Overview of the Fifth Text REtrieval Conference (TREC-5)

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

1 Introduction

The fifth Text REtrieval Conference (TREC-5) was held at the National Institute of Standards and Technology (NIST) on November 20–22, 1996. 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-5 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 [21], Tague-Sutcliffe and Blustein [23], and Harman [8]. 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-5. Thirty-eight groups including participants from nine different countries and ten 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 proven to 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-5. 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, were introduced. TREC-5 continued the tracks started in TREC-4 and added a new track on natural language processing (NLP). 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

Table 1: Organizations participating in TREC-5

Apple Computer	MITRE
Australian National University	Monash University
CLARITECH Corporation	New Mexico State University (two groups)
City University	Open Text Corporation
Computer Technology Institute	Queens College, CUNY
Cornell University	Rank Xerox Research Center
Dublin City University	Rutgers University (two groups)
FS Consulting	Swiss Federal Institute of Technology (ETH)
GE/NYU/Rutgers/Lockheed Martin	Universite de Neuchatel
GSI-Erli	University of California, Berkeley
George Mason University	University of California, San Diego
IBM Corporation	University of Glasgow
IBM T.J. Watson Research Center	University of Illinois at Urbana-Champaign
Information Technology Institute, Singapore	University of Kansas
Institut de Recherche en Informatique de Toulouse	University of Maryland
Intext Systems	University of Massachusetts, Amherst
Lexis-Nexis	University of North Carolina
MDS at RMIT	University of Waterloo

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 50 routing topics (selected by NIST from the pool of training 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. In TREC-5, the routing topics were selected by choosing topics that had many relevant documents in the Associated Press (AP) collection and the test documents were articles extracted from the Foreign Broadcast Information Service (FBIS).

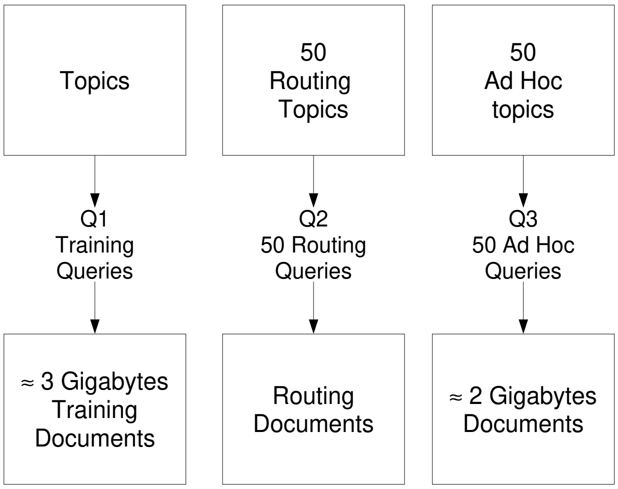

Figure 1: TREC main tasks.

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 251–300 were created for the TREC-5 ad hoc task. The set of documents used in the task were those contained on Tipster Disk 2 and the new TREC Disk 4; see Section 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 as fair as possible separation among the diverse query construction approaches. The allowable query construction methods in TREC-5 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. In contrast to previous TRECs, the definition of manual query construction methods in TREC-5 permitted users to look at individual documents retrieved by the ad hoc queries and then reformulate the queries based on the documents retrieved.[1]

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 two manual runs and up to two automatic runs. An additional constraint in this year's ad hoc task was that 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.

[1]Previous TRECs defined a third query construction method, *interactive*, for these types of runs. However, the interactive track in TREC-5 evolved such that these simple "manual feedback" runs did not fit well within the track's focus. The program committee redefined the manual query construction method to give the participants who were interested in studying manual feedback methods a home in TREC-5. Since both one-time and feedback runs are included in the single category of manual methods, care must be taken when comparing the results of manual runs.

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. Each of the tracks started in TREC-4 continued in TREC-5. In addition, a new track that focussed on using natural language processing techniques to improve retrieval performance was begun, and a "pre-track" laid the groundwork for the debut of the very large corpus (target of 20 GB of text) track in TREC-6.

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.

Confusion: The confusion track investigates how retrieval performance is affected by noisy or "confused" data. In this running of the track, participants performed *known-item* searches; that is, they searched for particular previously identified documents in three versions of documents. The three versions of the documents were the original documents, the documents that resulted after the originals were subjected to an optical character recognition (OCR) process with a character error rate of approximately 5%, and the documents produced through OCR with a 20% error rate (caused by down-sampling the image before doing the OCR).

Database Merging: The database merging track investigates methods for producing a single document ranking for queries when the underlying data sets consists of separate document collections. The TREC-5 track had an explicit focus on accomplishing the distributed search *without* searching every document collection for every query. That is, part of the task was to define a method that selects some proper subset of the document collections to be searched for a given query.

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.

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.

Multilingual: The multilingual track investigates retrieval performance when the text (both documents and topics) is in a language other than English. The TREC-5 track contained both a Spanish task, which had also been run in TREC-4, and a Chinese task, which was introduced in TREC-5.

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.

3 The Test Collections

Like most traditional retrieval 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-5 main tasks.

