	KS	MG-B	0.33	1.00
326i	LD	MG-B	0.44	1.00
326i	LS	ZP	0.44	1.00
326i	SM	ZP	0.33	0.67
339i	KS	ZP	0.70	0.60
339i	LD	ZP	0.90	1.00
339i	LS	MG-B	0.50	0.60
339i	SM	MG-B	0.10	1.00
347i	KS	ZP	0.15	0.50
347i	LD	ZP	0.19	0.83
347i	LS	MG-B	0.12	1.00
347i	SM	MG-B	0.15	1.00

Table 4 – Aspectual recall and precision for OHSU searchers by topic-system pair.

Searcher	LD	LS	KS	SM
Understand task (1-5)	4	5	5	5
Task similar to others (1-5)	3	5	5	1
Systems different (1-5)	4	3	4	5
System comparisons				
Easier to use	MG-B	ZP	ZP	ZP
Easier to learn	MG-B	ZP	ZP	ZP
Liked best	ZP	ZP	MG-B	ZP

Table 5 – Post-experiment questionnaire. Leikert-scale ratings from 1 (not at all) to 5 (complete).

Conclusions

The TREC-6 interactive task was designed to be a multi-site experiment. As with most other sites, there was a small sample size at the individual OHSU site, which makes generalization of the results difficult. However, these experiments did show that OHSU searchers in general did well on the TREC-6 interactive task, and that natural language searching performed better than Boolean searching.

Acknowledgments

The authors would like to thank Paul Over at NIST for all of his assistance in getting ZPRISE to run, managing the experiments, and helping with the collection of data.

References

Hersh, W., C. Buckley, et al. (1994). OHSUMED: an interactive retrieval evaluation and new large test collection for research. *Proceedings of the 17th Annual International ACM Special Interest Group in Information Retrieval*, Dublin, 192-201.

Hersh, W., D. Elliot, et al. (1995). Towards new measures of information retrieval evaluation. *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, WA, 164-170.

Hersh, W. and D. Hickam (1995). An evaluation of interactive Boolean and natural language searching with an on-line medical textbook. *Journal of the American Society for Information Science,* 46: 478-489.

Hersh, W., J. Pentecost, et al. (1996). A task-oriented approach to information retrieval evaluation. *Journal of the American Society for Information Science,* 47: 50-56.

Robertson, S. and C. Thompson (1990). Weighted searching: the CIRT experiment. *Informatics 10: Prospects for Intelligent Retrieval*, York, 153-166.

Turtle, H. (1994). Natural language vs. boolean query evaluation: a comparison of retrieval performance. *Proceedings of the 17th Annual International ACM Special Interest Group in Information Retrieval*, 212-220.

Witten, I., A. Moffat, et al. (1994). *Managing Gigabytes - Compressing and Indexing Documents and Images*. New York, Van Nostrand Reinhold.

Appendix – Narrative Description of Topic 326i for One Searcher

The OHSU "experimental" system was a Boolean interface on top of the MG system. We used a simple Web interface that comes with the software distribution. This interface performs an OR between all words on an individual line and an AND between lines. Results are displayed in arbitrary (i.e., non-ranked) order. This document represents our narrative description for one searcher on topic 326i.

Topic

```
<num> Number: 326i
<title> Ferry Sinkings

<desc> Description:
Any report of a ferry sinking where 100 or more people lost
their lives.

<narr> Narrative:
To be relevant, a document must identify a ferry that has
sunk causing the death of 100 or more humans.  It must
identify the ferry by name or place where the sinking
occurred.  Details of the cause of the sinking would be
helpful but are not necessary to be relevant.  A reference
to a ferry sinking without the number of deaths would not be
relevant.

<aspects> Aspects:
Please save at least one RELEVANT document that identifies EACH DIFFERENT
ferry sinking of the sort described above. If one document document discusses
several such sinkings, then you need not save other documents that repeat
those aspects, since your goal is to identify different sinkings of the sort
described above.
```

Searches

The searcher entered 6 searches into the system. Unfortunately, our logging function for MG did not record the Boolean operators and only saved the stemmed version of terms entered (i.e., not the original query text). However, it is apparent from the results below that the searcher used at least one OR for the first three searches and ANDs for the last three searchers.

