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

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

1 Introduction

The sixth Text REtrieval Conference (TREC-6) was held at the National Institute of Standards and Technology (NIST) on November 19–21, 1997. The conference was co-sponsored by NIST and the Information Technology Office of the Defense Advanced Research Projects Agency (DARPA) as part of the TIPSTER Text Program.

TREC-6 is the latest in a series of workshops designed to foster research in text retrieval. For analyses of the results of previous workshops, see Sparck Jones [6], Tague-Sutcliffe and Blustein [8], and Harman [2]. In addition, the overview paper in each of the previous TREC proceedings summarizes the results of that TREC.

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.

Table 1 lists the groups that participated in TREC-6. Fifty-one groups including participants from 12 different countries and 21 companies were represented. The diversity of the participating groups has ensured that TREC represents many different approaches to text retrieval. The emphasis on individual experiments evaluated within a common setting has provenpto be a major strength of TREC.

This paper serves as an introduction to the research described in detail in the remainder of the volume. The next section defines the common retrieval tasks performed in TREC-6. Sections 3 and 4 provide details regarding the test collections and the evaluation methodology used in TREC. Section 5 provides an overview of the retrieval results. The final section summarizes the main themes learned from the experiments.

2 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. TREC-6 continued four tracks from previous years and introduced four new tracks. This section describes the goals of the two main tasks in detail, and outlines the goals of each of the tracks. Readers are urged to consult the appropriate track report found later in these proceedings for details about individual tracks.

2.1 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 topics that have known relevant documents and testing on a completely new document set.

The training for the routing task is shown in the left-hand column of Figure 1. 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. 3.2 for details). The topics are used to create a set of queries (the actual input to the retrieval system) that are then used against

Table 1: Organizations participating in TREC-6.

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	

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 47 routing topics and run against the test documents.

The testing phase of the routing task is shown in the middle column of Figure 1. The output of running Q2 against the test documents is the official test result for the routing task. Due to the difficulty of obtaining appropriate data, the test and training documents were not well-matched in both TREC-4 and TREC-5. Since we wanted a good match for TREC-6, we used (mostly) the same routing topics as were used in TREC-5 for TREC-6, and obtained additional Foreign Broadcast Information Service (FBIS) documents as the test set. In particular, we included those TREC-5 routing topics that had at least six relevant documents in the TREC-5 FBIS data as TREC-6 routing topics. Additional relevance assessments were made on the TREC-5 FBIS corpus for several other topics deemed likely to have relevant documents in FBIS, and for four new topics specifically created for the track (topics 10001–10004). For these topics, the top 100 FBIS documents as retrieved by NIST's PRISE search engine were judged and those with at least six relevant were also included in the set of routing topics. The final set of routing topics contained 47 topics.

2.2 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 1 depicts how the ad hoc task is accomplished in TREC. Participants are given approximately two gigabytes worth of documents. They are also given 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. Topics 301–350

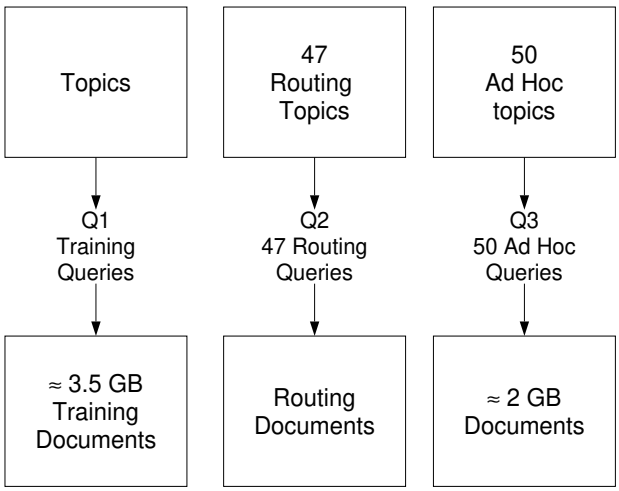

Figure 1: TREC main tasks.

were created for the TREC-6 ad hoc task. The set of documents used in the task were those contained on TREC Disks 4 and 5. See sec. 3.1 for details about this document set.

2.3 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. As in TREC-5, 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.

There are two levels of participation in TREC: category A, participation using the full dataset, or category B, participation using a reduced dataset (1/4 of the full document set). Groups could choose to do the routing task, the ad hoc task, or both, and were asked to submit the top 1000 documents retrieved for each topic for evaluation. Groups that performed the routing task were allowed to submit up to two official test results for judging. When two sets of results were sent, they could be made using different methods of creating queries, or different methods of searching with the same queries. Groups that performed the ad hoc task could submit up to three runs, though if any automatic results were submitted, at least one of the runs was required to use "short" topics (see sec. 3.2).

2.4 The tracks

One of the goals of TREC is to provide a common task evaluation that allows cross-system comparisons. This has proven to be a key strength in TREC. The second major strength is the loose definition of the two main tasks, allowing a wide range of experiments. The addition of secondary tasks (tracks) in TREC-4 combined these strengths by creating a common evaluation for tasks that are either related to the main tasks, or are a more focussed implementation of those tasks. TREC participants were free to turn in results for any, or all, or none, of the tracks. Each track had a set of guidelines developed under the direction of the track coordinator. The set of tracks and their primary goals are listed below. See the track reports elsewhere in this proceedings for a more complete description of each track.

Four tracks continued from previous years and had similar goals as in those years.

Chinese: In the Chinese track, participants performed an ad hoc search in which both the topics and the documents were in Chinese. Twenty-six new topics (CH29–CH54) were created for the track, and the document set was the same as for the TREC-5 track (articles selected from the *Peoples Daily* newspaper and the Xinhua newswire).

Filtering: The filtering task is a routing task in which the system must decide whether or not to retrieve each individual document. Instead of producing a list of documents ranked according to the presumed similarity to a query, filtering systems retrieve an unordered set of documents for each query. The quality of the retrieved set is computed as a function of the benefit of a retrieved relevant document and the cost of a retrieved irrelevant document. The TREC-6 version of the track differed from its predecessors in several ways. New utility functions were introduced to assess the quality of the search. More significantly, filtering track participants could train their systems using only FBIS data (as opposed to all available relevance assessments) and processed the test data in timestamp order.

Interactive: The high-level goal of the interactive track is the investigation of searching as an interactive task by examining the process as well as the outcome. The TREC-6 track used six slightly modified ad hoc topics and the *Financial Times* 1991–1994 collection. The experiment was designed to isolate the effect of topic and searcher from that of the search system and used a common control system to remove other site-specific effects. The searcher task involved six searches (three on control, three on an 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. System comparisons were based on recall and precision defined in terms of the set of all possible answers as determined by NIST assessors. Participants also reported extensive data on each searcher's interactions with both the control and experimental system.

NLP: The NLP track was initiated to explore whether the natural language processing (NLP) techniques available today are mature enough to have an impact on IR, and specifically whether they can offer an advantage over purely quantitative retrieval methods. The track used the 50 ad hoc topics and the *Financial Times* document set.

The remaining four tracks were introduced in TREC-6.

Cross Language (CLIR): The CLIR 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 *New Zurich Newspaper* (NZZ); and 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 twenty-five topics were 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.

High Precision (HP): The goal of the high precision track was to test the effectiveness, efficiency, and user interface of participating systems. Participants used the same 50 topics and document set as the ad hoc task. For each topic, a user was given the query and asked to find 10 documents that answer the 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.

Spoken Document Retrieval (SDR): The TREC-6 SDR 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. Participants performed known-item searches using three versions of the documents. The documents were transcripts of radio broadcast news shows: a "truth" transcript that was hand-produced, a transcript produced by a baseline speech recognition system, and a transcript produced by the participant's own speech recognition system.