3.1 Documents

TREC documents are distributed on CD-ROM's with approximately 1 GB of text on each, compressed to fit. For TREC-5, Disks 1, 2 and 3 were all available as training material (see Table 2) and Disk 2 and new Disk 4 were used for the ad hoc task. Additional new data (also shown in Table 2) was 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 4 and the FBIS routing test data are new TREC document sets. The *Federal Register* is the official record of the executive branch of the U.S. Government. Similarly, the *Congressional Record* is the proceedings of the legislative branch of the U.S. Government; the copy of the 103rd *Congressional Record* was obtained from Dean Wilder of the Library of Congress. The *Financial Times* articles were obtained from the Financial Times through the University of Glasgow. The Foreign Broadcast Information Service provides (English translations of) selected non-U.S. broadcast and print publications. The documents used in the routing test were mostly from the early 1990's and were provided for TREC use by the Foreign Broadcast Information Service.

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 not only much shorter, but also were missing 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 on page 7 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

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 # Terms/Doc	Mean # Terms/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
Routing Test Data				
Foreign Broadcast Information Service (FBIS)	470	130,471	322	543.6

and automatic systems trouble, and that there were issues associated with using short topics in TREC that needed further investigation [9]. Accordingly, the TREC-5 ad hoc topics re-introduced the title and narrative fields (see fig. 4, page 8), although, as shown in Table 3, the length of the topics as measured by number of words was generally shorter than in TREC-3.

Groups who performed automatic ad hoc runs were required to use a short version of the topics, just the "Description" field, for one of their runs. These runs are tagged as "short, automatic" runs in the results section; 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. The effect of the different lengths on retrieval performance is described in Section 5.

As was true for TREC-3 and TREC-4, each TREC-5 ad hoc topic was constructed by the same person who performed all relevance assessments for that topic (with a few exceptions). Assessors were asked to come to NIST with already-constructed topics. The assessors used these topics to search (part of) the TREC-5 ad hoc collection (using NIST's ZPRISE system) and to make an initial set of relevance assessments. NIST personnel used these assessments to select the final set of 50 topics from approximately 150 candidate topics, based mainly on how many relevant documents were found in the search. Candidate topics that retrieved too many or too few relevant documents were rejected. Candidate topics were also rejected if they seemed ambiguous. Once the final set of topics were selected, the initial relevance assessments were discarded.

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 — hopefully as comprehensive a list as possible. All TRECs have used the pooling method [22] 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.

```
<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*.

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 on page 9 summarizes the amount of overlap in the ad hoc and routing pools for each of the five 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 100 times the number of runs used to form 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 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. Various tracks in TREC-4 and TREC-5 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.[2]

Since participants were allowed to submit two manual and two automatic ad hoc runs in TREC-5, many more ad hoc runs were judged than in previous years. The average actual pool size is also much larger than before. Much of the increase in the pool size can be attributed to the increase in the number of category B runs: while ad hoc runs in general increased from 40 to 77 between TREC-4 and TREC-5, category B runs increased from 6 to 16 runs. The comparatively large number of category B runs decreases overlap among runs in two ways. First, since category B runs retrieved documents from only the *Wall Street Journal* (WSJ) collection and the category A runs retrieved documents from seven different collections, the category B runs contribute many WSJ documents that

[2] The interactive track also contributed some documents to the ad hoc pool in TREC-5. However, the track used only a few topics, and many fewer than 100 documents were retrieved per topic. Table 4 does not include any interactive results for TREC-5.

```
<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:  251
<title> Exportation of Industry

<desc> Description:
Documents will report the exportation of some part of U.S. Industry to another
country.

