Overview of the Sixth Text REtrieval Conference (TREC-6)*

Ellen M. Voorhees, Donna Harman
National Institute of Standards and Technology
Gaithersburg, MD 20899

Abstract

The Text REtrieval Conference is a workshop series designed to encourage research on text retrieval for realistic applications by providing large test collections, uniform scoring procedures, and a forum for organizations interested in comparing results. TREC contains two main retrieval tasks plus optional subtasks that allow participants to focus on particular common subproblems in retrieval. The emphasis on individual experiments evaluated in a common setting has proven to be very successful. In the six years since the beginning of TREC, the state of the art in retrieval effectiveness has approximately doubled, and technology transfer among research labs and between research systems and commercial products has accelerated. In addition, TREC has sponsored the first large-scale evaluations of Chinese language retrieval, retrieval of speech, and retrieval across different languages.

1 Introduction

As part of the U.S. Department of Commerce, the National Institute of Standards and Technology (NIST) has the mission to promote U.S. economic growth by working with industry to develop and apply technology, measurements, and standards. The Text REtrieval Conference (TREC), co-sponsored by NIST and the Defense Advanced Research Projects Agency (DARPA), was started in 1992 to support the text retrieval industry by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. In particular, the TREC workshop series has the following goals:

- to encourage research in text retrieval based on large test collections;
- to increase communication among industry, academia, and government by creating an open forum for the exchange of research ideas;
- to speed the transfer of technology from research labs into commercial products by demonstrating substantial improvements in retrieval methodologies on real-world problems; and
- to increase the availability of appropriate evaluation techniques for use by industry and academia, including development of new evaluation techniques more applicable to current systems.

For each TREC, NIST provides a test set of documents and questions. Participants run their retrieval systems on the data, and return to NIST a list of the retrieved top-ranked documents. NIST pools the individual results, judges the retrieved documents for correctness, and evaluates the results. The TREC cycle ends with a workshop that is a forum for participants to share their experiences.

This paper provides an overview of TREC-6, which was held at NIST on November 19–21, 1997. The next section summarizes the history of TREC to provide context. Subsequent sections define the TREC-6 tasks, describe the data collections used, and review the research findings. The final section discusses plans for future TRECs.

*Appears in **Information Processing and Management**, vol 36(1), pp. 3–35, January, 2000.

Table 1: TREC-6 participants

Apple Computer	MIT/IBM Almaden Research Center
AT&T Labs Research	NEC Corporation
Australian National University	New Mexico State U. (2 groups)
CEA (France)	NSA (Speech Research Branch)
Carnegie Mellon University	Open Text Corporation
Center for Information Research, Russia	Oregon Health Sciences U.
City University, London	Queens College, CUNY
CLARITECH Corporation	Rutgers University (2 groups)
Cornell U./SaBIR Research, Inc	Siemens AG
CSIRO (Australia)	SRI International
Daimler Benz Research Center Ulm	Swiss Federal Inst. of Tech.(ETH)
Dublin City University	TwentyOne (TNO/U-Tente/DFKI/Xerox/U-Tuebingen)
Duke U./U. of Colorado/Bellcore	U. of California, Berkeley
FS Consulting, Inc.	U. of California, San Diego
GE Corp./Rutgers U.	U. of Glasgow
George Mason U./NCR Corp.	U. of Maryland, College Park
Harris Corp.	U. of Massachusetts, Amherst
IBM T.J. Watson Research (2 groups)	U. of Montreal
ITI (Singapore)	U. of North Carolina (2 groups)
MSI/IRIT/U. Toulouse (France)	U. of Sheffield/U. of Cambridge
ISS (Singapore)	U. of Waterloo
APL, Johns Hopkins University	Verity, Inc.
Lexis-Nexis	Xerox Research Centre Europe
MDS at RMIT, Australia	

2 A brief history of TREC

The retrieval field has a long history of using retrieval experiments on test collections to advance the state of the art (Sparck Jones & Willett, 1997). However, by the early 1990's progress in the text retrieval field was hampered by the lack of realistically large test collections. Large test collections did exist, but they were proprietary, with each collection usually the result of a single company's efforts. The proprietary nature of the collections biased them in various ways, in addition to rendering them unavailable to most of the retrieval research community. TREC was conceived as a way to address this need for large, unbiased test collections.

The first TREC workshop was held in November, 1992 with 25 research groups participating. This first conference was largely dedicated to engineering issues: participants had to scale their systems to handle two gigabytes of documents rather than the two megabytes of documents that was typical of test collections at the time. However, the conference did demonstrate that the statistical methods used by these various groups were capable of handling operational amounts of text, and that research on these large test collections could lead to new insights in text retrieval.

TREC has since grown in both number of participants and number of different retrieval tasks (see sec. 7 for a discussion of the TREC tasks). Table 1 lists the 51 groups that participated in TREC-6. The 51 groups include representatives from 12 different countries, and academic, government, and commercial organizations. The diversity of the groups participating in TREC ensures that TREC represents many different approaches to text retrieval.

TREC provides a common test set to focus research on a particular retrieval task, yet actively encourages participants to do their own experiments within the umbrella task. The individual experiments broaden the scope of the research that is done within TREC and makes TREC more attractive to individual participants. This marshaling of research efforts has succeeded in improving the state of the art in retrieval technology.

The graph in Figure 1 shows that retrieval effectiveness has approximately doubled since the beginning of TREC. This means, for example, that retrieval engines that could retrieve three good documents within the

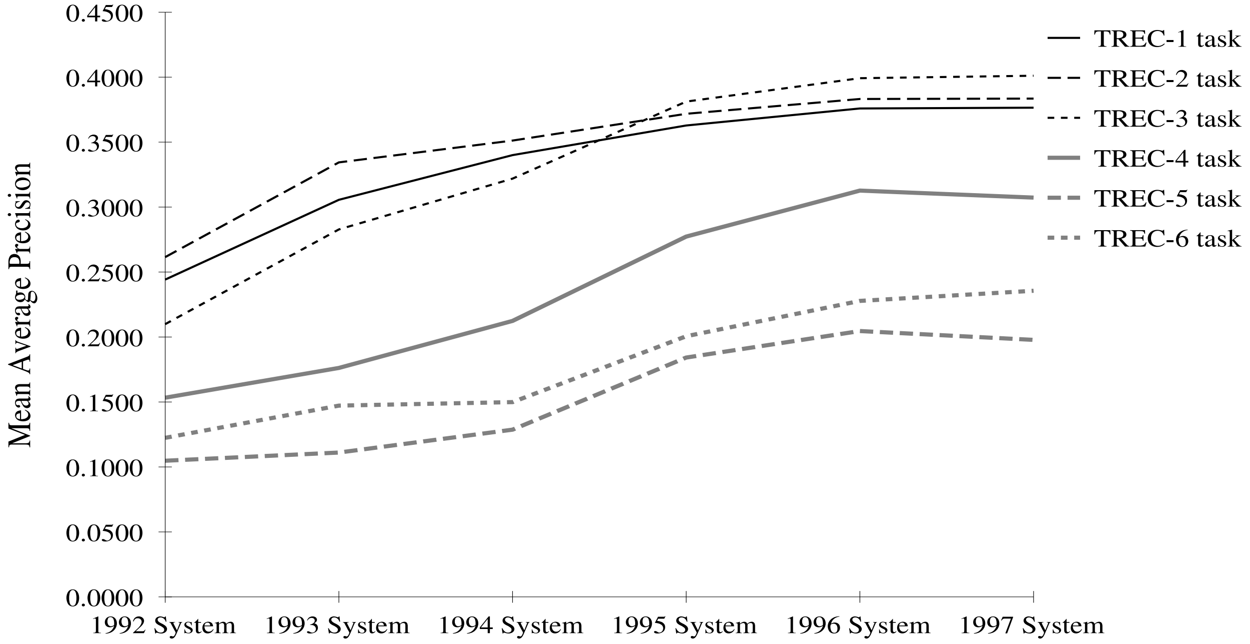

Figure 1: Retrieval effectiveness improvement for Cornell's SMART system, TREC-1 – TREC-6.

top ten documents retrieved in 1992 are now likely to retrieve six good documents in the top ten documents retrieved for the same search. The figure plots retrieval effectiveness for one well-known retrieval engine, the SMART system of Cornell University. The SMART system has consistently been one of the more effective systems in TREC, but other systems are comparable with it, so the graph is representative of the increase in effectiveness for the field as a whole.

Researchers at Cornell ran the version of SMART used in each of the six TREC conferences against each of the six ad hoc test sets (Buckley, Mitra, Walz, & Cardie, 1998). Each line in the graph connects the mean average precision scores produced by each version of the system for a single test. For each test, the TREC-6 system has a markedly higher mean average precision than the TREC-1 system. The recent decline in the absolute scores reflects the evolution towards more realistic, and difficult, test questions.

The six TREC conferences represent hundreds of retrieval experiments. The Proceedings of each conference captures the details of the individual experiments, and the Overview paper in each Proceedings summarizes the main findings of each conference. A special issue on TREC-2 was published in a previous volume of *Information Processing and Management* (Harman, 1995), which included an Overview of TREC-2 (Harman, 1994) as well as an analysis of the results by Sparck Jones (1995). Sparck Jones updates her analysis in another paper in this volume.

3 The Tasks

Each of the TREC conferences has centered around two main tasks, the routing task and the ad hoc task. In addition, starting in TREC-4 a set of "tracks," tasks that focus on particular subproblems of text retrieval, was introduced. This section describes the goals of the two main tasks. Details regarding the TREC-6 tracks are given in Section 7.

3.1 The routing task

The routing task in the TREC workshops investigates the performance of systems that use standing queries to search new streams of documents. These searches are similar to those required by news clipping services and library profiling systems. A true routing environment is simulated in TREC by using questions (called *topics* in TREC) for which the right set of documents to be retrieved is known for one document set, and then testing the systems' performance with those questions on a completely new document set.

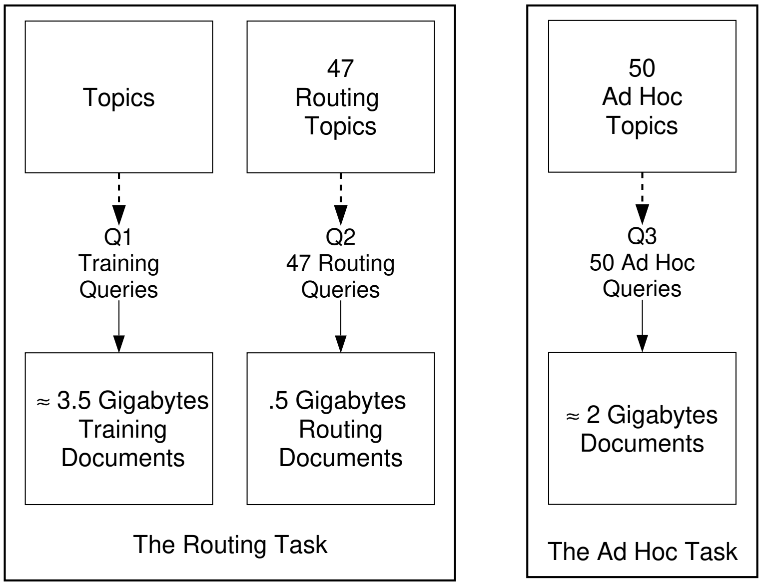

Figure 2: TREC main tasks.