Time	Input	Docs Viewed	Docs Seen
1:11:17 PM	fer sink	666	0
1:11:53 PM	fer sink ship 100 dead	234	0
1:12:41 PM	fer sink ship	1148	1
1:14:25 PM	fer sink	53	9
1:26:30 PM	fer traged	22	2
1:28:56 PM	fer sunk	14	1

Documents Viewed

A total of 2,137 documents were viewed, 1,468 of which were unique. The overwhelming majority of these were retrieved by the first three searches.

Documents Seen

The searcher only chose to see one document from the results of the first three searches, and was likely overwhelmed by the massive output. After the fourth search, the searcher began choosing documents to see.

Time	Document
1:13:42 PM	FT911-3535
1:17:02 PM	FT931-8485
1:17:02 PM	FT943-312
1:17:02 PM	FT943-316
1:17:02 PM	FT943-535
1:17:02 PM	FT943-536
1:17:02 PM	FT943-543
1:17:02 PM	FT944-15661
1:17:02 PM	FT944-18722
1:17:02 PM	FT944-18877
1:27:10 PM	FT931-16406
1:27:10 PM	FT944-18719
1:29:28 PM	FT944-18938

Searcher Aspects

The searcher selected six aspects. He actually did not follow the instructions correctly, as the latter 5 aspects represent a single ferry sinking. In fact, he did not explicitly identify NIST aspect 5, the Belgian sinking, but got credit for it because one of his designated documents describes that sinking as well.

Sequence	Document	Text
1	FT931-8485	Crowded Ferry Sinks off Haiti
2	FT943-316	Ferries face calls for safety curbs; Estonia disaster brings
3	FT943-536	Safety Rules (Estonia)
4	FT943-543	Bow Doors Leak Reported (Estonia)
5	FT944-18722	Review of Emergency Procedures (Estonia)
6	FT944-18722	Reactions to the Disaster (Estonia)

NIST Aspects

The table below lists the aspects identified by NIST assessors.

Sequence	Text
1	Zairean ferry accident
2	Neptune ferry sinks
3	Korean ferry sinks west coast of South Korea
4	Moby Prince ferry fire - (2.5 miles at sea)
5	Herald Free Enterprise ferry off Belgian coast
6	Ferry capsizes Port of Mombasa
7	Estonian sinks
8	Bangladesh ferrys sink in Bay of Bengal (Oct 94)
9	Philippine ferry sinking (apparently in the Philippines)

NIST Aspects Coverage

The table below lists the documents and aspects covered as identified by NIST assessors.

Document	Coverage
FT921-924	000000000
FT922-6072	000000000
FT923-4546	000000000
FT923-6558	000000000
FT931-5947	100000000
FT931-8485	010000000
FT934-15680	001000000
FT934-1954	000110000
FT941-9683	000000000
FT942-12305	000001000
FT942-14757	000000000
FT943-178	000010100
FT943-312	000000100
FT943-316	000000100
FT943-3240	000010000
FT943-3945	000000000
FT943-3954	000000000
FT943-535	000000100
FT943-536	000010100
FT943-543	000010100
FT943-569	000000000
FT944-10102	000000100
FT944-10109	000000100
FT944-10853	000000000
FT944-11013	000000100
FT944-11048	000010100
FT944-11367	000000000
FT944-12822	000000000
FT944-15057	000000100
FT944-15143	000000100

FT944-15661	000000010
FT944-1600	000000100
FT944-17957	000000100
FT944-17958	000000000
FT944-18204	000000100
FT944-18217	000000000
FT944-18499	000000100
FT944-18508	000000100
FT944-18722	000000100
FT944-18875	000010100
FT944-18938	000000100
FT944-3393	000000000
FT944-5084	000000100
FT944-5248	000000101
FT944-5773	000000001
FT944-6361	000000000
FT944-6974	000000100

Conventional Recall and Precision

We attempted to apply conventional (i.e., document level, not aspectual) recall and precision calculations to the data. We grouped documents based on the one or more aspects they contained. As is seen below, the searcher identified aspects 2, 5, and 7. He viewed and saw all documents with these aspects. He also viewed and saw the document with aspect 8, but did not designate it as an aspect. He did not view nor see the documents with aspects 1, 3, 4, 6, and 9.

Of the 26 "relevant" documents, he retrieved 23. By standard calculations, his recall would be 88.5%. Calculating his precision is more difficult, due to the large number of documents viewed by the initial Boolean searches. It is unclear whether or not to include those.

Aspect(s)	Relevant	Viewed/Seen
1	1	0
2	1	1
3	1	0
4&5	1	0
5	1	1
5&7	3	3
6	1	0
7	15	15
8	1	1
9	1	0