<narr> Narrative:
Relevant documents will identify the type of industry being exported, the country
to which it is exported; and as well will reveal the number of jobs lost as a
result of that exportation.
```

Figure 4: A sample TREC-5 topic.

are not retrieved by category A runs. The second column of Table 5 shows the total number and the percentage of documents from each data source across all 50 ad hoc pools. Nearly half of the documents in the pools were from the *Wall Street Journal*. Second, the WSJ collection had a relatively small number of relevant documents, yet 100 documents per topic were added to the pools from each category B run. The third column of Table 5 shows the total number and the percentage of relevant documents from each data source across all 50 ad hoc pools. The WSJ collection contributed only 19% of the relevant documents. In general, pools for topics with fewer relevant documents exhibit less overlap, since systems that retrieve the same relevant documents often differ in the nonrelevant documents they retrieve.

Table 4 also shows that the average number of relevant documents per topic has decreased over the years. As discussed below, NIST has deliberately chosen more tightly-focused topics to better guarantee the completeness of the relevance assessments. Larger pools coupled with fewer relevant documents means the percentage of the actual pool that is relevant has also been decreasing.

3.3.2 Quality of Relevance Assessments

Given the vital role relevance judgments play in a test collection, it is important to assess the quality of the judgments created using the pooling technique. 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 TREC-4 overview [9] reports on the results of an investigation of the completeness of the TREC-2 and TREC-3 relevance judgments. The relevance assessors judged the documents in new pools formed from the second 100 documents in the ranked results submitted by participants. On average, the assessors found approximately one new relevant document per run (i.e., one relevant document that was not in the pool created from the top 100 documents of each ranking). The distribution of the new relevant documents was roughly uniform across runs, but was skewed across topics — topics that had many relevant documents initially also had many more new relevant documents. This latter finding motivates the use of topics that have relatively few relevant documents in TREC, while the lack of bias against particular participants and the small number of new relevant documents found indicate the completeness is quite acceptable for TREC purposes.

A separate experiment to test the consistency of the relevance assessments was conducted after TREC-4. For each of the 49 TREC-4 ad hoc topics, a pool consisting of 200 randomly selected relevant documents (or all relevant documents if there were fewer than 200) and 200 randomly selected, judged nonrelevant documents was created. The pool was then given to two additional assessors who were each asked to judge the documents.

Of the 14,968 documents that were judged in this experiment, 71.7% received an unanimous judgement: 1992 (13.3%) unanimous relevant and 8742 (58.4%) unanimous nonrelevant. A three-way unanimous agreement is quite a stringent test, and these rates are somewhat higher than those found in other studies [10]. Nonetheless, there were areas of significant disagreement. On average, 30% of the documents that the primary assessor marked relevant were

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

judged nonrelevant by both secondary assessors. In contrast, less than 3% of the documents judged nonrelevant by the primary assessor were considered relevant by both secondary assessors.

A primary goal of TREC is to construct test collections to facilitate IR research. In this context the important question regarding relevance assessments is not inter-assessor consistency per se but whether the assessments accurately reflect the relative merits of different retrieval techniques. Earlier studies have concluded that the ranking of retrieval techniques by effectiveness was stable across different sets of relevance assessments despite marked differences in the individual sets [14, 4]. While these conclusions are encouraging, the studies used small document collections and compared variants of the same retrieval systems, whereas TREC involves significantly larger collections and a wide variety of retrieval approaches. We therefore investigated how the ranking of systems by effectiveness varied with respect to different relevance assessment sets.

As a preliminary test of the stability of system rankings, we created five different "qrels" sets, where each qrels set consists of a particular set of assessments for each of the 49 topics. The original qrels set

Table 4: Overlap of submitted results

Ad Hoc

	Possible	Actual	Relevant
TREC-1	3300	1279 (39%)	277 (22%)
TREC-2	4000	1106 (28%)	210 (19%)
TREC-3	2700	1005 (37%)	146 (15%)
TREC-4	7300	1711 (24%)	130 (08%)
ad hoc	4000	1345	115
confusion	900	205	0