The training for the routing task is shown in the left-hand column of Figure 2. Participants are given a set of topics and a document set that includes known relevant documents for those topics. The topics consist of natural language text describing a user's information need (see sec. 4.2 for details). The topics are used to create a set of queries (the actual input to the retrieval system) that are then used against the training documents. This is represented by Q1 in the diagram. Many Q1 query sets might be built to help adjust the retrieval system to the task, to create better weighting algorithms, and to otherwise prepare the system for testing. The result of the training is query set Q2, routing queries derived from the routing topics and run against the test documents.

The testing phase of the routing task is shown in the middle column of Figure 2. The output of running Q2 against the test documents is the official test result for the routing task.

3.2 The ad hoc task

The ad hoc task investigates the performance of systems that search a static set of documents using new topics. This task is similar to how a researcher might use a library—the collection is known but the questions likely to be asked are not known. The right-hand column of Figure 2 depicts how the ad hoc task is accomplished in TREC. Participants are given a document collection consisting of approximately 2 gigabytes of text and 50 new topics. The set of relevant documents for these topics in the document set is not known at the time the participants receive the topics. Participants produce a new query set, Q3, from the ad hoc topics and run those queries against the ad hoc documents. The output from this run is the official test result for the ad hoc task.

3.3 Task guidelines

In addition to the task definitions, TREC participants are given a set of guidelines outlining acceptable methods of indexing, knowledge base construction, and generating queries from the supplied topics. In general, the guidelines are constructed to reflect an actual operational environment and to allow fair comparisons among the diverse query construction approaches. The allowable query construction methods in TREC-6 are divided into *automatic* methods, in which queries are derived completely automatically from the topic statements, and *manual* methods, which includes queries generated by all other methods. The definition of manual query construction methods in TREC-6 permitted users to look at individual documents retrieved by the ad hoc queries and then reformulate the queries based on the documents retrieved.

Table 2: Document collection statistics. Words are strings of alphanumeric characters. No stop words were removed and no stemming was performed.

	Size (megabytes)	# Docs	Median # Words/Doc	Mean # Words/Doc
Disk 1				
Wall Street Journal, 1987–1989	267	98,732	245	434.0
Associated Press newswire, 1989	254	84,678	446	473.9
Computer Selects articles, Ziff-Davis	242	75,180	200	473.0
Federal Register, 1989	260	25,960	391	1315.9
abstracts of U.S. DOE publications	184	226,087	111	120.4
Disk 2				
Wall Street Journal, 1990–1992 (WSJ)	242	74,520	301	508.4
Associated Press newswire (1988) (AP)	237	79,919	438	468.7
Computer Selects articles, Ziff-Davis (ZIFF)	175	56,920	182	451.9
Federal Register (1988) (FR88)	209	19,860	396	1378.1
Disk 3				
San Jose Mercury News, 1991	287	90,257	379	453.0
Associated Press newswire, 1990	237	78,321	451	478.4
Computer Selects articles, Ziff-Davis	345	161,021	122	295.4
U.S. patents, 1993	243	6,711	4445	5391.0
Disk 4				
the *Financial Times*, 1991–1994 (FT)	564	210,158	316	412.7
Federal Register, 1994 (FR94)	395	55,630	588	644.7
Congressional Record, 1993 (CR)	235	27,922	288	1373.5
Disk 5				
Foreign Broadcast Information Service (FBIS)	470	130,471	322	543.6
the *LA Times*	475	131,896	351	526.5
Routing Test Data				
Foreign Broadcast Information Service (FBIS)	490	120,653	348	581.3

```
<DOC>
<DOCNO>FT911-3</DOCNO>
<PROFILE>AN-BEOA7AAIFT</PROFILE>
<DATE>910514
</DATE>
<HEADLINE>
FT 14 MAY 91 / International Company News:  Contigas plans DM900m east German
project
</HEADLINE>
<BYLINE>
By DAVID GOODHART
</BYLINE>
<DATELINE>
BONN
</DATELINE>
<TEXT>
CONTIGAS, the German gas group 81 per cent owned by the utility Bayernwerk, said
yesterday that it intends to invest DM900m (Dollars 522m) in the next four years
to build a new gas distribution system in the east German state of Thuringia.  ...
</TEXT>
</DOC>
```

Figure 3: A document extract from the *Financial Times*.

4 The Test Collections

Like most traditional retrieval test collections, there are three distinct parts to the collections used in TREC: the documents, the questions or topics, and the relevance judgments or "right answers." This section describes each of these pieces for the collections used in the TREC-6 main tasks.

4.1 Documents

TREC documents are distributed on CD-ROM's with approximately 1 GB of text on each, compressed to fit. For TREC-6, Disks 1–4 were all available as training material (see Table 2) and Disks 4 and new Disk 5 were used for the ad hoc task. New Foreign Broadcast Information Service (FBIS) data (also shown in Table 2) were used for testing in the routing task.

Documents are tagged using SGML to allow easy parsing (see Fig. 3). The documents in the different datasets have been tagged with identical major structures, but they have different minor structures. The philosophy in the formatting at NIST is to leave the data as close to the original as possible. No attempt is made to correct spelling errors, sentence fragments, strange formatting around tables, or similar faults.

4.2 Topics

In designing the TREC task, there was a conscious decision made to provide "user need" statements rather than more traditional queries. Two major issues were involved in this decision. First, there was a desire to allow a wide range of query construction methods by keeping the topic (the need statement) distinct from the query (the actual text submitted to the system). The second issue was the ability to increase the amount of information available about each topic, in particular to include with each topic a clear statement of what criteria make a document relevant.

The topics used in TREC-1 and TREC-2 (topics 1–150) were very detailed, containing multiple fields and lists of concepts related to the subject of the topics. The ad hoc topics used in TREC-3 (151–200) were much shorter and did not contain the complex structure of the earlier topics. Nonetheless, participants in TREC-3 felt that the topics were still too long compared with what users normally submit to operational retrieval systems. Therefore the TREC-4 topics (201–250) were made even shorter: a single field consisting of a one sentence description of the information need. Figure 4 gives a sample topic from each of these sets.

One of the conclusions reached in TREC-4 was that the much shorter topics caused both manual and automatic systems trouble, and that there were issues associated with using short topics in TREC that needed further investigation (Harman, 1996). Accordingly, the TREC-5 ad hoc topics re-introduced the title and narrative fields, making the topics similar in format to the TREC-3 topics. TREC-6 topics used this same format, as shown in Figure 5. While having the same format as the TREC-3 topics, on average both the TREC-5 and the TREC-6 topics were shorter (contained fewer words) than the TREC-3 topics.

Since TREC-3, the ad hoc topics have been created by the same person (or *assessor*) who performed the relevance assessments for that topic. Each assessor comes to NIST with ideas for topics based on his or her own interests, and searches the ad hoc collection (looking at approximately 100 documents per topic) to estimate the likely number of relevant documents per candidate topic. NIST personnel select the final 50 topics from among the candidates based on having a range of estimated number of relevant and balancing the load across assessors.

4.2.1 Runs using different topic fields Groups who performed automatic ad hoc runs were required to do at least one run using a short version of the topic, i.e., the Description field. In what follows, these runs are called "short" runs, automatic runs that used only the Title field are called "title" runs, and automatic runs that used the entire topic are called "long" runs. Manual runs had no length requirements, and are assumed to be based on the entire topic text. Unfortunately, NIST did not inform the assessors that the different pieces of the topics would be used differently when they created the topics, and this confounds the conclusions that can be drawn from runs using different topic lengths.

The assessors treated each topic as a single unit, and did not necessarily repeat themselves in the different parts. Thus, some Description fields do not contain "Title" words — words that were specifically chosen to represent the core meaning of the topic! Specifically, none of the title words occur in the description for 5 topics, at least one title word is missing from the description for 22 topics, and the description contains all of the title words for the remaining 23 topics. Given this construction of the topics, a valid comparison is title *vs.* description+title, not title *vs.* description. A more thorough discussion of the effect of using different topic sections is given in section 6.

4.3 Relevance assessments

Relevance judgments are of critical importance to a test collection. For each topic it is necessary to compile a list of relevant documents—as comprehensive a list as possible. All TRECs have used the pooling method (Sparck Jones & van Rijsbergen, 1975) to assemble the relevance assessments. In this method a pool of possible relevant documents is created by taking a sample of documents selected by the various participating systems. This pool is then shown to the human assessors. The particular sampling method used in TREC is to take the top 100 documents retrieved in each submitted run for a given topic and merge them into the pool for assessment. This is a valid sampling technique since all the systems used ranked retrieval methods, with those documents most likely to be relevant returned first. On average, an assessor judges approximately 2000 documents per topic.

Given the vital role relevance judgments play in a test collection, it is important to assess the quality of the judgments created in TREC. In particular, both the *completeness* and the *consistency* of the relevance judgments are of interest. Completeness measures the degree to which all the relevant documents for a topic have been found; consistency measures the degree to which the assessor has marked all the "truly" relevant documents relevant and the "truly" irrelevant documents irrelevant.

The completeness of the TREC relevance judgments has been investigated both at NIST (Harman, 1996) and independently at the Royal Melbourne Institute of Technology (RMIT) (Zobel, 1998). Both studies found that the completeness for most topics is adequate, though topics with very many relevant documents are likely to have yet more relevant documents that have not been found through pooling. For this reason, NIST has deliberately chosen more tightly focused topics in the more recent TRECs. Both studies also found that any lack of completeness did not bias the results of particular systems. Indeed, the RMIT study showed that systems that did not contribute documents to the pool can still be evaluated fairly with the resulting judgments.