Very Large Corpus (VLC): The VLC track explored the effectiveness and efficiency of retrieval in collections approximately 10 times the size of a normal TREC collection. The track's corpus consisted of 7.5 million texts for a total of 20.14GB of data. The TREC-6 ad hoc topics were used. Participants were evaluated on precision of the top 20 retrieved; query response time; data structure building time; and a cost measure of queries/minute/dollar (number of queries processed per minute per hardware dollar).

3 The Test Collections

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

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

3.1 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. Additional new 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. 2). 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 has been to preserve as much of the original structure as possible, while providing enough consistency to allow simple decoding of the data. Both as part of the philosophy of leaving the data as close to the original as possible, and because it is impossible to check all the data manually, many "errors" remain in the data. The error-checking done at NIST has concentrated on allowing readability of the data rather than on correcting content. This means that there have been automated checks for control characters, special symbols, foreign language characters, for correct matching of the begin and end document tags, and for complete "DOCNO" fields (the field that gives the unique TREC identifier for the document). The types of "errors" remaining include fragment sentences, strange formatting around tables or other "non-textual" items, misspellings, etc.

The data on disk 5 and the FBIS routing test data are new TREC document sets (although the FBIS data on disk 5 was used as routing test data in TREC-5). The Foreign Broadcast Information Service provides (English translations of) selected non-U.S. broadcast and print publications. The documents on disk 5 were mostly from the early 1990's, and those used in the routing test data were mostly from the mid 1990's. The documents were provided

```
<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 2: A document extract from the *Financial Times*.

for TREC use by the Foreign Broadcast Information Service. The *LA Times* documents are a sample of the articles that appeared in the newspaper in 1989 and 1990. The articles are used by permission of the LA Times and were obtained for TREC use by Lexis-Nexis.

3.2 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 3 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 [3]. 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. A sample TREC-6 topic is shown in Figure 4, while Table 3 summarizes the length of the topics as measured by number of words.

3.2.1 Building topic statements

Ad hoc topics have been constructed by the same person who performed the relevance assessments for that topic since TREC-3. For TREC-6, NIST introduced a new procedure for developing topics with the hope that the resulting topics would strike a good balance

```
<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 3: 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 4: A sample TREC-6 topic.

Table 3: Topic length statistics by topic section. Lengths count number of tokens in topic statement including stop words.

	Min	Max	Mean
TREC-1 (51–100)	44	250	107.4
title	1	11	3.8
description	5	41	17.9
narrative	23	209	64.5
concepts	4	111	21.2
TREC-2 (101–150)	54	231	130.8
title	2	9	4.9
description	6	41	18.7
narrative	27	165	78.8
concepts	3	88	28.5
TREC-3 (151–200)	49	180	103.4
title	2	20	6.5
description	9	42	22.3
narrative	26	146	74.6
TREC-4 (201–250)	8	33	16.3
description	8	33	16.3
TREC-5 (251–300)	29	213	82.7
title	2	10	3.8
description	6	40	15.7
narrative	19	168	63.2
TREC-6 (301–350)	47	156	88.4
title	1	5	2.7
description	5	62	20.4
narrative	17	142	65.3

between topics as diagnostic tools (i.e., neither too difficult nor too easy) and topics as realistic user inquiries.

The assessors came to NIST with an initial topic statement. These statements were prepared at home, and were treated as a user's statement of the information he or she was seeking. The statements usually reflected some consideration regarding the subject areas likely to be covered in the target documents, but otherwise were a simple description of the needed information without regard to retrieval system capabilities or document collection peculiarities.

Using these initial topic statements, the assessors explored (a subset of)[1] the TREC-6 ad hoc collection using NIST's PRISE retrieval system. There were two aims of the collection exploration phase: estimating the number of relevant documents in the collection and evaluating whether the topic could be judged consistently in the assessment phase. The assessors formed an initial PRISE query and judged the top 25 documents for relevance. If the top 25 contained no relevant documents or more than 20 relevant documents, the topic was abandoned. If the top 25 contained more than 5 but fewer than 21 relevant documents, the assessor continued to judge 75 more documents for a total of 100 documents judged. Finally, if the top 25 contained at least 1 relevant document but no more than 5 relevant documents, the assessor invoked the relevance feedback mechanism in PRISE, and judged the top 100 documents

[1] The collection used in the exploration phase consisted of the documents in the *Financial Times*, *LA Times*, and FBIS subcollections only. That is, the *Federal Register* and *Congressional Record* subcollections were excluded.

in the feedback result set. The total number of relevant documents found and the assessor's opinion as to how difficult the topic was to judge consistently were recorded for each topic.

The assessors came to NIST with a total of 120 candidate topics. Of those, 20 were discarded because there were no relevant documents in the top 25, and 9 were discarded because there were more than 20 relevant documents in the top 25. NIST selected 50 of the remaining 91 candidate topics based on having a range of estimated number of relevant, balancing the load across assessors, and eliminating topics that were considered difficult to judge.

Each of the final 50 topic statements were then reviewed by the assessors and NIST staff to ensure that the Narrative field of the topic statement accurately reflected how the assessor would judge documents for relevance. By judging 100 documents in the exploration phase, the assessors were able to see many of the issues they would have to deal with when assessing participants' results. Approximately five topics' Narrative fields were modified during this review, usually by removing restrictive clauses. The review also ensured that the Title field of the topics would meet the needs of those interested in exploring very short queries. Using guidelines suggested by Mark Sanderson of Glasgow University, the assessors created titles that contained up to three words that best described the topic.

3.2.2 Predicting topic difficulty

Recall that one of the goals for the TREC-6 topics was that they be neither too difficult nor too easy so they would be useful as diagnostic tools. In practice, predicting the difficulty of a topic is quite challenging. As an experiment to see whether NIST staff members could predict the difficulty of a topic based simply on the topic statement, nine members of the Natural Language Processing and Information Retrieval Group at NIST (including the authors) predicted how difficult each ad hoc topic would be. These predictions were made before the relevance assessments were performed, so the true answer could not be known at the time of prediction. Each person divided the topics into disjoint sets of easy topics, middling topics, and hard topics.

Once the relevance assessments were available and the participants' runs evaluated, a hardness measure was computed for each topic. The hardness measure used was introduced in TREC-2 and explored further in the TREC-5 Overview [9]. The hardness score for a topic T is computed as

mean Prec(100) for T	if T has 100 or more relevant documents
mean R-Prec for T	otherwise

where Prec(100) is precision at rank 100 and R-Prec is precision at rank R when there are R relevant documents. The means were computed over all Category A ad hoc submissions, including both manual and automatic runs. To arrive at "hard", "middling", and "easy" classifications of the topics, the hardness scores were sorted and divisions were made based on gaps in the hardness scores. This resulted in 12 hard topics, 11 easy topics, and 27 middling topics. These classifications were considered to be "the truth".

The Pearson correlation coefficient was computed between each person's prediction and the truth, and between different predictions. The Pearson correlation coefficient is suitable for interval values (so the difference between hard and easy was treated as more significant than the difference between middling and easy or middling and hard), and takes on a value between -1 and 1 inclusive. A value of 1 indicates perfect agreement, a value of -1 perfect disagreement, and a value of 0 chance agreement. The largest correlation between a prediction and the truth was .257, and the largest correlation between any two predictions was .387. To set this in context, in TREC-5 the (Spearman) correlation between hardness and topic number (that is, two items that have no actual correlation) was computed as .20. Thus, essentially none of the NIST staff members agreed with the truth or with one another.