dbmerge	800	77	2
interactive	1600	84	13
TREC-5	10,100	2671 (27%)	110 (04%)
ad hoc	7700	2310	104
dbmerge	600	72	2
NLP	1800	289	3

Routing

	Possible	Actual	Relevant
TREC-1	2200	1067 (49%)	371 (35%)
TREC-2	4000	1466 (37%)	210 (14%)
TREC-3	2300	703 (31%)	146 (21%)
TREC-4	3800	957 (25%)	132 (14%)
routing	2600	930	131
filtering	1200	27	1
TREC-5	3100	955 (31%)	113 (12%)
routing	2200	854	94
filtering	900	100	19

Table 5: Number and percentage of documents in pool and relevant documents across all 50 ad hoc topics by document source

	Docs in Pool		Relevant Docs	
AP	16,460	(13%)	1644	(30%)
CR	10,467	(8%)	844	(15%)
FR88	5,341	(4%)	38	(1%)
FR94	8,347	(6%)	200	(4%)
FT	19,515	(15%)	1582	(29%)
WSJ	62,017	(47%)	1049	(19%)
ZIFF	9,601	(7%)	123	(2%)

Table 6: Mean average precision values for selected runs across five qrels sets

Run	Orig	A	B	∪	∩
uwgcl1	.2994	.2724	.2775	.3133	.2607
CLARTF	.2669	.2620	.2955	.3022	.2551
CLARTN	.2576	.2493	.2794	.2898	.2433
crnlAE	.2944	.2884	.2887	.3165	.2705
crnlAL	.2829	.2767	.2809	.3010	.2689
fsclt1	.1303	.1190	.1392	.1327	.1338
fsclt2	.1248	.1124	.1337	.1271	.1271

consists of the primary assessments for each topic — this is the qrels set released after TREC-4. The second and third qrels sets consist of a secondary relevance set for each topic. These qrels are equivalent to a qrels that might have been produced after a TREC conference if that set of assessors had been assigned those topics. Finally, we created a "union" qrels in which a document is considered to be relevant to a topic if any assessor judged it relevant to that topic, and an "intersection" qrels in which a document is considered relevant to a topic if all three assessors judged it relevant to that topic.

We evaluated each of the 33 TREC-4 category A ad hoc runs using each of the five qrels sets, and ranked the runs by decreasing mean average precision. Table 6 gives the mean average precision values for a subset of the 33 runs. Each pair of runs with similar names was submitted by a single participant. The final rankings produced by the different relevance assessments are very similar, though not identical. As was found in the earlier studies, in all cases where variants of a single system are compared, the ranking of the variants is the same across all qrels sets, even when the runs differ by a comparatively small margin. However, the ranking is somewhat less stable for different systems, even when the percentage difference in the original evaluation is considerably larger. For example, using the original qrels set, the `uwgcl1` run is approximately 11% better than the `CLARTF` run, yet for the Set B qrels set, the `CLARTF` run is approximately 6% better than `uwgcl1`. Note that the neither the intersection nor the union qrels appear to be materially different from the original qrels. Indeed, the Set B qrels appears to differ the most from the original qrels set.

We intend to perform a more detailed analysis of how system rankings vary with the qrels set used. However, these preliminary results suggest that the TREC relevance assessments reliably measure retrieval effectiveness when comparing variants of the same system, and thus support the use of the TREC test collection as a research vehicle. Comparing effectiveness across systems is more difficult in that a single comparison is unlikely to be meaningful. As Tague-Sutcliffe and Blustein found in their analysis of the TREC-3 data [23], seemingly large differences in average effectiveness may not be statistically significantly different.

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 timing, 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 these proceedings. The system descriptions are available on the TREC web site (currently http://www-nlpir.nist.gov/trec).

5 Retrieval Results

5.1 Introduction

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.

The 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.

5.2 TREC-5 ad hoc automatic results

The TREC-5 ad hoc evaluation used new topics (topics 251-300) against two disks of training documents (disks 2 and 4). A dominant feature of the ad hoc task was the desire of groups to continue work with the short topics like those used in TREC-4, but with the luxury of being able to compare results with those from a longer or full topic, such as the topics in TREC-3. The systems doing automatic query building were required to submit at least one run using only the short version of the topic (the description field) to allow this comparison. Groups doing manual query building could use the full topic.

There were 77 sets of results for ad hoc evaluation in TREC-5, with 61 of them based on runs for the full (category A) data set. Of these, 32 used automatic construction of queries, and 29 used manual construction. Fourteen of the category B runs used automatically constructed queries, and two used manually constructed queries.

Figure 5 shows the recall/precision curves for the eight TREC-5 groups with the highest non-interpolated average precision using automatic construction of queries for the short version of the topics. The runs are ranked by the average precision and only one run is shown per group.