The consistency of the TREC judgments was investigated at NIST by obtaining multiple independent assessments for a set of topics and evaluating systems using each of the different judgment sets (Voorhees,

```
<num> Number:  051

<dom> Domain:  International Economics

<title> Topic:  Airbus Subsidies

<desc> Description:
Document will discuss government assistance to Airbus Industrie, or mention
a trade dispute between Airbus and a U.S. aircraft producer over the issue of
subsidies.

<narr> Narrative:
A relevant document will cite or discuss assistance to Airbus Industrie by the
French, German, British or Spanish government(s), or will discuss a trade dispute
between Airbus or the European governments and a U.S. aircraft producer, most
likely Boeing Co.  or McDonnell Douglas Corp., or the U.S. government, over
federal subsidies to Airbus.

<con> Concept(s):
1.  Airbus Industrie
2.  European aircraft consortium, Messerschmitt-Boelkow-Blohm GmbH, British
Aerospace PLC, Aerospatiale, Construcciones Aeronauticas S.A.
3.  federal subsidies, government assistance, aid, loan, financing
4.  trade dispute, trade controversy, trade tension
5.  General Agreement on Tariffs and Trade (GATT) aircraft code
6.  Trade Policy Review Group (TPRG)
7.  complaint, objection
8.  retaliation, anti-dumping duty petition, countervailing duty petition,
sanctions
```

```
<num> Number:  168
<title> Topic:  Financing AMTRAK

<desc> Description:
A document will address the role of the Federal Government in financing the
operation of the National Railroad Transportation Corporation (AMTRAK).

<narr> Narrative:
A relevant document must provide information on the government's responsibility
to make AMTRAK an economically viable entity.  It could also discuss the
privatization of AMTRAK as an alternative to continuing government subsidies.
Documents comparing government subsidies given to air and bus transportation with
those provided to AMTRAK would also be relevant.
```

```
<num> Number:  207

<desc> What are the prospects of the Quebec separatists achieving independence
from the rest of Canada?
```

Figure 4: The evolution of TREC topic statements. Sample topic statement from TRECs 1 and 2 (top), TREC-3 (middle), and TREC-4 (bottom).

```
<num> Number:  312
<title> Hydroponics

<desc> Description:
Document will discuss the science of growing plants in water or some substance
other than soil.

<narr> Narrative:
A relevant document will contain specific information on the necessary nutrients,
experiments, types of substrates, and/or any other pertinent facts related to the
science of hydroponics.  Related information includes, but is not limited to, the
history of hydroponics, advantages over standard soil agricultural practices,
or the approach of suspending roots in a humid enclosure and spraying them
periodically with a nutrient solution to promote plant growth.
```

Figure 5: A sample TREC-6 topic.

1998). The study confirmed that the comparative results for different runs remains stable despite changes in the underlying judgments. Taken together, these studies validate the use of the TREC collections for retrieval research.

5 Evaluation

An important element of TREC is to provide a common evaluation forum. A standard evaluation package, called `trec_eval`, is used to evaluate each of the submitted runs. `trec_eval` was developed by Chris Buckley at Cornell University and is available by anonymous ftp from `ftp.cs.cornell.edu` in the `pub/smart` directory. TREC reports a variety of recall- and precision-based evaluation measures for each run to give a broad picture of the run.

The single-valued evaluation measure most frequently used in TREC is the mean (non-interpolated) average precision. The average precision for a single topic is the mean of the precision obtained after each relevant document is retrieved (using zero as the precision for relevant documents that are not retrieved). The mean average precision for a run consisting of multiple topics is the mean of the average precision scores of each of the individual topics in the run.

The average precision measure has a recall component in that it reflects the performance of a retrieval run across all relevant documents, and a precision component in that it weights documents retrieved earlier more heavily than documents retrieved later. Geometrically, mean average precision is the area underneath a non-interpolated recall-precision curve.

Unless otherwise stated, the retrieval results reported below are given in terms of mean average precision.

6 Retrieval Results

One of the important goals of the TREC conferences is that the participating groups freely devise their own experiments within the TREC task. For some groups this means doing the routing and/or ad hoc task with the goal of achieving high retrieval effectiveness performance. For other groups, however, the goals are more diverse and may mean experiments in efficiency or unusual ways of using the data. This overview of the results discusses the effectiveness of the systems and analyzes some of the similarities and differences in the approaches that were taken. In all cases, readers are referred to the system papers for more details.

6.1 TREC-6 ad hoc automatic results

The TREC-6 ad hoc evaluation used new topics (topics 301–350) against two disks of training documents (disks 4 and 5). A dominant feature of the ad hoc task was the desire of groups to continue to work with both the short and long versions of the topics (as in TREC-5), and in addition to try a title only version. All three parts of the topics were built by the assessors, with the title being constrained to three words. Systems doing automatic query building were required to submit at least one run using the short version of

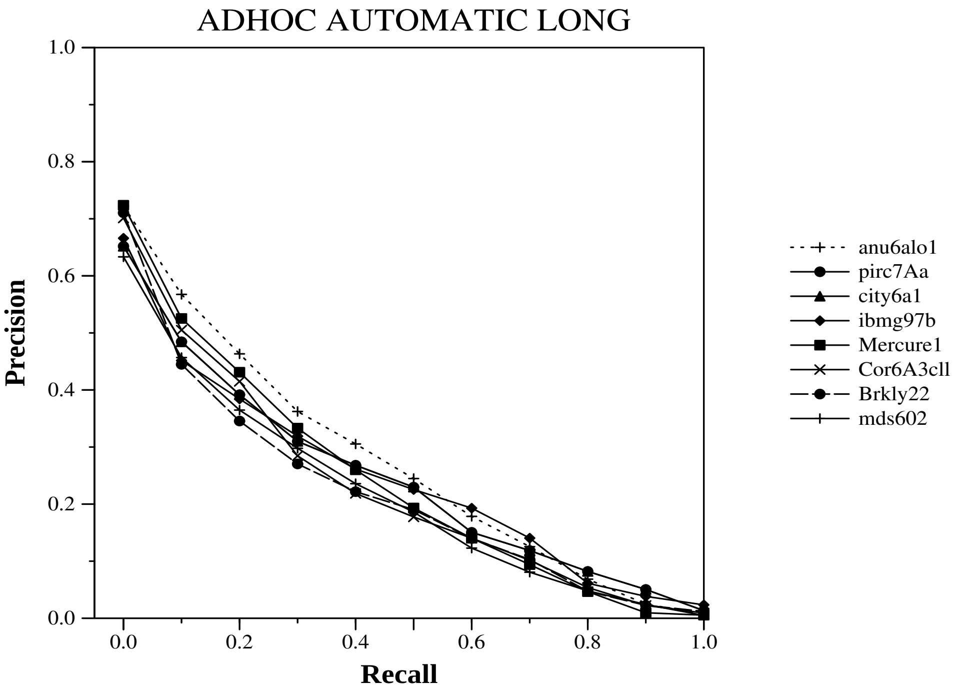

Figure 6: Recall-Precision graph for the top eight automatic ad hoc runs using the full topic.

the topic (only the description field), but in addition they could submit runs using either the title version or the long (full topic) version. Groups doing manual query building were assumed to be using the full topic.

There were 79 sets of official results for ad hoc evaluation in TREC-6, with 74 of them based on runs for the full data set. Of these, 57 used automatic construction of queries, with 12 official title runs, 29 short runs, and 16 long runs. Seventeen groups used manual construction.

6.2 Long (Full Topic) Runs

Figure 6 shows the recall-precision curves for the eight TREC-6 groups with the highest mean average precision using automatic construction of queries for the long (full topic) version of the topics. The runs are ranked by average precision and only one run is shown per group. These graphs (and others in this section) are not intended to show specific comparison of results across sites but rather to provide a focal point for discussion of methodologies used in TREC. For more details on the various runs and procedures, please see the cited papers.

The following runs are represented in the graph:

anu6al01 – from the Australian National University (Hawking, Thistlewaite, & Craswell, 1998) who used a parallel architecture with an emphasis on efficiency; new for TREC-6 was the use of the Cornell variant of the OKAPI BM25 term weighting, and major experiments to determine correct parameters for pseudo-relevance feedback (automatic relevance feedback using the top retrieved documents).

pirc7Aa – from Queens College, CUNY (Kwok, Grunfeld, & Xu, 1998) who used their spreading activation model for a two-stage search (initial search for doing pseudo-relevance feedback and a final search including expansion terms).

city6a1 – from City University, London (Walker, Robertson, Boughanem, Jones, & Sparck Jones, 1998) who ran many experiments investigating the various parameters in the BM25 weighting technique, including adding provisions for using non-relevant documents.

ibmg97b – from IBM T.J. Watson Research Center (Brown & Chong, 1998) who used a probabilistic system that includes the lexical affinities (statistical phrases) in the topic.

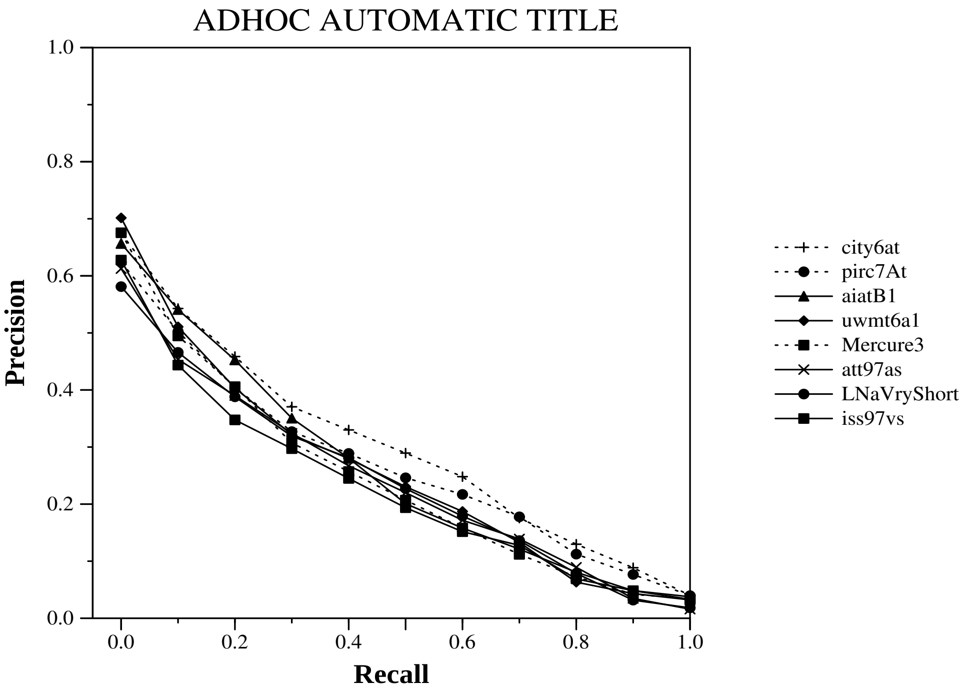

Figure 7: Recall-Precision graph for the top eight automatic ad hoc runs using the title only.

Mercure1 – from MSI/IRIT/SIG/CERISS (Boughanem & Soulé-Dupuy, 1998) who worked with a spreading activation model that incorporated the OKAPI/SMART BM25 weighting algorithm.

Cor6A3cl – from Cornell/SaBIR Research (Buckley et al., 1998) who concentrated on better initial retrieval and improved expansion, including a clustering of candidate documents for topic expansion to help improve term selection.

Brkly22 – from the University of California, Berkeley (Gey & Chen, 1998) who used a probabilistic system based on logistic regression.

mds602 – one of a comprehensive series of factor analysis runs from MDS/RMIT (Fuller et al., 1998) who investigated various known successful components of retrieval, including stopwords, stemming, passage retrieval, term expansion, methods of combining results, and query length.