This lack of agreement illustrates how little is known about what makes a topic difficult in the context of a particular document collection. Without an understanding of the factors that make a topic difficult, it is not possible to create test collections that balance difficulty (the ideal for the "diagnostic tool" test collection goal). The lack of understanding also impedes retrieval effectiveness. The Query Track, a new track to be introduced in TREC-7, was created to address this need. Each participant in the Query Track will create several different versions of queries for existing TREC topics. All participants will then run all versions of the queries. The goal of the track is to create a large enough pool of queries such that it will be possible to investigate query-dependent retrieval strategies.

3.2.3 Runs using different topic fields

As in TREC-5, 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.

These runs are tagged as "short, automatic" runs in the results section. Automatic runs that used only the Title field are tagged as "title, automatic" runs, and automatic runs that used the entire topic are tagged as "long, automatic" 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 5.

3.3 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 [7] 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.

3.3.1 Overlap

The effect of pooling can be measured by examining the overlap of retrieved documents. Table 4 summarizes the amount of overlap in the ad hoc and routing pools for each of the six TRECs. The first column in the table gives the maximum possible size of the pool. Since the top 100 documents from each run are judged, this number is usually 100 times the number of runs used to form the pool. However, in TREC-6 there were 13 High Precision runs that contributed a maximum of 10 documents each to the pool. The second column shows the number of documents that were actually in the pool (i.e., the number of unique documents retrieved in the top 100 across all judged runs) averaged over the number of topics. The percentage given in that column is the size of the actual pool relative to the possible pool size. The final column gives the average number of relevant documents in the pool and the percentage of the actual pool that was relevant. Starting in TREC-4, various tracks also contributed documents to the ad hoc or routing pools. These are broken out in the appropriate rows within Table 4. The order of the tracks is significant in the table—a document retrieved in a track listed later is not counted for that track if the document was also retrieved by a track listed earlier.

The tremendous drop in the size of the ad hoc pool between TREC-5 and TREC-6 reflects the difference in the number of runs NIST was able to assess. In TREC-5, participants were allowed to submit two manual and two automatic ad hoc runs, and all submitted runs were judged. However, many more participants submitted runs in TREC-6 than in TREC-5 and the amount of time available for assessing was two weeks shorter due to scheduling around other IR activities. Thus only one ad hoc run per participant was judged in TREC-6. (Participants were allowed to submit up to three ad hoc runs. They ranked the runs in order of preference as to which runs should be judged first when submitting the results. NIST judged every group's first choice. An investigation of the size of the pools if everyone's second choice were also merged into the pools showed that the pools would be too large for the assessors to finish in the available time.)

Table 4 shows that the average number of relevant documents per topic continues to decrease over the years. NIST has deliberately chosen more tightly-focused topics to better guarantee the completeness of the relevance assessments.

4 Evaluation

An important element of TREC is to provide a common evaluation forum. Standard recall/precision figures and some single evaluation measures have been calculated for each run and are shown in Appendix A. A detailed explanation of the measures is also included in the appendix.

Additional data about each system was collected that describes system features and system tim-

Table 4: Overlap of submitted results.

Ad Hoc				Routing			
	Possible	Actual	Relevant		Possible	Actual	Relevant
TREC-1	3300	1279 (39%)	277 (22%)	TREC-1	2200	1067 (49%)	371 (35%)
TREC-2	4000	1106 (28%)	210 (19%)	TREC-2	4000	1466 (37%)	210 (14%)
TREC-3	2700	1005 (37%)	146 (15%)	TREC-3	2300	703 (31%)	146 (21%)
TREC-4	7300	1711 (24%)	130 (08%)	TREC-4	3800	957 (25%)	132 (14%)
ad hoc	4000	1345	115	routing	2600	930	131
confusion	900	205	0	filtering	1200	27	1
dbmerge	800	77	2				
interactive	1600	84	13				
TREC-5	10,100	2671 (27%)	110 (04%)	TREC-5	3100	955 (31%)	113 (12%)
ad hoc	7700	2310	104	routing	2200	854	94
dbmerge	600	72	2	filtering	900	100	19
NLP	1800	289	3				
TREC-6	3,430	1445 (42%)	92 (06%)	TREC-6	4400	1306 (30%)	146 (11%)
ad hoc	3100	1326	89	routing	3400	979	105
NLP	200	113	2	filtering	1000	327	41
HP	130	6	1				

ing, and allows some primitive comparison of the amount of effort needed to produce the corresponding retrieval results. Due to the size of these system descriptions, they are not included in the printed version of the proceedings. The system descriptions are available on the TREC web site (http://trec.nist.gov).

5 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 in this proceedings for more details.

5.1 TREC-6 ad hoc 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 very short (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 the topic (only the description field), but in addition they could submit runs using either the very short (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 (Category A) data set. Of these, 57 used automatic construction of queries, with 12 official very short runs, 29 short runs, and 16 long runs. Seventeen groups used manual construction. There were only five Category B runs from two groups.

5.1.1 Long (full topic) automatic runs

Figure 5 shows the recall/precision curves for the eight TREC-6 groups with the highest non-interpolated average precision using automatic construction of queries for the long (full topic) version of the topics (see Appendix A of this volume for definitions of the evaluation metrics). 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

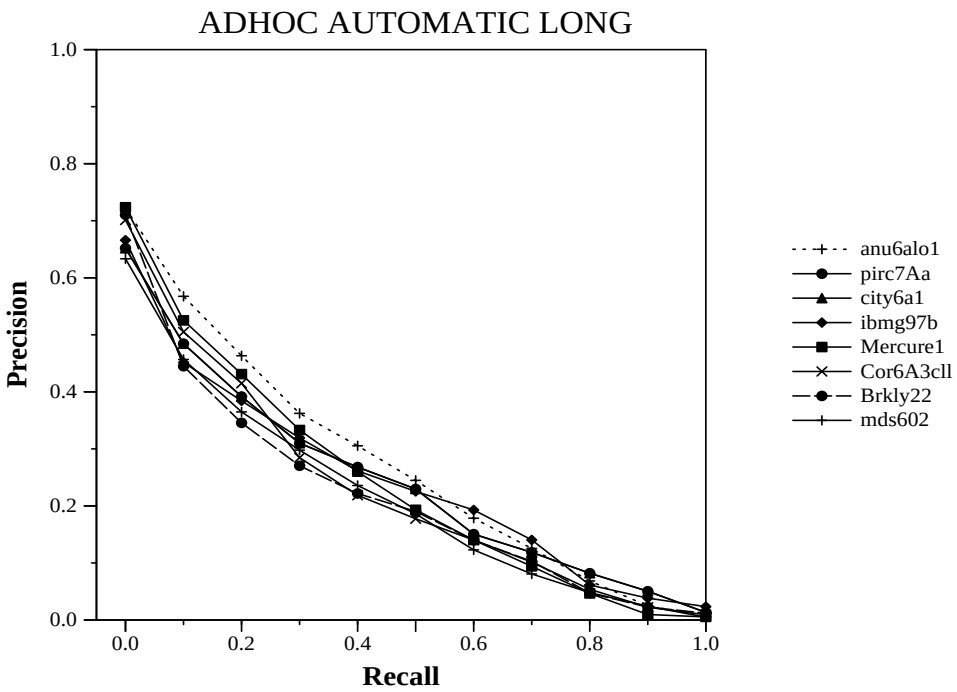

Figure 5: Recall/Precision graph for the top eight automatic ad hoc runs using the full topic.