A brief summary of the techniques used in these runs shows the breadth of the approaches and the changes in approach from TREC-4. For more details on the various runs and procedures, please see the cited papers in this proceedings.

Cor5A2cr – Cornell ("Using Query Zoning and Correlation Within SMART: TREC 5" by Chris Buckley, Amit Singhal and Mandar Mitra) used the same term weighting scheme (Lnu.ltu) [19] developed for TREC-4, but worked on better query expansion techniques. This took two paths: a more careful selection of the pseudo-relevant documents for use in query expansion, and the use of both relevant and non-relevant documents to compute Rocchio weights. A smaller number of terms and phrases were added (25 terms, 5 phrases) than in TREC-4 (50 terms, 10 phrases). Breakdown of the results shows minimal improvement from the use of non-relevant documents in reweighting, but a 12% improvement from the use of a "query coverage" algorithm to more accurately pick the top 20 documents declared to be relevant. The run on the full topics using this same technique (*Cor5mlle*) showed improved performance of 23% over using only the short version of the topics.

INQ301 – University of Massachusetts at Amherst ("INQUERY at TREC-5" by James Allan, Jamie Callan, Bruce Croft, Lisa Bellesteros, John Broglio, Jinxi Xu and Hongmin Shu) took advantage of their inference net architecture to combine the results of three different input queries for each topic. The first query was constructed in the same highly structured manner [2] as for the INQUERY TREC-4 queries. The second query contained only the most critical terms, and the third query used their local context analysis expansion method [24] to expand the initial query. Each of the three different types of queries performed better than using only the words in the short topic, with the largest improvement coming from the structuring of the query terms (24.5% improvement in average precision). Fusion of the three methods gave a nearly 40% improvement over using only the basic topic terms without structure and without expansion. The three methods perform differently at various recall levels, with the core query method performing the best at high recall (surprisingly) and the local context expansion performing the worst at the low recall (as would be expected). A similar run (revised) on the full topics (*INQ302c*) had 13.3% higher average precision than the run using only the short topics.

vtwnA1 – Apple Research Laboratories ("V-Twin: A Lightweight Engine for Interactive Use" by Daniel E. Rose and Curt Stevens) used an information access toolkit called V-Twin that was particularly built for use in interactive environments with very short queries. V-Twin is a vector-space model system using a variant of tf.idf weighting and length normalization. One aspect of this engine is its minimal memory requirement and very low indexing overhead: the index for TREC-5 was only 22.5% of the text size. A second aspect is the use of a new weighting function, called SQR, that is especially designed for interactive use with short queries. This function showed useful results during training from the TREC-4 data, but was inadvertently not used in the official TREC-5 results. The TREC-5 results did use automatic relevance feedback similar to other systems. Later unofficial runs showed that the SQR weighting function did not actually help performance, but more importantly, retained its more user-intuitive weighting without hurting performance.

pircsAAS – Queens College, CUNY ("TREC-5 English and Chinese Retrieval Experiments using PIRCS" by K.L. Kwok and L. Grunfeld) used a special term weighting scheme developed for short queries after TREC-4 [12] that is a way of

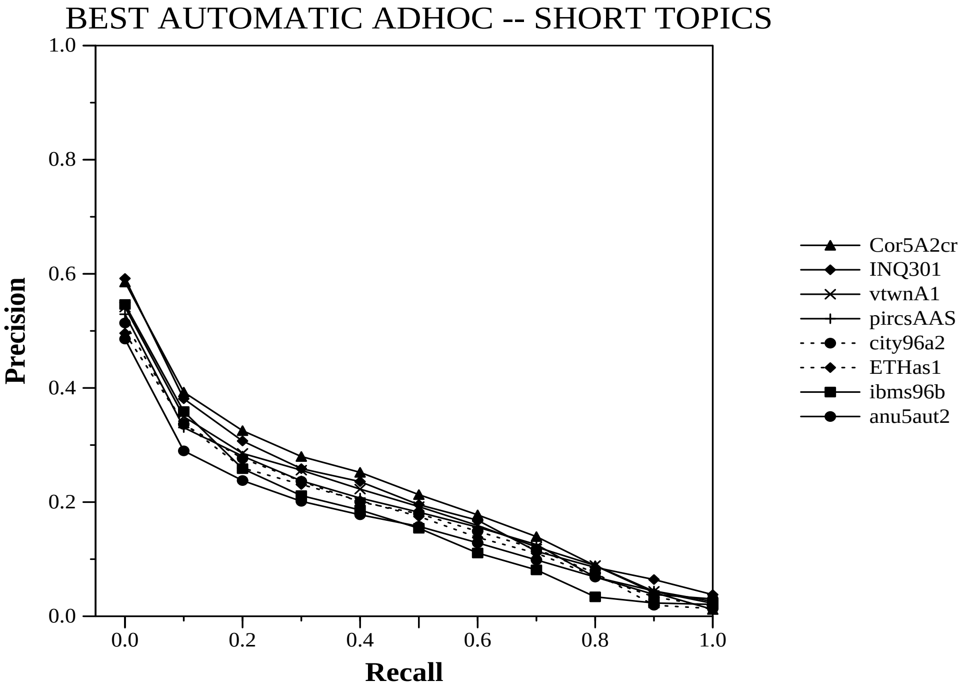

Figure 5: Recall/Precision graph for the top eight ad hoc, automatic, short runs.

simulating the manual reweighting of terms done in TREC-4. Also new for TREC-5 was an experiment in building 2-word phrases that are concatenations of high frequency and low frequency query terms to help match statistically-produced phrases in the documents. As in TREC-4, 50 expansion terms are picked from the top-ranked 40 subdocuments. The new term weighting scheme gave an almost 15% improvement for the short topics, with expansion adding an additional 13%. The experiment in phrases did not work and hurt performance by 6%. When these techniques were used on the full topics (*pircsAAL*), there was a 36% improvement in average precision over the run for short topics.