6.3 Short (Description Only) Runs

The method used at NIST to construct the topics for TREC-6 (discussed in section 4.2.1) caused very unusual results for the required short runs. The titles of the topics generally contained excellent topic descriptors, but in over half the topics these terms were not included in the short (description section) of the topic. For many of the topics, therefore, the input to the short run consisted of a very poor set of terms. Results from the "short" runs using the only the description section should be viewed with great caution, and most groups redid their short runs to include the title (see individual papers for results). For this reason, these runs are not included in this paper, although they are included in the TREC-6 proceedings (Voorhees & Harman, 1998).

6.4 Title Only Runs

Figure 7 shows the recall-precision curves for the eight TREC-6 groups with the highest mean average precision using automatic construction of queries for the title only version of the topics.

The following runs are represented in the graph:

city6at – from City University, London (Walker et al., 1998) who used the same methods as for their "long" run, but with different parameter settings (particularly applying less weight to negative feedback terms and no weighting for query term frequency).

pirc7Aat – from Queens College, CUNY (Kwok et al., 1998)) who also used the same system as for their long run, but tried an unsuccessful experiment in document reranking before topic expansion.

aiatB1 – from Apple Computing who did not submit a paper for TREC-6, so nothing is known about how this run was made.

uwmt6a1 – from the University of Waterloo (Cormack, Clarke, Palmer, & To, 1998) who did their first automatic runs in TREC-6; this run is based on a merging of three runs: a cover density run, a run using a modified OKAPI weighting, and a third run using a modified OKAPI expansion method.

Mercure3 – from MSI/IRIT/SIG/CERISS (Boughanem & Soulé-Dupuy, 1998) who used the same system as for their long run using only title words.

att97as – from AT&T Labs Research (Singhal, 1998) who extended the basic Cornell ad hoc approach by using a new method to reweight the topic terms prior to term expansion.

LNaVryShort – from Lexis-Nexis (Lu, Meier, Rao, Miller, & Pliske, 1998) who used their EUREKA toolbox in a simplified version of their description-only run; in this run only the OKAPI algorithm was used, and 26 terms were added from an internally-built thesaurus before a relevance feedback process (not Rocchio) was used to produce the final results.

iss97vs – from the Institute for Systems Science (Leong, 1998) who used this run as a baseline automatic run for the major experiments they ran using manual queries.

The INQUERY system from the University of Massachusetts (Allan et al., 1998) did not make the above charts since they ran with only one-half the data (by mistake). Correct runs (see the cited paper) show that they performed as well as most of the systems shown.

6.5 TREC-6 ad hoc manual results

Figure 8 shows the recall-precision curves for the eight TREC-6 groups with the highest mean average precision using manual construction of queries. Note that manual query construction included user interaction in TREC-6, i.e. the rules allowed initial results to be viewed and the queries changed, with no restrictions on how much time could be spent. Therefore the amount of human effort required for these various techniques should be considered when comparing the retrieval results. A short summary of the techniques used in these runs follows; for more details on the various runs and procedures, see the cited papers.

The following runs are represented in the graph:

uwmt6a0 – from the University of Waterloo (Cormack et al., 1998) who submitted this run as part of an experiment in new ways to build test collections (Cormack, Palmer, & Clarke, 1998). The group spent an average of 2.1 hours per topic searching the collection, and submitted to NIST the list of documents the searchers felt were relevant.

CLAUG – from CLARITECH Corp. (Milic-Frayling, Zhai, Tong, Jansen, & Evans, 1998) who spent an average of about 20 minutes per topic to perform testing on two variations of relevance feedback.

anu6min1 – from the Australian National University (Hawking et al., 1998) who investigated how well a relatively naive user could perform a series of edits on automatically generated queries.

LNmShort – from Lexis-Nexis (Lu et al., 1998) who also experimented with manual editing of automatically generated queries.

geruna1 – from GE Corporate R&D/Rutgers University (Strzalkowski, Lin, & Perez-Carballo, 1998)) who experimented with users selectively adding automatically-generated query-related summaries for the top 30 documents retrieved by the original topic.

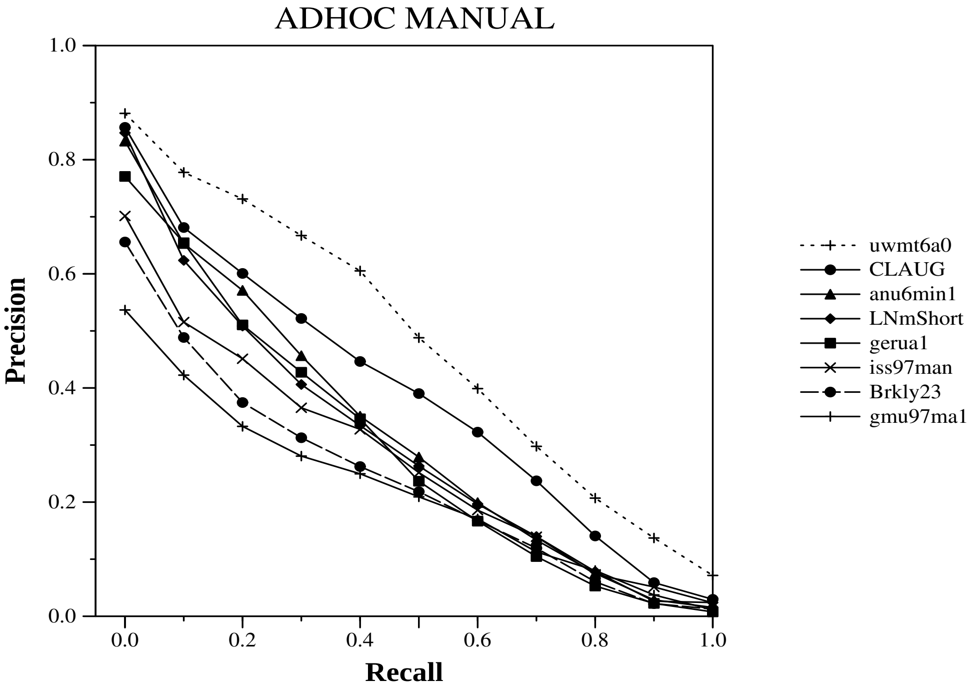

Figure 8: Recall-Precision graph for the top eight manual ad hoc runs.

iss97man – from the Institute for Systems Science (Leong, 1998) who ran an elaborate user experiment to test the hypothesis that expert users would be able to use very precise search terms.

Brkly23 – from the University of California, Berkeley (Gey & Chen, 1998) who did a manual reformation of their automatic queries.

gmu87ma1 – from George Mason University/OIT/NCR (Lundquist, Holmes, Grossman, Frieder, & McCabe, 1998) who used their relational database model information retrieval system to experiment with manually pre-defined concept lists combined in different ways.

6.5.1 Discussion of TREC-6 ad hoc results Since a dominant feature of the TREC-6 ad hoc task was the use of three different versions (lengths) of the topic, it is interesting to note the somewhat unexpected effects of this. The results using the title only (Very Short version) were surprisingly good, whereas those that used only the description (Short version) were considerably worse. Results using all three parts of the topic (Full version) were approximately the same as the results using the title only. These effects were generally consistent across all participating groups.

However, it would be unwise to generalize these results by claiming that systems do as well with very short (three word) topics as with much longer ones. As with most information retrieval testing, there is a huge variation across topics. For example, the table below shows the number of topics for which a given length was better than the other two lengths as measured by average precision for two sets of runs, the City

	Title	Short	Long
City	21	11	18
PIRCS	23	9	18

University runs and the CUNY (PIRCS system) runs. The counts in the table show that each topic length had some topics for which it formed the best query.[1]

[1] While the counts are very similar, the set of topics for which one length is better than the others differs between the two groups.

Many of the TREC-6 topics turned out to have very few relevant documents, and in most of these cases all of the relevant documents were retrievable using only the keywords in the title. In these cases the full topic simply adds more "noise" to the query. An extreme example of this is Topic 312 shown in Figure 5. The single title word, "hydroponics," appears in all of the 11 relevant documents and in 18 documents total. This simple separation between relevant and non-relevant documents is not true of all the TREC-6 topics, and is probably not true of user requests in general, but the highly precise terms in the TREC-6 titles both illustrate the power of a well-constructed user query and create biased results.

Some of the participating groups used the same retrieval techniques for all topic lengths. Given the above discussion, this is likely to be less effective than adapting techniques to the specific parts of the topic being used. For example, City University used fewer documents (top 7 *vs.* top 10 *vs.* top 15) for mining of expansion terms, added fewer expansion terms (20 *vs.* 30 *vs.* 30), and gave more weight to the original topic terms (3.5 *vs.* 2.5 *vs.* 2.5) for the long, short, and title versions of the topic. Lexis-Nexis, Cornell, and the University of Waterloo tried different techniques for different lengths of topics. Cornell used a new technique, SuperConcepts, for the description only runs and not the full topic runs. Lexis-Nexis used a much simpler version of their elaborate data fusion techniques for the title runs, whereas the University of Waterloo used more data fusion for their title-only run.

A second theme that dominated the TREC-6 ad hoc task was the continued spread of the newer, better techniques across most participating groups. Some techniques have now become standard usage, and TREC-6 saw both some interesting adaptations of these techniques to new retrieval models, and some further elaboration of these techniques by their originators. Table 3 shows some of these now-standard techniques, along with their spread and elaboration history.

Six different research areas are shown in the table, with research in many of these areas triggered by changes in the TREC evaluation environment. For example, the use of subdocuments or passages was caused by the initial difficulties in handling full text documents, particularly excessively long ones. The use of better term weighting, including correct length normalization procedures, made this technique less used in TRECs 4 and 5, but it resurfaced in TREC-6 to facilitate better input to relevance feedback.

The table also shows the rapid spread of successful technology across groups. Most groups spent TREC-1 simply struggling to scale-up their systems from searching several megabytes of documents to searching two gigabytes of documents. However, two new techniques were already being used by TREC-2. The Okapi system from City University, London was compelled to experiment with new term weighting algorithms since their initial algorithm did not scale to full text. By TREC-3 this algorithm had been "perfected" into the BM25 algorithm now in use by many of the systems in TREC-6. Continuing along this same row in table 3, two other systems (the SMART system from Cornell and the INQUERY system from the University of Massachusetts) changed their weighting algorithms in TREC-4 based on analysis comparing their old algorithms to the new BM25 algorithm. By TREC-5 many of the groups had adopted these new weighting algorithms, with the early adopters being those systems with similar structural models to the Okapi, SMART, or INQUERY systems.

TREC-6 saw even further expansion of the use of these new weighting algorithms (alternatively called the Okapi/SMART algorithm, or the Cornell implementation of the Okapi algorithm). In particular, many groups adapted these algorithms to new models, often involving considerable experimentation to find the correct fit. For example IRIT modified the Okapi algorithm to fit a spreading activation model, IBM modified it to deal with unigrams and trigrams, and the Australian National University and the University of Waterloo used it in conjunction with various types of proximity measures. Of major note is the fact that City University also ran major experiments with the BM25 weighting algorithm in TREC-6, including extensive exploration of the various existing parameters, and addition of some new ones!