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 in this proceedings.

anu6alo1 – Australian National University ("ANU / ACSys TREC-6 Experiments" by David Hawking, Paul Thistlewaite, and Nick Craswell) used a parallel architecture with an emphasis on efficiency. Improvements for TREC-6 include 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). These experiments included the use of "hot spots" in the top 20 documents for locating expansion terms and the use of the Robertson formula for term selection. The hot spots were defined as contiguous passages of text within a specified $p = 500$ characters of topic terms or phrases.

pirc7Aa – Queens College, CUNY ("TREC-6 English and Chinese Retrieval Experiments using PIRCS" by K.L. Kwok, L. Grunfeld, and J.H. Xu) used their spreading activation model for a two-stage search (initial search for doing pseudo-relevance feedback and a final search including expansion terms). They continued to work with 550-word subdocuments rather than dealing with multi-length documents, and generated the final score of a document as a weighted average of the scores of its three highest ranked subdocuments.

city6a1 – City University, London ("Okapi at TREC-6: automatic ad hoc, VLC, routing and filtering" by S. Walker, S.E. Robertson, and M. Boughanem) ran many experiments investigating the various parameters in the BM25 weighting technique, including adding provisions for using non-relevant documents. Additionally a new formula for selecting expansion terms that considers 500 non-relevant documents was tried. Note that the availability of many parameters for tuning allows the City group to systematically adjust their runs to specific functions. The first stage run (to get the expansion terms) was done as a high precision run; the final run was done with parameters appropriate to the length of the topic section being used. Expansion for the full topics selected the top 30 terms from the top 15 documents, multiplying weights in the original topic terms by 2.5 before doing the final retrieval. No phrases or pairs were used for TREC-6 (only single terms), however passages of between 4 and 30 paragraphs were used for the final runs only (not the initial runs for term expansion).

ibmg97b – IBM T.J. Watson Research Center ("The GURU System in TREC-6" by E. Brown and H. Chong) ran a probabilistic system that includes the use of lexical affinities (statistical phrases) in the topic. Through a series of experiments they found that performance was fairly insensitive to the distance between terms (up to a distance of 5 words was tested), but was very sensitive to the weighting of those terms. The best weight they found for these phrases was about 10% of that for the single terms in the topic. Note that no expansion was used in this run, and that only the title and description section were used as input (not the full topic).

Mercure1 – MSI/IRIT/SIG/CERISS ("Mercure at trec6" by M. Boughanem and C. Soule-Dupuy) continued their work with a spreading activation model. For TREC-6 they incorporated the Okapi/SMART BM25 weighting algorithm. Parameters in this algorithm were first set to achieve a high precision in the initial search to gather information for query expansion (similar to the City technique). Negative feedback using 500 low-ranked documents was also used in query expansion.

Cor6A3cl – Cornell/SaBIR Research ("Using Clustering and SuperConcepts Within SMART: TREC 6" by Chris Buckley, Mandar Mitra, Janet Walz, and Claire Cardie) concentrated on better initial retrieval and improved expansion. Many (unsuccessful) experiments were tried with phrases and Boolean filters, but were not used in the final run. The run for the full topics performed a clustering of candidate documents for topic expansion to help improve term selection. Note that this run did not use the title (by mistake), and the inclusion of the title gives an additional 13% improvement.

Brkly22 – University of California, Berkeley ("Phrase Discovery for English and Cross-language Retrieval at TREC-6" by Fredric C. Gey and Aitao Chen) used a probabilistic system involving heavy use of logistic regression. In TREC-6 they decided to try a new method for identifying phrases based on a mutual information measure that had been very successful in Chinese retrieval. The addition of phrases to their retrieval terms meant that the log-odds formula that they have used since TREC-2 needed to be modified to deal with the different patterns of occurrence associated with phrases as opposed to single terms. The use of phrases did not improve results in English over those that could be obtained from single terms.

mds602 – MDS/RMIT ("MDS TREC6 Report" by M. Fuller, M. Kaszkiel, C. Ng, P. Vines, R. Wilkinson, and J. Zobel) did a comprehensive factor analysis of various known successful components of retrieval, including stopwords, stemming, passage retrieval, term expansion, methods of combining results, and query length. This particular run combined four different sets of results: a baseline run, a run using the best 30 documents for expansion, a run using the best 150-word passages for expansion, and finally a run using the best 150-word passages from an already expanded query for additional expansion.

5.1.2 Short (description only) automatic runs

The method used at NIST to construct the topics for TREC-6 (discussed in section 3.2.3) caused very unusual results for the required short runs. The titles of the topics generally contained excellent topic descriptors, but for over half the topics some of these terms were not included in the description section of the topic. For many of the topics, therefore, the input to the short run consisted of a poor set of terms. Results from the "short" runs using only the description section should be viewed with great caution, therefore, and most groups redid their short runs to include the title (see individual papers for results).

However, as a way of continuing the discussion of the ad hoc results, results from the top eight short runs are shown in Figure 6.

city6ad – City University, London ("Okapi at TREC-6: automatic ad hoc, VLC, routing and filtering" by S. Walker, S.E. Robertson, and M. Boughanem) used the same methods as for the "long" run, but with different parameter settings (particularly applying less weight to negative feedback terms). Only the top ten documents were used for expansion, with 30 new terms being added.

LNaShort – Lexis-Nexis ("Query Processing in TREC-6)" by A. Lu, E. Meier, A. Rao, D. Miller, and D. Pliske) used their EUREKA toolbox to perform investigations in topic expansion and data fusion. Modified versions of two different search algorithms (Cornell's cosine Lnu.ltu and the Okapi BM25) were used along with three different methods of topic expansion

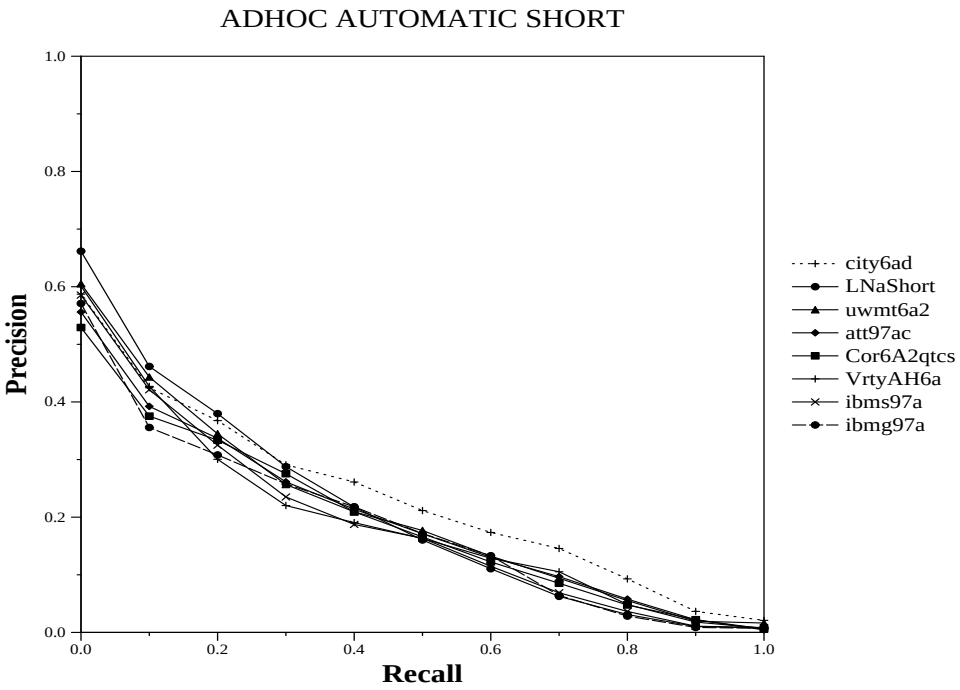

Figure 6: Recall/Precision graph for the top eight automatic ad hoc runs using the description only.