city96a2 – City University, London ("Okapi at TREC-5" by M.M. Beaulieu, M. Gatford, Xiangji Huang, S.E. Robertson, S. Walker, and P. Williams) used essentially the same OKAPI weighting and expansion schemes [17, 18] used in TREC-4. An emphasis was placed on efficiency this year, resulting in runs at 4 times the TREC-4 speeds. Additionally many experiments were tried (but failed), including variations of the term expansion algorithms and several experiments with adjacent term pairs. The *city96a2* run used the top 15 documents to get additional terms, for a total of 30 terms per query. Unlike the groups mentioned earlier, City experimented with the full topic, and then ran the short topic using a variation of the best techniques. The full topic run used the top 30 documents to get additional terms, and had a maximum of 55 terms per query. The full topic results showed a 30% improvement over the short version of the topic.

ETHas1 – Swiss Federal Institute of Technology (ETH) ("SPIDER Retrieval System at TREC-5" by Jean-Paul Ballerini, Marco Büchel, Ruxandra Domenig, Daniel Knaus, Bojidar Mateev, Elke Mittendorf, Peter Schäuble, Páraic Sheridan, and Martin Wechsler) used the basic Cornell TREC-4 term weighting schemes (Lnu.ltn), but a somewhat different feature selection method for picking the top 50 terms and 20 phrases. The full topic results showed a 40% improvement over the short version of the topic.

ibms96b – IBM T.J. Watson Research Center ("TREC-5 Ad Hoc Retrieval using K Nearest-Neighbors Re-Scoring" by Ernest P. Chan, Santiago Garcia and Salim Roukos) built a two-pass retrieval system using the OKAPI weighting formula for a first scoring/ranking of the documents, and then a second pass algorithm that rescored the top 1000 documents using a combination of the first score and a score based on using the top 10 documents as queries. They used single terms plus statistical two-word phrases, and a new sigmoidal suppression factor for length normalization.

anu5aut2 – Australian National University ("ANU/ACSys TREC-5 Experiments" by David Hawking, Paul Thistlewaite and Peter Bailey) used a parallel architecture with an emphasis on efficiency. The automatic queries were generated by adding multi-word terms to the original words, using statistical co-occurrence to locate these multi-word terms. Term frequency weights were used in the ranking algorithm, replacing the semantics-based techniques used in TREC-4.

Shortly before this overview was finalized, NIST was notified that the Lexis-Nexis run that had previously been categorized as automatic actually had some (minimal) human intervention. This run has been removed from Figure 5, but will be included in this section for discussion to ease the confusion and because the techniques are more related to the automatic techniques than to the other manual runs.

LNaDesc2 – Lexis-Nexis ("Ad Hoc Experiments Using EUREKA" by Allan Lu, Maen Ayoub and Jianhua Dong) used their experimental toolbox, the EUREKA system, for experiments to improve the ranking of their top set of retrieved documents. This was done to improve performance at the high precision end of the performance curve (of importance for interactive systems), and also to improve the pseudo-relevance feedback necessary for query expansion. They experimented with three complementary techniques: 1) use of inter-term distance between nouns in the short version of the topic, 2) fusion (using logistic regression) of the results from three different ranking measures, and 3) clustering of documents in the top 20 documents initially retrieved in order to improve accuracy.

Three experimental themes dominate these top groups. The first is the continued investigation into query expansion. All groups except ANU used the top n documents/subdocuments/passages to pick m terms, where n ranged from 15 to 40, and m ranged between 25 and 70 (not all groups identified these numbers). There were interesting variations in this theme. For example, Lexis-Nexis clustered the top documents in hopes of finding better features, INQUERY used noun groups based on their occurrence in the top ranked passages, and IBM used the top 10 documents as individual queries. Cornell tried some experiments using negative weights, but these were not successful. ANU expanded the terms in the short topics by using statistically co-occurring terms across the corpus, rather than using the top n documents as a source of expansion terms. As can be seen, differing amounts of improvement were found for these techniques, but the wide range of performance improvements is likely to come from interaction with the other features in the underlying systems rather than the particular expansion technique used.

Most (but not all) of the TREC-5 ad hoc experiments used some type of query expansion. Of particular note are two groups that tried unique methods. The first group, Open Text Corporation ("Experiments with TREC using the Open Text Livelink Engine" by Larry Fitzpatrick, Mei Dent, and Gary Promhouse) gathered terms for expansion by looking at relevant documents from past topics that were loosely similar to the TREC-5 topics. This worked very well, although their results were lowered by further shortening the topics to "imitate" real user requests. (For further work by Open Text on this method see [7].) The second group, the University of Kansas ("Corpus Analysis for TREC 5 Query Expansion" by Susan Gauch and Jianying Wang), used a complicated corpus linguistics approach involving context vectors and mutual information values. This approach was not as successful, likely because of the number of parameters that had to be discovered and tuned.