The second new technique started back in TREC-2 (the second line of table 3) was the use of smaller sections of documents, called subdocuments, by the PIRCS system at City University of New York. Again this issue was forced by the difficulty of using the PIRCS spreading activation model for documents having a wide variety of lengths. By TREC-3 many of the groups were also using subdocuments, or passages, to help with retrieval. But, as mentioned before, TREC's 4 and 5 saw far less use of this technique as many groups dropped the use of passages due to minimal added improvements in performance.

TREC-6 saw a revival in the use of passages, but generally only for specific uses. Whereas the PIRCS system continued to use 550-word subdocuments for all its processing, most systems used passages only in

	TREC-2	TREC-3	TREC-4	TREC-5	TREC-6
Term weighting	baseline for most systems beginning of Okapi weighting experiments	Okapi perfects BM25 algorithm	new SMART weighting algorithm new INQUERY weighting algorithm	use of Okapi / SMART weighting algorithms by other groups	adaptations of Okapi / SMART algorithms in most systems
Passages	use of subdocuments by PIRCS system	heavy use of passages / subdocuments	decline in use of passages		use of passages in relevance feedback
Automatic query expansion		beginning of expansion using top X documents	heavy use of expansion using top X documents	beginning of more complex expansion schemes	more sophisticated expansion experiments by many groups
Manual query modification		beginning of manual expansion using other sources	major experiments in manual editing / user-in-the-loop	extensive user-in-the-loop experiments	return to manual editing & simple user-in-the-loop experiments
Other new areas		initial use of “data fusion”		beginning of more concentration on initial topic	more complex use of “data fusion” continued focus on initial topic, including title

Table 3: Use of new techniques in the ad hoc task

the topic expansion phase. The Australian National University worked with "hot spots" of 500 characters surrounding the original topic terms to locate new expansion terms. AT&T used overlapping windows of 50 words to help rerank the top 50 documents before selecting the final documents for use in expansion. The University of Waterloo used passages of maximum length 64 words to select expansion terms, whereas Verity used their automatic summarizer for this purpose. Two groups (Lexis-Nexis and MDS) performed major experiments in the use of passages, particularly when employed in conjunction with other methods as input to data fusion.

The query expansion/modification techniques shown in the third and fourth lines of the table were started when the topics were substantially shortened in TREC-3. As described in section 4.2, the format of the topics was modified to remove a valuable source of keywords: the concept section. In the search for some technique that would automatically expand the topic, several groups revived an old technique of assuming that the top retrieved documents are relevant, and then using them in relevance feedback. This technique, which had not worked on smaller collections, turned out to work very well in the TREC environment.

By TREC-6 almost all groups were using variations on expanding queries using information from the top retrieved documents (often called *pseudo-relevance feedback*). There are many parameters needed for success here, and groups continue to investigate the best settings for these parameters. Whereas there is general system convergence on some of these parameters, such as how many top documents to use for mining terms, how many terms to select, and how to weight those terms, these still need to be tested by systems adopting these techniques. Additionally there continue to be elaborations on these techniques, such as the several groups (City University, AT&T, and IRIT) that successfully got information from negative feedback in TREC-6.

Groups that built their queries manually also looked into better query expansion techniques starting in TREC-3. By TREC-5 these had evolved into very extensive user-in-the-loop experiments. Many of the manual experiments seen in TREC-6, however, go back to the simpler scenario of having users edit the automatically-generated query, or having users select documents to be used in automatic relevance feedback. Several of the groups had specific user strategies that they tested in TREC-6.

Other areas have seen concentrated research in the ad hoc task. Data fusion has been used in TREC by many groups in various ways, but has increased in complexity over the years. In TREC-6, for example, several groups such as Lexis-Nexis used multiple stages of data fusion, including merging results from different term weighting schemes and from different query expansion schemes. Starting in TREC-5, there have been experiments by several groups to "mine" more information from the initial topic, rather than simply treating the topic as a bag of potential keywords for input to the system. The INQUERY system from the University of Massachusetts has worked in all TREC's to automatically build more structure into their queries, based on information they have mined from the topic. In an effort to further improve performance, more groups have experimented with other information in the initial topic. This includes making more use of term proximity features (Australian National University, University of Waterloo, and IBM), clustering potential query expansion terms to maintain the initial topic balance (Cornell University), and looking for clues that would suggest a need for more emphasis on certain topic terms (AT&T and CUNY).

6.6 TREC-6 routing results

The routing evaluation used a specifically selected subset of the training topics against a new set of test documents. Due to the difficulty of obtaining appropriate documents, the test set in the TREC-4 and TREC-5 routing tests came from different sources than the training data. As a result, the training and test sets had different characteristics, including very different numbers of relevant documents for topics. The main emphasis in the TREC-6 task, therefore, was to investigate routing performance when the training and testing data were taken from the same source. Since the TREC-5 routing task had used a document stream from the Foreign Broadcast Information Service (FBIS) as its test set, a new stream of FBIS documents was selected as the TREC-6 test set. The TREC-6 routing topics consisted of 38 topics used in TREC-5 that had at least 5 relevant documents in the original FBIS stream, plus nine new topics (that had minimal training data on the original FBIS stream). The histogram in Figure 9 shows that the training and test data do have similar numbers of relevant documents for most topics.

There were 33 sets of results for routing evaluation, 28 that used automatic construction of queries and 5 that used manual construction. Figure 10 shows the recall-precision curves for the eight TREC-6 groups

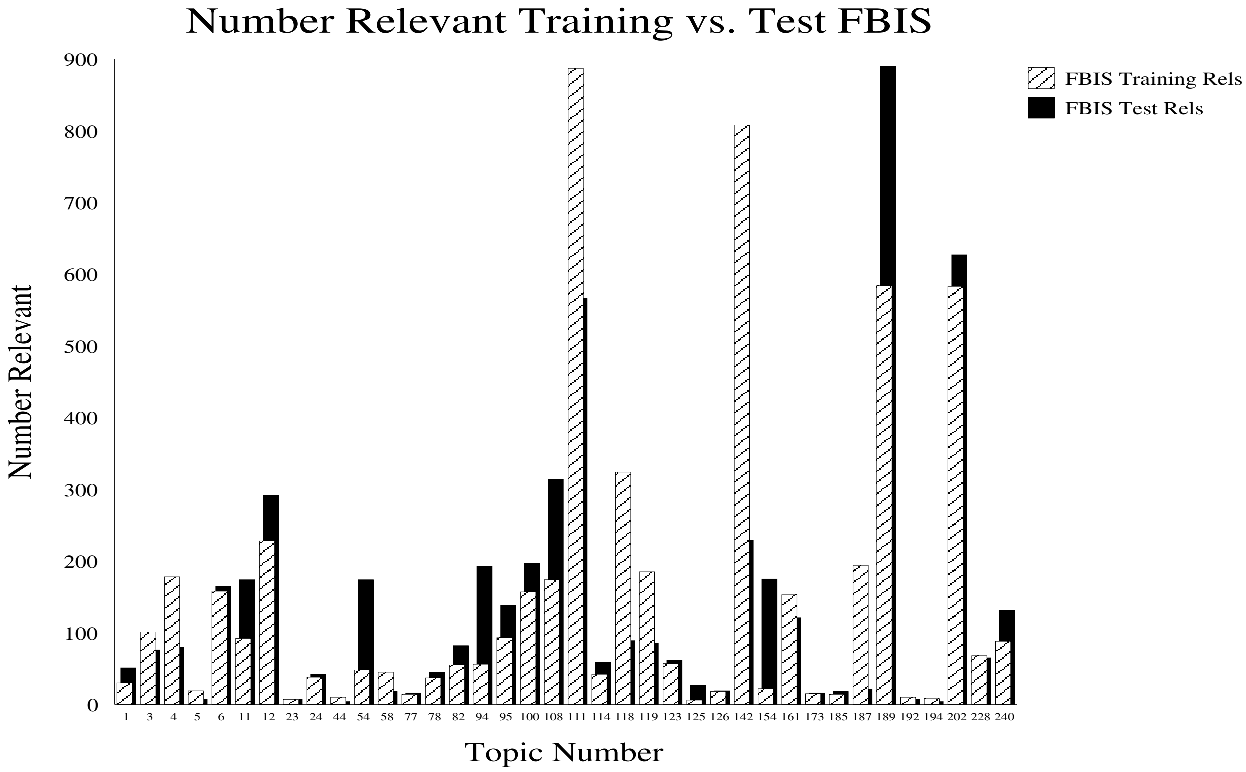

Figure 9: Comparison of the number of relevant documents in the training and test FBIS collections.

with the highest mean average precision for the routing queries. The runs are ranked by the mean average precision over the 47 topics. Manual and automatic runs are combined in the graph since previous TREC experiments have shown no benefits from manual construction of routing queries when large amounts of relevance judgments are available. Indeed, the top four runs in Figure 10 are automatic runs.

The following runs are represented in the graph:

att97re – from AT&T Labs Research (Singhal, 1998) who added the machine learning technique of boosting to the query refinement phase of the Cornell TREC-5 routing algorithm (which includes the use of word pairs, DFO optimization, and query zones).

city6a1 – from City University, London (Walker et al., 1998) who explored iterative methods of term weighting with the goal of avoiding overfitting.

Cor6A3cl – from Cornell/SaBIR Research (Buckley et al., 1998) who also used a variant of the basic Cornell TREC-5 routing approach, adding SuperConcepts to the routing query.

pirc7R2 – from Queens College, CUNY (Kwok et al., 1998) who combined the of results from five separate component runs; this combined result is superior to each of the individual components.

uwmt6r1 – from the University of Waterloo (Cormack et al., 1998) who interactively refined a set of Boolean queries into a single tiered Boolean query for each topic.

CLMAX – from Claritech Corporation (Milic-Frayling et al., 1998) who explored the benefits of using different term selection methods in different parts of the query refinement process. For this run they developed different queries using different term selection strategies and then, for each topic, selected the query that performed the best on the training data.