(WordNet, a Lexis-Nexis thesaurus, and Rocchio feedback) in a complex set of experiments involving merging results at several points in the process. This particular run involved first a merge of results from three runs using two weighting algorithms and WordNet or the thesaurus. The results of this were piped into a Rocchio feedback process, and then a final merge made of this output and the output of the first merging process.

uwmt6a2 – University of Waterloo ("Passage-Based Refinement (MultiText Experiments for TREC-6)" by G. Cormack, C. Clarke, C. Palmer, and S. To) made their first automatic runs in TREC-6. All the Waterloo runs used passage retrieval, with no collection-wide statistics, as the system is built for distributed architectures. The core technique for the short runs was a cover density method, which uses coordination-level matching for terms, followed by a secondary ranking using shortest substrings. The cover density technique was used to make the initial search to locate appropriate passages (of maximum length of 64 words) for use in expansion. To incorporate expansion, a modified implementation of the Okapi measure was used in the final search.

att97ac – AT&T Labs Research ("AT&T at TREC-6" by A. Singhal) is an outgrowth of the basic Cornell ad hoc approach. Two new techniques were tried in TREC-6. The first was the use of negative feedback in the Rocchio formula, based on documents ranked (in the initial search) at ranks 501-1000. This improved results from 3 to 4%. More improvement (6-7%) was gained from a new method of reweighting the topic terms and reranking the top 50 documents prior to location of expansion terms. The top 20 documents were used for this expansion, adding 25 new terms and 5 phrases.

Cor6A2qtcs – Cornell/SaBIR Research ("Using Clustering and SuperConcepts Within SMART: TREC 6" by Chris Buckley, Mandar Mitra, Janet Walz, and Claire Cardie) tried a new method called "SuperConcepts" in the short run. The idea here was to divide the expansion terms into sets clustered around the initial topic terms, and adjust their weights, with the goal of producing a more balanced query that makes maximal use of the expansion terms without skewing the query. These SuperConcepts were then used for matching against the documents rather than using an expanded set of topic terms.

VrtyAH6a – Verity, Inc. ("Verity at TREC 6: Out-

of-the-Box and Beyond" by J. Pedersen, C. Silverstein, and C. Vogt) ran a series of experiments using several tools from the Verity system. Their baseline system used a variation of tf.idf weighting, but in addition they used a commercial shallow parser to find phrases and parts of speech. They used the Verity summarizer for both length normalization and as a method of finding expansion terms for relevance feedback (5 new terms added from the top 20 documents). They also used the Verity clustering tool to help decide whether to use feedback for a given topic, based on the distribution of the top 20 documents in 5 clusters from the top 1000 documents.

ibms97a – IBM T.J. Watson Research Center ("TREC-6 Ad-Hoc Retrieval" by M. Franz and S. Roukos) used a multi-pass strategy with a combination of unigrams (single terms) and bigrams (defined as order-dependent two-word phrases). The Okapi scoring algorithm was used with different parameter settings for the unigrams and bigrams, and the scores linearly combined in the first pass. The top 40 documents from this pass were used to find expansion unigrams and bigrams, which were then used in a second pass. The final pass used expansion terms from the second pass, but combined the scores with those from the first two passes.

ibmg97a – IBM T.J. Watson Research Center ("The GURU System in TREC-6" by E. Brown and H. Chong) used the same methods as for the long run, but took only the description as input.

5.1.3 Very short (title only) automatic runs

Figure 7 shows the recall/precision curves for the eight TREC-6 groups with the highest non-interpolated average precision using automatic construction of queries for the very short (title only) version of the topics.

city6at – City University, London ("Okapi at TREC-6: automatic ad hoc, VLC, routing and filtering" by S. Walker, S.E. Robertson, and M. Boughanem) used the same methods as for the long run, but with different parameter settings (particularly, applying less weight to negative feedback terms and no weighting for query term frequency). Only the top seven documents were used for expansion, with 20 new terms being added. Weights for the original topic terms were multipled by 3.5 instead of 2.5.

pirc7Aat – Queens College, CUNY ("TREC-6 English and Chinese Retrieval Experiments using PIRCS" by K.L. Kwok, L. Grunfeld and J.H. Xu) used the same system as for the long run. However for this title version they tried (without success) an experiment in document reranking before topic expansion that used selected topic term pairs from the description.

aiatB1 – Apple Computing. No paper was submitted for this run, so nothing is known about how it was made.

uwmt6a1 – University of Waterloo ("Passage-Based Refinement (MultiText Experiments for TREC-6)" by G. Cormack, C. Clarke, C. Palmer and S. To) This run was similar to their short topic run, but the final ranked list was based on a merging of three runs: a cover density run, a run using a modified Okapi weighting and third run using a modified Okapi expansion method. The expansion used 24 new terms.

Mercure3 – MSI/IRIT/SIG/CERISS ("Mercure at trec6" by M. Boughanem and C. Soule-Dupuy) did a similar run to their long run, but took only the title as input.

att97as – AT&T Labs Research ("AT&T at TREC-6" by A. Singhal) did a similar run to their short run, but took only the title as input.

LNaVryShort – Lexis-Nexis ("Query Processing in TREC-6)" by A. Lu, E. Meier, A. Rao, D. Miller, and D. Pliske) 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 the internally-built thesaurus before a relevance feedback process (not Rocchio) was used to produce the final results.

iss97vs – Institute for Systems Science ("Concrete Queries in Specialized Domains: Known Item as Feedback for Query Formulation" by M. Leong) ran their major experiments in a manual mode. This run represents their baseline and was simply a query automatically constructed from the title. No query expansion was done.

The INQUERY system from the University of Massachusetts ("INQUERY Does Battle with TREC-6" by J. Allan, J. Callan, W.B. Croft, L. Ballesteros, D. Byrd, R. Swan, and J. Xu) did not make the above charts since they ran with only one-half the data (by mistake). Correct runs (see the paper) show that

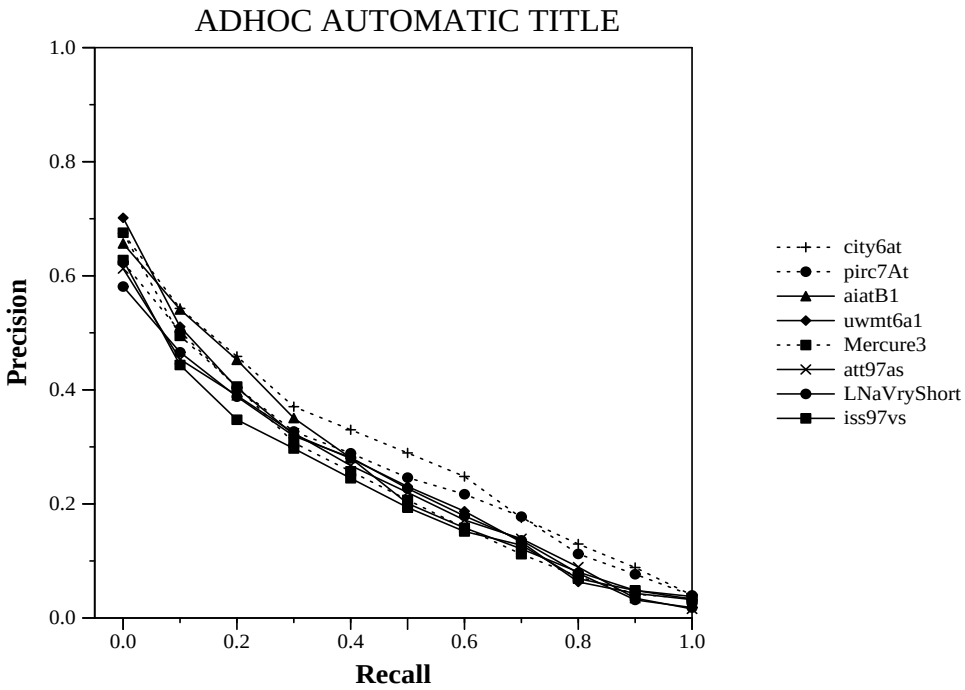

Figure 7: Recall/Precision graph for the top eight automatic ad hoc runs using the title only.

they performed as well as most of the systems shown. The basic retrieval model used in this system is a probabilistic belief network using weighting similar to the Okapi/SMART weighting, but employing complex query structures generated automatically. For TREC-6 they ran experiments in building phrase tables to help in phrase identification for input to that query structure.