The second experimental theme in the TREC-5 ad hoc runs is the growing interest in getting more information from the initial topic, even the short version of the topic. Most of the top groups tried various schemes to improve on the "bag of words" approach to basic topic processing. INQUERY used their elaborate automatic query structuring method that has been enhanced over the years, with an improvement over the baseline query of 24.5%. Lexis-Nexis used a distance relationship between nouns in a topic to improve term weighting, and PIRCS used an automatic reweighting scheme on key concepts in the topic to gain 15% performance. Cornell used an analysis of the match between key concepts in the topic and those in small windows of the top 50 documents to rerank these initial documents (12% improvement). Note that these schemes not only improve performance in the initial ranked list (critical to interactive system performance), but also improve the set of documents used for query expansion (and therefore the high recall performance). More intense work with the TREC topics is a theme that is likely to expand in TREC-6.

The third experimental theme is the continued interest in data fusion. Three category A groups tried major data fusion experiments. The Lexis-Nexis group tried fusing results from three different ranking algorithms (variations on the OKAPI algorithm and the Cornell algorithms). The INQUERY system used fusion results from three different queries: a basic structured query, a "key concepts" query, and an expanded query. Experiments were also done by RMIT ("The MDS Experiments for TREC5" by Marcin Kaszkiel, Phil Vines, Ross Wilkinson and Justin Zobel) combining results from various parts of the original topic (and its expansion) when used as input to various cosine and OKAPI measures.

Two category B groups also experimented with data fusion. The Universite de Neuchatel ("Report on the TREC-5 Experiment: Data Fusion and Collection Fusion" by Jacques Savoy, Anne Le Calvé, and Dana Vrajitoru) tried fusion of results from multiple weighting algorithms (OKAPI and several different SMART weighting schemes), using logistic regression to create the final ranks. The University of California, San Diego ("Using Relevance to Train a Linear Mixture of Experts" by Christopher Vogt, Garrison Cottrell, Richard Belew and Brian Bartell) investigated the mixing of results from three "experts." These experts consisted of two weightings of SMART (binary and tf.idf), and the LSI techniques, and their experiments concentrated on examining the effects of various training techniques.

The experimental work in TREC-5 using the short version of the topic for automatic query construction shows considerable progress over that done in TREC-4. Some of this growth is due to more experience with (and acceptance of) the shorter topics, but much of it is due to accumulated knowledge in areas like query expansion and data fusion. Query expansion using the top-retrieved documents was started in TREC-3 by several groups, and by TREC-5 most groups have devised variations suitable for their particular systems. Data fusion experiments became more common, both in the ad hoc task and in the routing task. New for TREC-5 was more work with the initial topic.

Figure 6 shows the comparison of results between using only the short version of the topic (the description) and using the full topic, where both sets of results are shown for four of the systems previously described. Note that for all four of the systems there is a sharp rise in performance using the full topic as opposed to using the short topic only – PIRCS goes up by 36%, ETH by 40%, City by 30%, and the V-Twin Apple run by 9%. With the exception of the Apple runs, this performance difference appears at all levels of recall, i.e., there is a 25 to 30% difference even looking only at the top 10 documents retrieved.

Several reasons have been given for the generally large improvement in performance using the full topics. These center around two main causes: improved statistical power from the additional words, and more term variations or expansions in the longer topics. The increased statistical power of the full topic has many subtle effects, including improved term frequency information to give more weight to some terms, and more syntactic information to produce better phrases. The addition of term variations and expansions provides not only useful synonyms, but also "corrections" to stemmers and stopword lists via redundant words.

It is interesting to note that the Apple runs show the least difference, and the Lexis-Nexis run on the short form of the topics demonstrates potential ability to perform well automatically using minimal input. These commercial groups have clearly adapted to short user topics, and this adaptation is an important research area. Not only will the TREC-6 topics contain both a short and full version, but an even shorter title version (on the order of three words) will also be added.