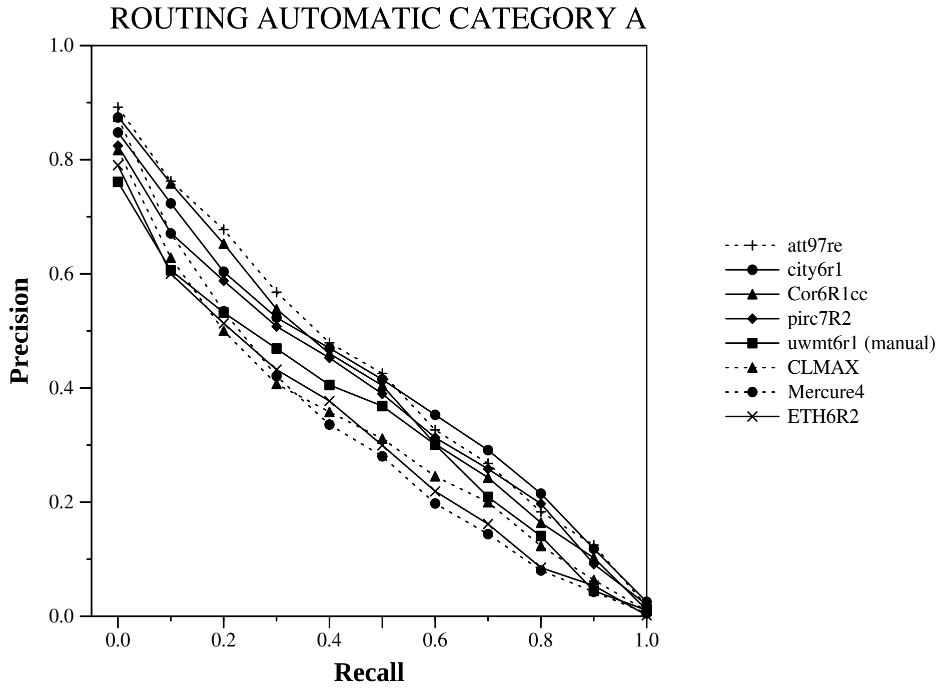

Figure 10: Recall-Precision graph for the top eight routing runs.

Mercure1 – from MSI/IRIT/SIG/CERISS (Boughanem & Soulé-Dupuy, 1998) who continued their work with a spreading activation model by expanding queries with the top 30 terms from relevance back-propagation.

ETH6R2 – from the Swiss Federal Institute of Technology (ETH) (Mateev, Munteanu, Sheridan, Wechsler, & Schäuble, 1998) who also performed a combination run where one component run selected query word and phrases based on the U-measure.

The best mean average precision for a routing run in TREC-6 was .420, a 9% improvement over TREC-5's best of .386. However, given that the TREC-6 task was designed to use a homogeneous data set whereas the TREC-5 test data were different from the training data, a greater improvement was expected. At this point, it is unclear why the difference is not greater than it is. It is possible that while the numbers of relevant documents in the training and test set are comparable, the relevant documents in each set don't "look like" each other. However, this is unlikely since both sets of documents come from a common source. It is also possible that the mismatch between training and test sets is not as significant a factor as was thought.

Another hypothesis suggested by Singhal (1998) is that the relevance judgments are less consistent for routing than they are for the ad hoc task, and that this inconsistency prevents the machine learning methods that are prevalent in the task from performing well. Since some routing topics have been used many times, and therefore have relevance judgments spanning many years, the judgments are likely to be less consistent than for the ad hoc task. It may be instructive to explore the stability of the routing techniques in the face of different relevance judgments, especially given that real user judgments are known to be extremely volatile (Schamber, 1994).

The routing guidelines allowed participants to use any/all of the relevance judgments available for a topic in the training for that topic. In contrast, the filtering track (described below) specified that training could only be done with previous FBIS judgments. Some groups ran routing experiments comparing the results from the two different training sets, and reached contradictory conclusions regarding which was better. For example, the Daimler Benz group concluded that using all of the training examples made the training set too unlike the test set for their classifier and using only the FBIS examples would be better (Bayer, Mogg-Schneider, Renz, & Schäfer, 1998). CSIRO concluded precisely the opposite: that the FBIS training examples were too limiting and the variety introduced by judgments on other sources improved their results

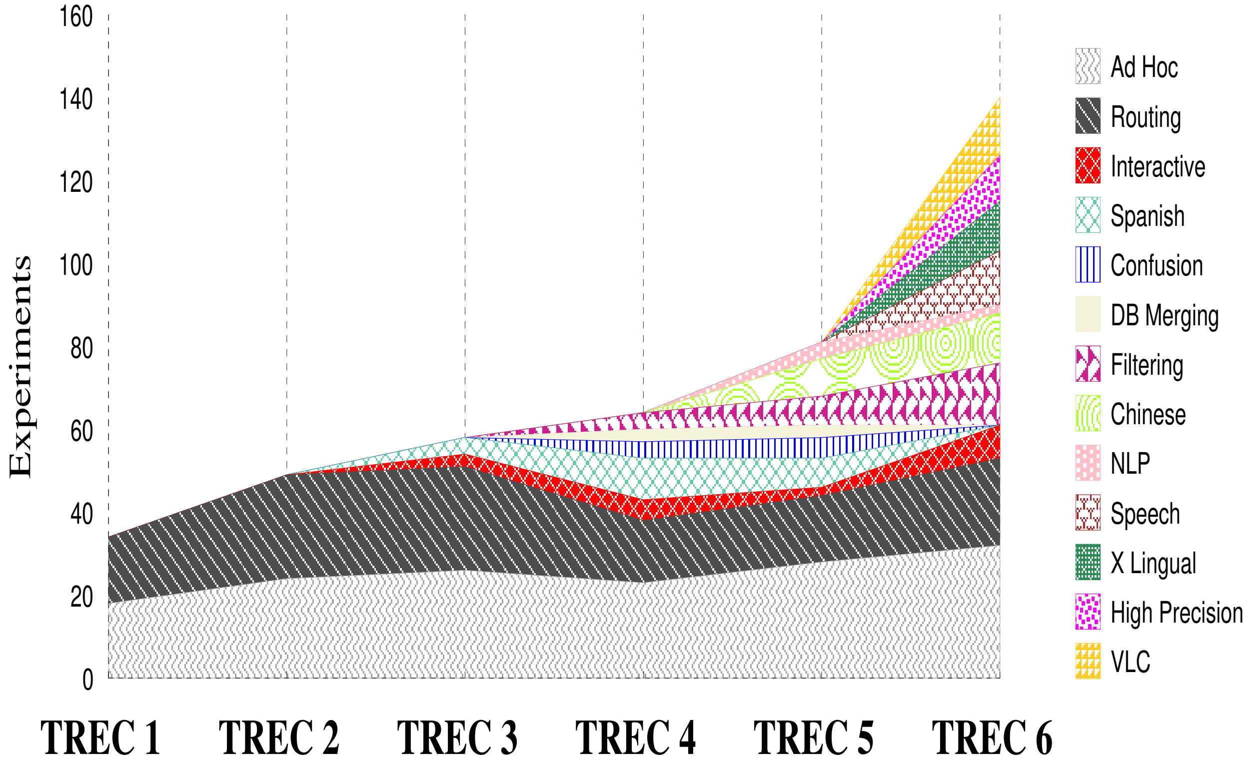

Figure 11: Number of TREC experiments by TREC task.

(Kosmynin, 1998). Verity suggested a compromise of using all the judgments while emphasizing a particular source (Pedersen, Silverstein, & Vogt, 1998). The optimum trade-off between specificity and generality of the training data is clearly different for different techniques, and should be explored further.

7 The Tracks

The main TREC tasks provide an entry point for new participants and provide a baseline of retrieval performance. A set of additional tasks called "tracks", each focused on a particular retrieval subproblem, invigorates TREC by encouraging research in new areas of information retrieval. Tracks were first introduced into TREC in TREC-4 (1995) and have been a primary source of growth in TREC. Figure 11 shows how the number of experiments has grown in each TREC. (The set of retrieval runs submitted for a track by one group is counted as one experiment.)

The set of tracks run in any particular year depends on the interests of the participants and sponsors, as well as on the suitability of the problem to the TREC environment. Some initial tracks have been discontinued because the goals of the track were met. For example, the Spanish track, an ad hoc task in which both topics and documents are in Spanish, was discontinued when the results demonstrated that current retrieval systems can retrieve Spanish documents as effectively as English documents. Other tracks, such as the interactive track, have been run each year, but have changed their focus in different years. This section provides a brief summary of the goals and results of the tracks that ran in TREC-6. More detailed analyses of the tracks can be found in the track reports in the TREC-6 proceedings.

7.1 The Chinese track

The purpose of the Chinese track is to investigate retrieval performance for a language whose orthographics is not word-oriented. Participants performed an ad hoc search in which both the topics and the documents were in Chinese. Twenty-six new topics were created for the track. The document set was a collection of articles selected from the *Peoples Daily* newspaper and the Xinhua newswire created for the TREC-5 track.

Twelve groups submitted runs to the track. The majority of the experiments compared different methods of segmenting the text into retrieval features. In general, approaches that used single characters or bi-grams as features were competitive with word-based approaches and have the advantage of not requiring complicated segmentation schemes.

A confounding factor in the analysis of the retrieval results is that the retrieval effectiveness was quite high (the median mean average precision was greater than .5), and was similar across systems. It is difficult to distinguish more effective techniques when all techniques appear to work equally as well. Without more testing, it is not possible to determine whether the TREC-6 topics were simply easy, or if there is something inherent in Chinese that facilitates retrieval.

7.2 The cross-language (CLIR) track

The cross-language information retrieval task is an ad hoc task in which the focus is on searching for documents in one language using topics in a different language. Three document sets were used in the track: a set of French documents from the Swiss news agency *Schweizerische Depeschen Agentur* (SDA); a set of German documents from SDA plus a set of articles from the newspaper *Neue Zuercher Zeitung* (NZZ); and a set of English documents from the AP newswire. All of the document sets contain news stories from approximately the same time period, but are not aligned or specially coordinated with one another. A set of 25 topics was created by NIST assessors for the track. The authors of the topics created English, French, and German versions of the topics (these were translations of one another). In addition, participants contributed Spanish and Dutch translations of the topics. Participants searched for documents in one target language using topics written in a different language. In addition, participants were asked to perform a monolingual run in the target language to act as a baseline.

Thirteen groups participated in the CLIR track. Three major approaches to cross-language retrieval were represented: machine translation, where either the topics or the documents were translated into the target language; the use of machine-readable bilingual dictionaries or other existing linguistic resources; and the use of corpus resources to train or otherwise enable the cross-language retrieval mechanism. The approaches all behaved similarly in that some group obtained good cross-language performance for each method. In general, the best cross-language performance was between 50%–75% as effective as a quality monolingual run.

The TREC-6 CLIR track was the first large-scale experimental evaluation of cross-language retrieval, and as such the results should be viewed as preliminary. The test collection created in the track is an important contribution that will enable continued research in cross-language retrieval.

7.3 The filtering track

The filtering task is a routing task in which systems are forced to make a binary decision as to whether the current document should be retrieved. Making an immediate decision to retrieve a document is more difficult than ranking a set of documents, but it is also a more realistic version of the routing problem. The TREC-6 filtering track used the same topics and test data as the TREC-6 routing task. Unlike the routing task, filtering participants could train their systems using only previous FBIS data, instead of all available relevance assessments, and processed the test data in time-stamp order.