5.1.4 TREC-6 ad hoc manual results

Figure 8 shows the recall/precision curves for the eight TREC-6 groups with the highest non-interpolated 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 in this proceedings.

uwmt6a0 – University of Waterloo ("Passage-Based Refinement (MultiText Experiments for TREC-6)" by G. Cormack, C. Clarke, C. Palmer, and S. To) used TREC-6 as an opportunity for experimentation on the correlation between the amount of user interaction and performance. The interfaces built for TREC-5 allowed extensive interaction with the system, and an average of 2.1 hours per topic was spent. Note that no actual ideal query was constructed during this time; the ranked list submitted to NIST was simply a list of all the documents that the searchers thought were relevant to the topic. This was done as part of an experiment in new ways of building test collections [1] rather than an investigation into manual query building.

CLAUG – CLARITECH Corp. ("Experiments in Query Optimization: The CLARIT System TREC-6 Report" by Natasa Milic-Frayling, Chengxiang Zhai, Xiang Tong, Peter Jansen, and David A. Evans) tested two different variations of relevance feedback. The searchers spent an average of about 20 minutes per topic and were constrained to constructing the initial manual query, modifying (adding/deleting/reweighting) terms based on inspecting documents, modifying Boolean query constraints (if used at all) and making relevance judgments. The CLAUG run represents a second pass using automatic pseudo-relevance feedback (from the top 50 documents) on top of a first pass (CLREL) which used 50 positive and

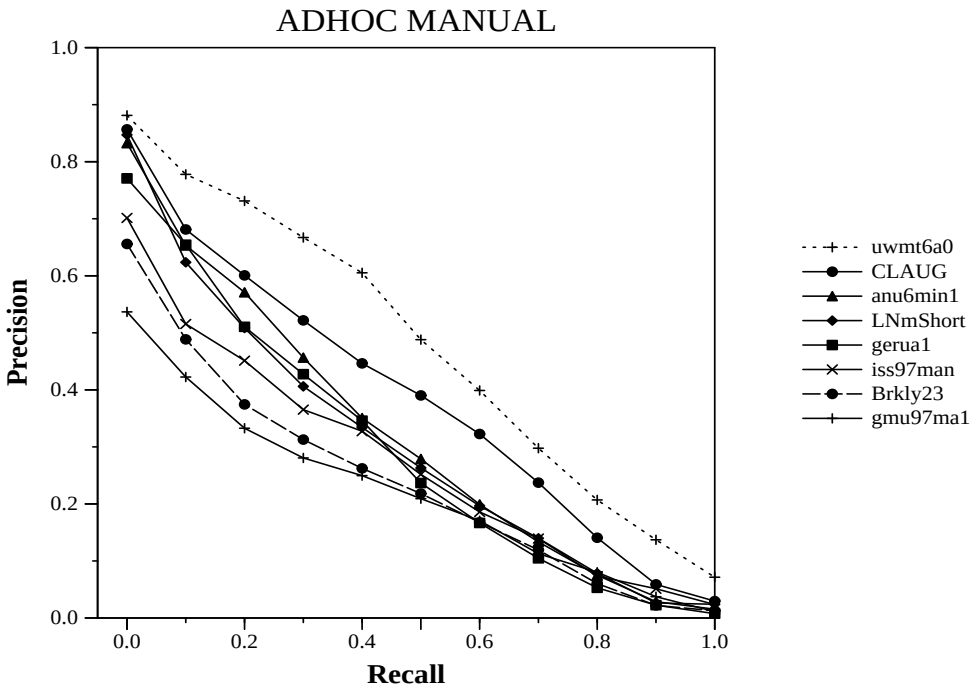

Figure 8: Recall/Precision graph for the top eight manual ad hoc runs.

30 negative terms selected via probabilistic term selection from user-judged documents.

anu6min1 – Australian National University ("ANU/ ACSys TREC-6 Experiments" by David Hawking, Paul Thistlewaite, and Nick Craswell) performed experiments investigating how well a relatively naive user could perform a series of edits on automatically generated queries, including removing or adding obvious terms, combining terms into phrases, altering weights, and grouping terms into concepts. These edits added 14% to average precision performance, and the further use of interactive modification improved precision by an additional 12% (with minimal improvement in recall).

LNmShort – Lexis-Nexis ("Query Processing in TREC-6)" by A. Lu, E. Meier, A. Rao, D. Miller, and D. Pliske) also performed experiments in manual editing of automatically generated queries. Their modified version of the Okapi BM25 algorithm was used to rank documents from automatic queries, and the top 20 documents were read looking for additional useful terms. The editing of the queries involved adding additional terms, removing negated terms, and doubling the frequency of the original query terms. This edited query was then used as input to some of the same automatic experiments used in the automatic runs. The addition of the manually selected terms gave major improvements.

geruna1 – GE Corporate R&D/Rutgers University ("Natural Language Information Retrieval TREC-6 Report" by T. Strzalkowski, F. Lin and J. Perez-Carballo) continued their investigations into contributions of natural language processing. This particular run represents experiments with the automatic generation of query-related summaries for the top 30 documents retrieved by the original topic. Users then added these summaries to the query if they "captured some aspects of relevant documents". These manually-expanded queries were run through the natural language processing modules to generate the final results.

iss97man – Institute for Systems Science ("Concrete Queries in Specialized Domains: Known Item as Feedback for Query Formulation" by M. Leong) used TREC as an environment to perform a specific experiment in manual query building. Their hypothesis was that expert users would be able to use very precise search terms, and this was tested using a two-stage search. In the first stage the users were given 20 minutes to find one or more highly relevant documents. In the second

stage the users were given 10 minutes to build a query that would return one of these highly relevant documents within the top 10 documents in the ranked list. This query was then used as the input to the manual run.

Brkly23 – University of California, Berkeley ("Phrase Discovery for English and Cross-language Retrieval at TREC-6" by Fredric C. Gey and Aitao Chen) did a manual reformulation of their queries.

gmu87ma1 – George Mason University/OIT/NCR ("Expanding Relevance Feedback in the Relational Model" by C. Lundquist, D. Holmes, D. Grossman, and O. Frieder) used their relational database model information retrieval system to experiment with pre-defined concept lists combined in different ways. These concepts were generated by first running a manual query, and then using relevance feedback and term-term association lists to generate more potential terms. These terms were then manually grouped into concept lists.

5.1.5 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

	Title	Short	Long
City	21	11	18
PIRCS	23	9	18

City 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[2].

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 Fig. 4 on page 8. 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 5 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

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

Table 5: Use of new techniques in the ad hoc task.

TREC-2	TREC-3	TREC-4	TREC-5	TREC-6
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
use of subdocuments by PIRCS system	heavy use of passages / subdocuments			use of passages in relevance feedback
	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
	beginning of manual expansion using other sources	major experiments in manual editing / user-in-the-loop	continued user-in-the-loop experiments	extensive user-in-the-loop experiments
	initial use of "data fusion"	continued use of "data fusion"	continued use of "data fusion"	more complex use of "data fusion"
			beginning of more concentration on initial topic	continued focus on initial topic, including title

use of better term weighting, including correct length normalization procedures, made this technique less used in TREC's 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 2 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. 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 5, 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 5) 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 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 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 3.2, 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 (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.

Data fusion (line 5 in table 5) 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.

The final major research area shown in this table started in TREC-5. This area is illustrated in the 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).

5.2 TREC-6 routing results

The routing evaluation used a specifically selected subset of the training topics against a new set of test documents. The routing tests in TREC-4 and TREC-5 had serious mismatches in the training and the test data, and it was determined to try routing in TREC-6 using very similar training and testing data. To this end the topics were the TREC-5 topics that had reasonable numbers of relevant documents from the FBIS data. To replace the "bad" topics, nine new topics, with minimal training data, were created, bringing the total to 47 topics. The test data for TREC-6 was additional FBIS documents.

There was a total of 34 sets of results for routing evaluation, with 33 of them based on runs for the full data set. Of the 33 systems using the full data set, 28 used automatic construction of queries, and 5 used manual construction. The single Category B routing run used automatic construction of queries.

Figure 9 shows the recall/precision curves for

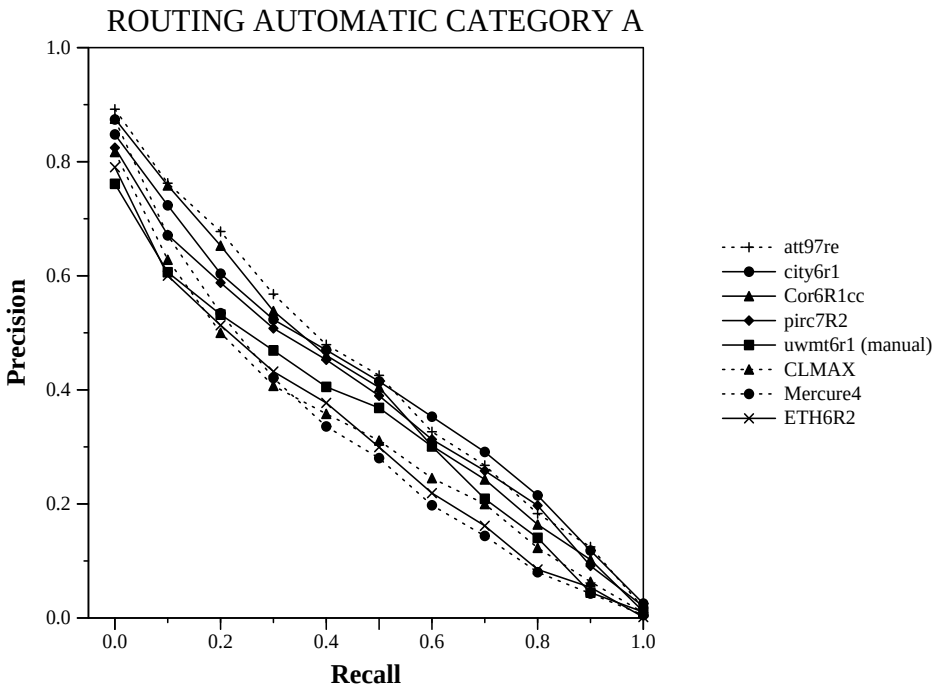

Figure 9: Recall/Precision graph for the top eight routing runs.

the eight TREC-6 groups with the highest non-interpolated average precision for the routing queries. The runs are ranked by the mean average precision over the 47 topics. A summary of the techniques used in these runs follows. For more details on the various runs and procedures, please see the cited papers in this proceedings.

att97re – AT&T Labs Research ("AT&T at TREC-6" by A. Singhal) used a variant of the Cornell TREC-5 routing algorithm. The modification added a version of the machine learning technique of boosting to the query refinement phase of the basic algorithm that includes the use of word pairs, DFO optimization, and query zones. The boosting added a small advantage (approximately 4%) compared to the algorithm without boosting.

city6a1 – City University, London ("Okapi at TREC-6: Automatic ad hoc, VLC, routing, filtering and QSDR" by S. Walker, S.E. Robertson, and M. Boughanem) explored iterative methods of term weighting with a major goal of avoiding overfitting the training data. This run is the result of merging 24 queries generated by picking various numbers of terms from the training set. For half the queries, the full FBIS training set was used; for the other half the training set was split in half, and one part was used to pick the terms and the other part was used to weight the terms.

Cor6A3cl – Cornell/SaBIR Research ("Using Clustering and SuperConcepts Within SMART: TREC 6" by C. Buckley, M .Mitra, J .Walz, and C . Cardie) added the SuperConcept technique to their basic TREC-5 routing algorithm. The SuperConcept technique attempts to maintain a balance between the different concepts represented in the original query by having expansion terms related to a particular concept of the original query share the total weight allocated to the concept. This technique did not improve the routing results as compared to the basic TREC-5 routing algorithm. However, the DFO optimization had not been modified to work with SuperConcept weighting, so improvements may still be possible.

pirc7R2 – Queens College, CUNY ("TREC-6 English and Chinese Retrieval Experiments using PIRCS" by K.L. Kwok, L. Grunfeld, and J.H. Xu) continued experimentation with merging of results from multiple runs. Five runs using different retrieval methods were used: one run using the topic only, one run using the training data only (only FBIS documents), two runs using

the genetic algorithms from TREC-5, and one using a new back propagation algorithm. This run combined the results of a combination of the first four methods with the back propagation run. This combined result was superior to all of the component runs.

uwmt6r1 – University of Waterloo ("Passage-Based Refinement (MultiText Experiments for TREC-6)" by G. Cormack, C Clarke, C. Palmer, and S. To) submitted a manual run using tiered Boolean queries that were refined interactively. An initial manual query was decomposed into basic components and combinations of these components were assigned to tiers such that combinations that retrieved relevant documents occurred in early tiers. The refinement produced small, but consistent, improvements over the original queries, and a future goal is to automate the process.

CLMAX – Claritech Corporation ("Experiments in Query Optimization: The CLARIT System TREC-6 Report" by N. Milic-Frayling, C. Zhai, X. Tong, P. Jansen, and D.A. Evans) 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. They discovered that the query that performed best on the training data was not always the query that performed best on the test data: the results of the *CLMAX* run were not better than some of the component runs, while the results of the combined run using the actual best-performing queries were significantly more effective than each of the component runs.

Mercure1 – MSI/IRIT/SIG/CERISS ("Mercure at trec6" by M. Boughanem and C. Soulé-Dupuy) continued their work with a spreading activation model. The initial queries were automatically built from the topics and then expanded using the top 30 terms from relevance backpropagation. To prevent the query from becoming too much like the already retrieved relevant documents, terms that occurred in relevant documents that were not retrieved in the top 1000 by this system were given a small extra weight.

ETH6R2 – Swiss Federal Institute of Technology (ETH) ("ETH TREC-6: Routing, Chinese, Cross-Language and Spoken Document Retrieval" by B. Mateev, E .Munteanu, P. Sheridan, M. Wechsler, and P. Schäuble) ran further experiments with the U-measure. The top 300 single-word features and top 300 phrases were selected based on this measure. These features were then grouped using a similarity thesaurus and used as one component of a combined run. The other components consisted of a straight Lnu.ltn query expansion run and a run using feature co-occurrence matrices.

The best mean average precision for a routing run in TREC-5 was .386 (using 39 topics) and for TREC-6 was .420. While this is a 9% improvement, a greater improvement was generally expected. As stated earlier, the test data in the TREC-4 and TREC-5 routing tasks were not very similar to the training data, whereas the TREC-6 task was designed to use a homogeneous data set. Indeed, the histogram given in Fig. 10 shows that the training and test data do have similar numbers of relevant documents for most topics.