The question of how to evaluate filtering runs has been a focus of the filtering track since its inception. Since filtering results are an unordered set of documents, the rank-based measures used in the ad hoc and routing tasks are not appropriate. The main approach used in TREC is to use utility functions as measures of the quality of the retrieved set—the quality is computed as a function of the benefit of retrieving a relevant document and the cost of retrieving an irrelevant document. In TREC-6 two different utility functions were used:

$$\begin{aligned} \mathrm{F1} &= 3\mathrm{R}^+ - 2\mathrm{N}^+ \\ \mathrm{F2} &= 3\mathrm{R}^+ - \mathrm{N}^+ - \mathrm{R}^- \end{aligned}$$

where R^+ is the number of relevant documents that are retrieved, R^- is the number of relevant documents that are not retrieved, and N^+ is the number of non-relevant documents that are retrieved.

A problem with utilities as measures is that different topics have widely varying possible utility values, and these utilities cannot be normalized. Thus, utilities cannot be meaningfully averaged or compared across topics. A second measure, *average set precision* (ASP) defined as the product of recall and precision, was therefore introduced in TREC-6. Unfortunately, ASP suffers from its own drawback. When no relevant documents are retrieved, ASP is 0 regardless of how many non-relevant documents are retrieved. This is a problem in filtering evaluation since knowing when to not retrieve documents is an important part of the filtering task.

The F1 utility measure defined above rewards a system with 3 "points" for every relevant document document it retrieves, and penalizes the system two "points" for every nonrelevant document it retrieves. While these benefit and cost values seem reasonable, they define a level of performance that is quite challenging for current systems to meet. Ten groups participated in the TREC-6 filtering track and submitted a total of 17 runs that were optimized for the F1 measure. The best of these runs had a positive utility for 33 (of 47) topics, and the median of the 17 runs had a positive utility for just 20 topics. Since retrieving no documents has an F1 utility of 0, retrieving no documents would result in a better F1 utility than current systems obtain on average. The F2 utility measure is even more demanding since systems are penalized for not retrieving relevant documents (and thus retrieving no documents also results in a negative utility). Of the 17 runs optimized for the F2 utility, only 10 topics had a positive median F2 utility. The track will continue to explore the relationship between utility measures and user satisfaction.

7.4 The high precision (HP) track

TREC-6 was the first running of the high precision track. The task in the track was to retrieve ten relevant documents for a topic within five minutes (wall clock time). Users could not collaborate on a single topic, nor could the system (or user) have previous knowledge of the topic. Otherwise, the user was free to use any available resources as long as the five minute time limit was observed. The task is an abstraction of a common retrieval problem: quickly find a few good documents to get a feel for the topic area.

Since the track guidelines put no limits on who the user could be, an implicit assumption of the track is that the runs were performed by system experts. As such, the track provides an upper-bound on the effectiveness obtainable by the systems. The 5-minute time limit was selected so that the intrinsic effectiveness of the system, the system efficiency, and the user interface would all be tested by the task. The same 50 topics and document set as used in the TREC-6 ad hoc task were used for the HP track.

Five groups participated in the HP track, submitting a total of 13 runs. The mean over 50 topics of the precision after ten documents were retrieved ranged from a high .6020 to a low of .3360. The least effective runs were a set of completely automatic runs submitted to see how automatic runs would fare; the results confirm that user involvement is indeed beneficial. However, the best result was a run in which the user simply provided yes/no relevance judgments as input for a sophisticated (automatic) relevance feedback algorithm. This suggests that user involvement does not need to be extensive to get good retrieval results.

7.5 The interactive track

One of the first tracks to be started in TREC, the interactive track studies text retrieval systems in interaction with users and is interested in the process as well as the results. Effectively supporting the users of a retrieval system has become an increasingly important problem as more and more text is made electronically accessible, and larger numbers of end users (as opposed to a relatively small group of trained intermediaries) perform searches. Yet designing retrieval experiments that can be fairly evaluated and that produce interpretable results when humans are included in the loop is especially challenging since it is difficult to isolate the effects of the different factors that contribute to overall effectiveness.

Interactive experiments include a third factor, the searcher, to the topic and retrieval system factors inherent in all retrieval experiments. An ideal experimental design tests all combinations of all settings of all factors with repetitions, but with human subjects such a design is not feasible within a single site and certainly not across sites. For example, the same user cannot perform a search for a topic more than once because the experience gained during the first search biases the second search, but logistics prevents randomly assigning searchers from one site to perform searches on another site's experimental system. Finding a sufficient number of subjects is also difficult: experience indicates that reliably detecting significant system effects requires relatively many searches. Unfortunately, reducing the number of required searches by narrowing the focus of the investigation makes generalizing any conclusions difficult.

The goal of the TREC-6 interactive track was to compare systems across sites. To this end, the track developed and employed a new method for comparing interactive IR systems across different sites (Over, 1998). The method involved comparing the particular retrieval system used at a site (an *experimental system*) to a common *control system* that was also run at each site. The direct comparison between the experimental and control systems was used to derive a measure of how much better the experimental system was than the control, independent of topic, searcher, and any other site-specific effects. Different experimental systems could then be indirectly compared across sites relative to the common control.

The experiment used six slightly modified ad hoc topics and the *Financial Times* newspaper data as the document collection. The searcher task involved six searches (three on control, three on the experimental system) to find and save documents which taken together contained as many answers as possible to the question stated or implied by the topic. Nine participants used this experimental framework to pursue their own research goals, and to contribute data to a cross-site comparison of systems. The evaluation measures used were recall and precision defined in terms of the set of all possible answers as determined by NIST assessors. Participants also reported extensive data on the characteristics of each searcher and of each searcher's interactions with both the control and experimental system.

As a first step in analyzing the cross-site data, the best model for each site's results in terms of which factors and interactions to include was determined. Then a cross-site analysis of variance (ANOVA) was performed, which indicated that there was a significant difference between some systems. However, a multiple comparisons test (Tukey's), run to determine which systems differed, found no significant pair-wise differences (Over, 1998).

The effectiveness of using a control system to remove the site effect from cross-site comparisons was an assumption of the track design and so could not be tested by it. Additional experiments before and after TREC-6 did address the effectiveness of the control (i.e., the equivalence of the direct and indirect comparison of systems) but neither confirmed nor refuted its effectiveness (Lagergren & Over, 1998; Swan & Allan, 1998). As a practical matter, it is difficult to justify the cost of adding a control system to an experimental design in the absence of clear positive evidence for its effectiveness.

7.6 The natural language processing (NLP) track

The NLP track was started in TREC-5 to explore whether the natural language processing techniques available today are mature enough to have an impact on IR, and specifically whether they can offer an advantage over more conventional methods. The TREC-6 track used the 50 TREC-6 ad hoc topics and a reduced document set consisting of just the *Financial Times* newspaper data. The track had limited participation, with just two groups submitting NLP runs.

To date, specific NLP processing has not proved essential to obtaining effective retrieval in TREC. The most useful NLP techniques for text retrieval generally have been methods that recognize and normalize names and other multi-word terms. However, the TREC topics do not require processing at this level of detail. Other information seeking tasks such as fact extraction or story summarization may be a more appropriate test of current NLP technology.

7.7 The spoken document retrieval (SDR) track

The TREC-6 spoken document retrieval track was the first running of a track intended to foster research on retrieval methodologies for spoken documents (i.e., recordings of speech). The track is a successor to the "confusion tracks" of earlier TREC conferences, which investigated methods for retrieving document surrogates whose true content has been confused or corrupted in some way. In the SDR track, the document surrogates are produced by speech recognition systems.

The document set was a set of transcripts from 50 hours of broadcast news originally collected by the Linguistic Data Consortium for DARPA Hub-4 speech recognition evaluations (Garofolo, Fiscus, & Fisher, 1997). Three versions of the transcripts were used: a "truth" transcript that was hand-produced; a transcript produced by an IBM baseline speech recognition system; and a transcript produced by the participant's own speech recognition system. Document boundaries were given in the hand-produced transcript, and the same boundaries were used in the other two versions. While recognizing fifty hours of news presented a serious challenge, the resulting document set was small by retrieval standards, consisting of only 1451 stories.

Like the earlier confusion tracks, the task in the SDR track was a *known-item search*. In a known-item search, the goal is to retrieve a single specific document, rather than a set of relevant documents. The search simulates a user seeking a particular, half-remembered document. NIST created 50 topics each designed to describe precisely one document. Half of the topics were created to target speech conditions, and half to target retrieval conditions. Within each half, half were designed to be easy and half difficult. Difficult speech conditions included background noise, non-native speakers, low-bandwidth channels, and the like. Difficult retrieval conditions included the use of synonyms (e.g., cinema for movie theater) and rare senses of common words (e.g., looking for the document describing cigarette pants when many stories were about cigarette smoking).

Thirteen groups submitted SDR track runs. The results suggest that speech recognition and IR technologies are sufficiently advanced to do a credible job of retrieving specific documents. The better systems were able to retrieve the target document at rank 1 over 70% of the time using their own recognizer transcripts, compared to the best performance on the truth transcripts of 78.7%. Search performance was a bigger factor in the overall results than recognition accuracy, although the best results were obtained by groups that included both speech and IR experts. Note, however, that the results from this track must be viewed as preliminary. The small collection size and the known-item search task limit the generality of conclusions that can be drawn.

7.8 The very large corpus (VLC) track

The Very Large Corpus track explored the effectiveness and efficiency of retrieval in collections approximately 10 times the size of the TREC ad hoc collection. The track's corpus consisted of 7.5 million texts for a total of 20.14 GB of data, including the five TREC CDs; USENET news postings; Canadian and Australian Hansards; HTML-formatted documents including university websites, and laws and judgments from the Australian Attorney General's Department; and the *Glasgow Herald* and *Financial Times* newspapers. The TREC-6 ad hoc topics were used.

Because of the difficulty of obtaining sufficient relevance judgments for recall-based measures, the main effectiveness measure used for VLC runs was precision after 20 documents were retrieved. Also reported were query response time; data structure (e.g., inverted index) building time; and a cost measure of number of queries processed per minute per hardware dollar. Participants were required to submit two runs: one run over the entire VLC corpus and a second run over a baseline collection that consisted of a random 10% sample of the full corpus. The focus of the evaluation was on the ratio of the measures between the baseline and full corpus runs.

Seven groups submitted VLC track runs. All of the participants were able to complete the VLC task with the hardware available to them (i.e., no special hardware purchases were made for the track). Indeed, the major conclusion of the track is that current systems are able to obtain good (high precision) retrieval effectiveness on a 20 GB collection with reasonable resources. For example, one of the best runs, from the University of Waterloo (Cormack et al., 1998), retrieved an average of 12.8 relevant documents in the top twenty processing at the rate of 2678 queries per hour using a cluster of four commodity PCs.

8 The Future

Evaluating competing technologies on a common problem set is a powerful way to improve the state of the art and hasten technology transfer. TREC has been able to build on the text retrieval field's tradition of experimentation to significantly improve retrieval effectiveness and extend the experimentation to new subproblems. By defining a common set of tasks, TREC focuses retrieval research on problems that have a significant impact throughout the community. The conference itself provides a forum in which researchers can efficiently learn from one another and thus facilitates technology transfer. TREC also provides a forum in which methodological issues can be raised and discussed, resulting in improved text retrieval research.