At this point, it is unclear why the routing results are not better than they are. 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. Another hypothesis suggested by Amit Singhal [5] is that the relevance judgments are less consistent for routing than they are for the ad hoc task. 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. On the one hand, a ceiling of .42 in retrieval effectiveness because of relevance judgment inconsistency is extremely unlikely. On the other hand, the techniques used to create the routing queries from the training data may magnify the effects of inconsistent judgments. 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 [4].

The routing guidelines allowed participants to use any/all of the relevance judgments available for a topic in the training for that topic. The filtering track, in contrast, 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

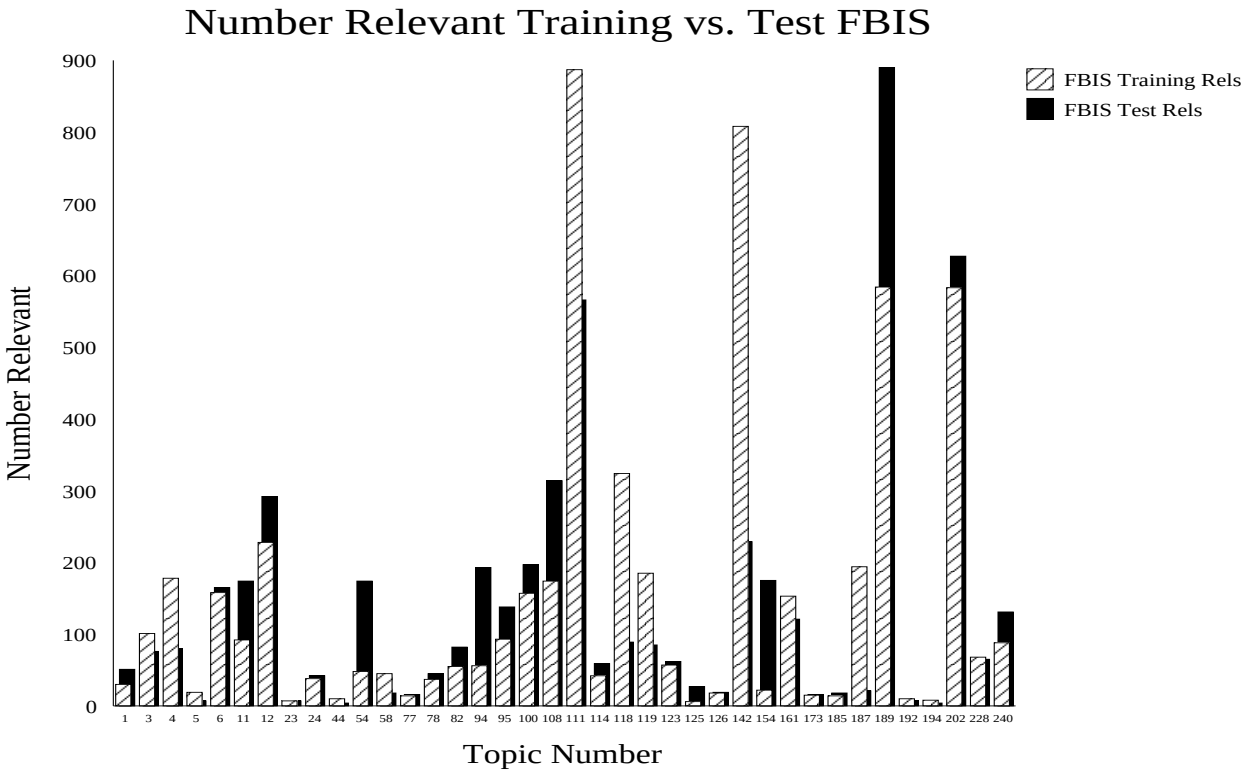

Figure 10: Comparison of the number of relevant documents in the training and test FBIS collections.

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. CISRO concluded precisely the opposite: that the FBIS training examples were too limiting and the variety introduced by judgments on other sources improved their results. Verity suggested a compromise of using all the judgments while emphasizing a particular source. The optimum trade-off between specificity and generality of the training data is clearly different for different techniques, and should be explored further.

6 Summary

TREC continues to grow both in number of participants and in number of tasks. The main tasks provide an entry point for new participants and provide a baseline of retrieval performance; the tracks invigorate TREC by introducing research in new areas of information retrieval. The Chinese track and the earlier Spanish track were the first (large-scale) formal tests of retrieval systems for languages other than English. The new Cross-Language Track exploits the current high interest in cross-language retrieval and serves as a testing platform both in the United States and Europe. The Spoken Document Retrieval Track, another track introduced in TREC-6, has joined the speech recognition and information retrieval communities, providing opportunities for rich interaction.

As always, it is difficult to summarize the many retrieval experiments that were performed in the context of TREC-6. Each group ran multiple experiments that resulted in their TREC submission, and readers are urged to explore the individual papers in this proceedings. In addition, Appendix B, "Summary Performance Comparisons TREC-2, TREC-3, TREC-4, TREC-5, TREC-6" by Karen Sparck Jones presents a snapshot of various system performances, particularly in the high precision end of the retrieval spectrum.

Several general conclusions can nevertheless be drawn from the main task experiments. The routing results suggest that there is still much to be learned about the stability of methods used to construct routing queries. The surprisingly good performance of the very short (titles only) ad hoc runs demonstrates the power of a few well-chosen query words—just as the relatively poor performance of the short ad hoc runs demonstrates how important it is to include those words. While this difference between the very short and short versions of the topics confounds results, there are suggestions that changing retrieval strate-

gies according to query length is beneficial.

The final session of each TREC workshop is a planning session for future TREC's. One of the tasks in this year's session was to contain the growth of TREC tasks in the face of finite resources at NIST to support TREC. Accordingly, the routing task was retired as a main task, though it will continue as a sub-task of the filtering track in TREC-7. The decision to retire the routing task was based on both the general agreement that the filtering task is a more realistic routing-type problem than the routing task as it has been defined in TREC, and that routing research can continue with the six routing collections that have already been built. Two tracks, NLP and Chinese, have also been discontinued for TREC-7, while a new Query Track will be introduced in TREC-7. The Query Track is designed to foster research on the effects of query variability on retrieval performance by creating and distributing many different queries derived from existing TREC topics.

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

[1] Gordon V. Cormack, Christopher R. Palmer, and Charles L. A. Clarke. Efficient construction of large test collections. In *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-98)*, 1998. To appear.

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

[3] Donna Harman. Overview of the fourth Text REtrieval Conference (TREC-4). In D. K. Harman, editor, *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*, pages 1–23, October 1996. NIST Special Publication 500-236.

[4] Linda Schamber. Relevance and information behavior. *Annual Review of Information Science and Technology*, 29:3–48, 1994.

[5] Amit Singhal. AT&T at TREC-6. In *Proceedings of TREC-6*, 1998. In this volume.

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

[7] K. Sparck Jones and C. van Rijsbergen. 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, 1975.

[8] Jean Tague-Sutcliffe and James Blustein. A statistical analysis of the TREC-3 data. In D. K. Harman, editor, *Overview of the Third Text REtrieval Conference (TREC-3) [Proceedings of TREC-3.]*, pages 385–398, April 1995. NIST Special Publication 500-225.

[9] Ellen Voorhees and Donna Harman. Overview of the fifth Text REtrieval Conference (TREC-5). In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the Fifth Text REtrieval Conference (TREC-5)*, pages 1–28, November 1997. NIST Special Publication 500-238.