The TREC workshop series will continue for the forseeable future. The set of tasks included in a conference will vary depending on the interests of the participants and sponsors, and the suitability of the problem to the TREC environment. New tracks will be introduced as the need arises. For example, NIST is currently investigating the feasibility of creating a test collection that accurately reflects aspects of Web searching.

In addition to managing current TREC tasks, NIST will increasingly use TREC as a vehicle for creating and testing new evaluation methodologies, and for conducting deeper investigations into the many issues involved in such testing. A problem of particular interest is to identify factors that make a topic difficult in the context of a particular document collection. Current measures of topic difficulty require that the set of relevant documents be known. Unfortunately, this means the TREC collections cannot be designed to control for topic difficulty, which reduces the effectiveness of the test collections as diagnostic tools. A deeper understanding of what makes a topic difficult would allow better test collections to be built and would improve retrieval performance in its own right.

Acknowledgments

The authors gratefully acknowledge the continued support of the TREC conferences by the Intelligent Systems Office of the Defense Advanced Research Projects Agency. Thanks also go to the TREC program committee and the staff at NIST. The TREC tracks could not happen without the efforts of the track coordinators; our special thanks to them.

References

Allan, J., Callan, J., Croft, W. B., Ballesteros, L., Byrd, D., Swan, R., & Xu, J. (1998). INQUERY does battle with TREC-6. In E. Voorhees & D. Harman (Eds.), *Proceedings of the sixth Text REtrieval Conference (TREC-6)* (pp. 169–206). NIST Special Publication 500-240.

Bayer, T., Mogg-Schneider, H., Renz, I., & Schäfer, H. (1998). Daimler Benz research: System and experiments, routing and filtering. In E. Voorhees & D. Harman (Eds.), *Proceedings of the sixth Text REtrieval Conference (TREC-6)* (pp. 329–346). NIST Special Publication 500-240.

Boughanem, M., & Soulé-Dupuy, C. (1998). Mercure at trec6. In E. Voorhees & D. Harman (Eds.), *Proceedings of the sixth Text REtrieval Conference (TREC-6)* (pp. 321–328). NIST Special Publication 500-240.

Brown, E. W., & Chong, H. A. (1998). The GURU system in TREC-6. In E. Voorhees & D. Harman (Eds.), *Proceedings of the sixth Text REtrieval Conference (TREC-6)* (pp. 535–540). NIST Special Publication 500-240.

Buckley, C., Mitra, M., Walz, J., & Cardie, C. (1998). Using clustering and SuperConcepts within SMART: TREC-6. In E. Voorhees & D. Harman (Eds.), *Proceedings of the sixth Text REtrieval Conference (TREC-6)* (pp. 107–124). NIST Special Publication 500-240.

Cormack, G. V., Clarke, C. L., Palmer, C. R., & To, S. S. L. (1998). Passage-based refinement (MultiText experiments for TREC-6. In E. Voorhees & D. Harman (Eds.), *Proceedings of the sixth Text REtrieval Conference (TREC-6)* (pp. 303–319). NIST Special Publication 500-240.

Cormack, G. V., Palmer, C. R., & Clarke, C. L. (1998). Efficient construction of large test collections. In W. B. Croft, A. Moffat, C. van Rijsbergen, R. Wilkinson, & J. Zobel (Eds.), *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 282–289). Melbourne, Australia: ACM Press, New York.

Croft, W. B., Moffat, A., Rijsbergen, C. van, Wilkinson, R., & Zobel, J. (Eds.). (1998, August). *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval.* Melbourne, Australia: ACM Press, New York.

Fuller, M., Kaszkiel, M., Ng, C. L., Vines, P., Wilkinson, R., & Zobel, J. (1998). MDS TREC6 report. In E. Voorhees & D. Harman (Eds.), *Proceedings of the sixth Text REtrieval Conference (TREC-6)* (pp. 241–257). NIST Special Publication 500-240.

Garofolo, J., Fiscus, J., & Fisher, W. (1997). Design and preparation of the 1996 Hub-4 broadcast news benchmark test corpora. In *Proceedings of the DARPA speech recognition workshop* (pp. 15–21).

Gey, F. C., & Chen, A. (1998). Phrase discovery for English and cross-language retrieval at TREC-6. In E. Voorhees & D. Harman (Eds.), *Proceedings of the sixth Text REtrieval Conference (TREC-6)* (pp. 637–647). NIST Special Publication 500-240.

Harman, D. (1994). Analysis of data from the second Text REtrieval Conference (TREC-2). In *Proceedings of RIAO94* (pp. 699–709).

Harman, D. (1995). Special issue: The second Text REtrieval Conference (TREC-2). *Information Processing and Management*, *31*(3).

Harman, D. (1996). Overview of the fourth Text REtrieval Conference (TREC-4). In D. K. Harman (Ed.), *Proceedings of the fourth text REtrieval conference (TREC-4)* (pp. 1–23). (NIST Special Publication 500-236.)

Hawking, D., Thistlewaite, P., & Craswell, N. (1998). ANU/ACSys TREC-6 experiments. In E. Voorhees & D. Harman (Eds.), *Proceedings of the sixth Text REtrieval Conference (TREC-6)* (pp. 275–290). NIST Special Publication 500-240.

Kosmynin, A. (1998). CSIRO routing and ad-hoc experiments at TREC6. In E. Voorhees & D. Harman (Eds.), *Proceedings of the sixth Text REtrieval Conference (TREC-6)* (pp. 455–460). NIST Special Publication 500-240.

Kwok, K., Grunfeld, L., & Xu, J. (1998). TREC-6 English and Chinese retrieval experiments using PIRCS. In E. Voorhees & D. Harman (Eds.), *Proceedings of the sixth Text REtrieval Conference (TREC-6)* (pp. 207–214). NIST Special Publication 500-240.

Lagergren, E., & Over, P. (1998). Comparing interactive information retrieval systems across sites: The TREC-6 interactive track matrix experiment. In W. B. Croft, A. Moffat, C. van Rijsbergen, R. Wilkinson, & J. Zobel (Eds.), *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 164–172). Melbourne, Australia: ACM Press, New York.

Leong, M.-K. (1998). Concrete queries in specialized domains: Known item feedback for query formulation. In E. Voorhees & D. Harman (Eds.), *Proceedings of the sixth Text REtrieval Conference (TREC-6)* (pp. 541–549). NIST Special Publication 500-240.

Lu, A., Meier, E., Rao, A., Miller, D., & Pliske, D. (1998). Query processing in TREC6. In E. Voorhees & D. Harman (Eds.), *Proceedings of the sixth Text REtrieval Conference (TREC-6)* (pp. 567–576). NIST Special Publication 500-240.

Lundquist, C., Holmes, D. O., Grossman, D. A., Frieder, O., & McCabe, M. C. (1998). Expanding relevance feedback in the relational model. In E. Voorhees & D. Harman (Eds.), *Proceedings of the sixth Text REtrieval Conference (TREC-6)* (pp. 489–502). NIST Special Publication 500-240.

Mateev, B., Munteanu, E., Sheridan, P., Wechsler, M., & Schäuble, P. (1998). ETH TREC-6: Routing, Chinese, cross-language, and spoken document retrieval. In E. Voorhees & D. Harman (Eds.), *Proceedings of the sixth Text REtrieval Conference (TREC-6)* (pp. 623–635). NIST Special Publication 500-240.

Milic-Frayling, N., Zhai, C., Tong, X., Jansen, P., & Evans, D. A. (1998). Experiments in query optimization: The CLARIT system TREC-6 report. In E. Voorhees & D. Harman (Eds.), *Proceedings of the sixth Text REtrieval Conference (TREC-6)* (pp. 415–454). NIST Special Publication 500-240.

Over, P. (1998). TREC-6 interactive track report. In E. Voorhees & D. Harman (Eds.), *Proceedings of the sixth Text REtrieval Conference (TREC-6)* (pp. 73–81). NIST Special Publication 500-240.

Pedersen, J. O., Silverstein, C., & Vogt, C. C. (1998). Verity at TREC-6: Out-of-the-box and beyond. In E. Voorhees & D. Harman (Eds.), *Proceedings of the sixth Text REtrieval Conference (TREC-6)* (pp. 259–273). NIST Special Publication 500-240.

Schamber, L. (1994). Relevance and information behavior. *Annual Review of Information Science and Technology*, *29*, 3–48.

Singhal, A. (1998). AT&T at TREC-6. In E. Voorhees & D. Harman (Eds.), *Proceedings of the sixth Text REtrieval Conference (TREC-6)* (pp. 215–225). NIST Special Publication 500-240.

Sparck Jones, K. (1995). Reflections on TREC. *Information Processing and Management*, *31*(3), 291–314.

Sparck Jones, K., & van Rijsbergen, C. (1975). *Report on the need for and provision of an "ideal" information retrieval test collection.* British Library Research and Development Report 5266. Computer Laboratory, University of Cambridge.

Sparck Jones, K., & Willett, P. (Eds.). (1997). *Readings in information retrieval.* Morgan Kaufmann Publishers, Inc., San Francisco, CA.

Strzalkowski, T., Lin, F., & Perez-Carballo, J. (1998). Natural language information retrieval: TREC-6 report. In E. Voorhees & D. Harman (Eds.), *Proceedings of the sixth Text REtrieval Conference (TREC-6)* (pp. 347–366). NIST Special Publication 500-240.

Swan, R. C., & Allan, J. (1998). Aspect windows, 3-d visualizations, and indirect comparisons of information retrieval systems. In W. B. Croft, A. Moffat, C. van Rijsbergen, R. Wilkinson, & J. Zobel (Eds.), *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 173–181). Melbourne, Australia: ACM Press, New York.

Voorhees, E., & Harman, D. (Eds.). (1998, August). NIST Special Publication 500-240.

Voorhees, E. M. (1998). Variations in relevance judgments and the measurement of retrieval effectiveness. In W. B. Croft, A. Moffat, C. van Rijsbergen, R. Wilkinson, & J. Zobel (Eds.), *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 315–323). Melbourne, Australia: ACM Press, New York.

Walker, S., Robertson, S., Boughanem, M., Jones, G., & Sparck Jones, K. (1998). Okapi at TREC-6: Automatic ad hoc, VLC, routing, filtering and QSDR. In E. Voorhees & D. Harman (Eds.), *Proceedings of the sixth Text REtrieval Conference (TREC-6)* (pp. 125–136). NIST Special Publication 500-240.

Zobel, J. (1998). How reliable are the results of large-scale information retrieval experiments? In W. B. Croft, A. Moffat, C. van Rijsbergen, R. Wilkinson, & J. Zobel (Eds.), *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 307–314). Melbourne, Australia: ACM Press, New York.