TREC has not particularly emphasized efficiency, although some minimal efficiency is needed just to search the large text collections. Several groups, including RMIT [16], ANU [11], and George Mason University, have examined efficiency issues in past TRECs. The group from George Mason University ("Using Relevance Feedback within the Relational Model for TREC-5" by David Grossman, Carol Lundquist, John Reichart, David Holmes, Abdur Chowdhury and Ophir Frieder) has based their work on using efficient parallel database systems, and for TREC-5 investigated methods of incorporating relevance feedback into SQL. Two groups specifically investigated efficiency algorithms for TREC-5. The Computer Technology Institute ("Parallel Techniques for Efficient Searching over Very Large Text Collections" by B. Mamalis, P. Spirakis, and B. Tampakas) reported on various experiments using a new parallel version of their VSM-based traditional system. Dublin City University ("TREC-5 Experiments at Dublin City University: Query Space Reduction, Spanish and Character Shape Encoding" by Fergus Kelledy and Alan F. Smeaton) experimented with reducing the number of terms to be processed for each query by using several thresholding techniques. Both these groups were able to enhance efficiency without sacrificing effectiveness.

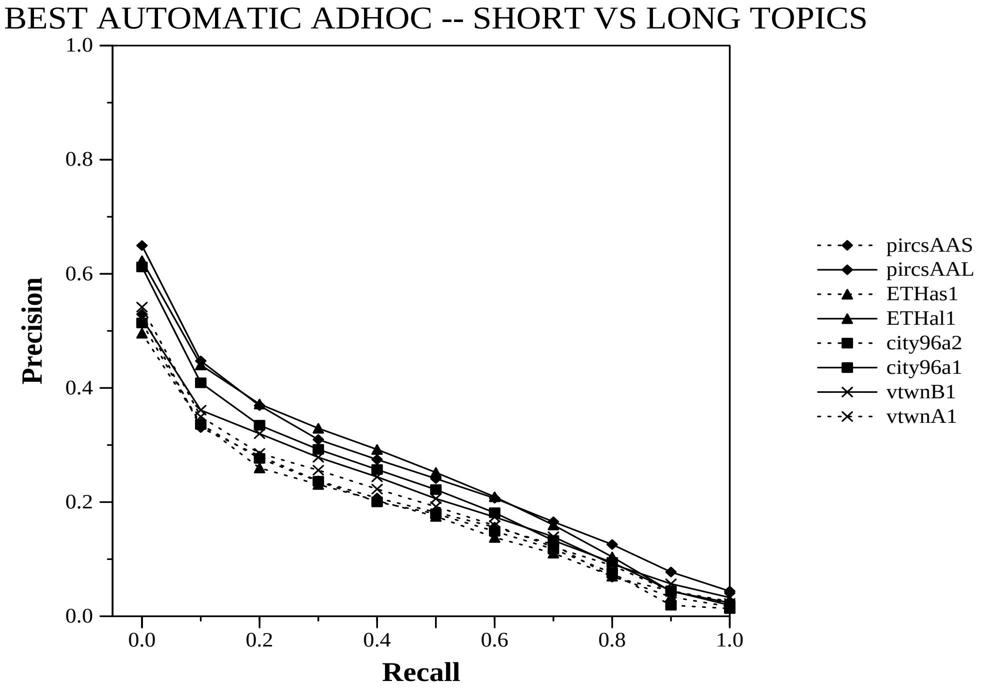

Figure 6: Comparison of short vs. long topics for selected systems.

5.3 TREC-5 ad hoc manual results

Figure 7 shows the recall/precision curves for the eight TREC-5 groups with the highest non-interpolated average precision using manual construction of queries. Note that manual query construction included user interaction in TREC-5, i.e., the rules were modified so that initial results could 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.

ETHme – Swiss Federal Institute of Technology (ETH) ("SPIDER Retrieval System at TREC-5" by Jean-Paul Ballerini, Marco Büchel, Ruxandra Domenig, Daniel Knaus, Bojidar Mateev, Elke Mittendorf, Peter Schäuble, Páraic Sheridan, and Martin Wechsler) did a completely manual search operation, including manual construction of the queries and query expansion with all found "relevant" documents. The users (who were information science students) had no time constraints, and spent about 30 to 40 minutes per topic. The basic retrieval system used was the same as for the automatic ad hoc runs, with an improvement in performance of 31% over the automatic run using the full topics.

uwgcx1 – University of Waterloo ("Interactive Substring Retrieval (MultiText Experiments for TREC-5)" by Charles L.A. Clarke and Gordon V. Cormack) used queries that were manually built in a special query language called GCL. This query language uses Boolean operators and proximity constraints to create intervals of text that satisfy specific conditions, and the concentration of these intervals of text is used to produce the ranking. The TREC-5 experiments involved several ranking method trials, including one with length normalization. The main experiment, however, tested the interactive use of the system to find suitable expansion terms.

LNmFull2 – Lexis-Nexis ("Ad Hoc Experiments Using EUREKA" by Allan Lu, Maen Ayoub and Jianhua Dong) repeated their *LNaDesc2* run using manually-edited versions of the full topic. The inter-term distance between nouns was not used because it was specifically built for the short version of the topics. No interaction was involved; results from the manually-edited queries were submitted without query modification. The manually-edited queries showed a 9% improve-

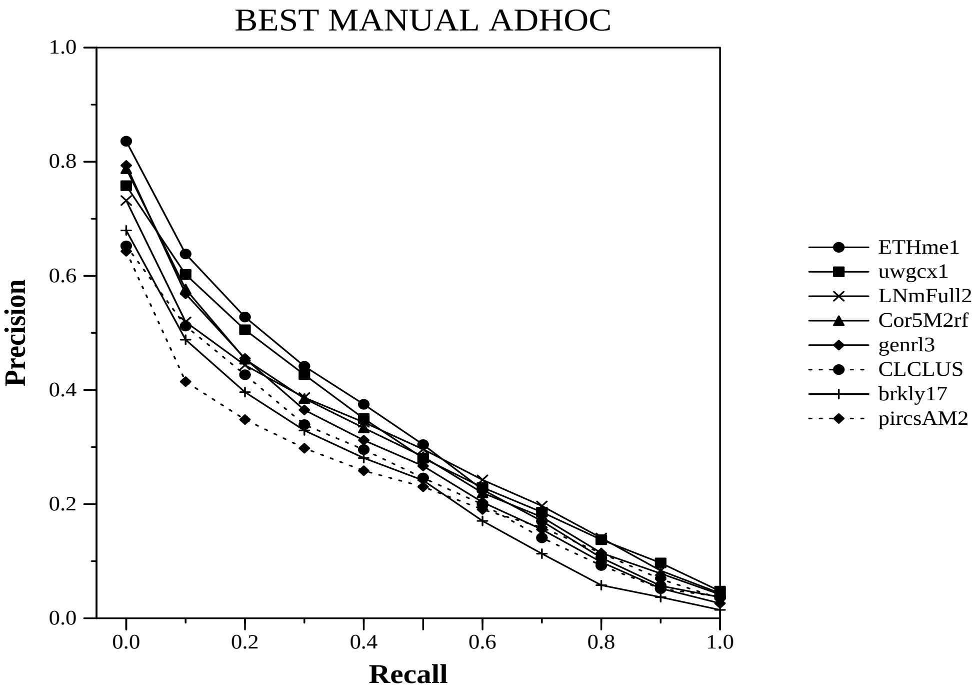

Figure 7: Recall/Precision graph for the top 8 ad hoc, manual runs.

ment over using only the short version of the topics.

Cor5M2rf – Cornell ("Using Query Zoning and Correlation Within SMART: TREC 5" by Chris Buckley, Amit Singhal and Mandar Mitra) used the same methods as the automatic runs with the full topics, but made manual relevance judgments on the top documents before using these documents for relevance feedback. On average only about 5 minutes was used to judge 25 documents per topic, and the improvement in performance was 15% over using the full set of top documents.

genrl3 – GE / Lockheed Martin / NYU / Rutgers ("Natural Language Information Retrieval: TREC-5 Report" by Tomek Strzalkowski, Louise Guthrie, Jussi Karlgren, Jim Leistensnider, Fang Lin, Jose Perez-Carballo, Troy Straszheim, Jin Wang and Jon Wilding) continued their investigations into contributions of natural language processing. This particular run represents experiments with users finding phrases and sentences to add to the initial query based on the top 10 documents retrieved from the initial queries (5 to 10 queries used per topic). These manually-expanded queries were run through the natural language processing modules to generate the final results.

CLCLUS – CLARITECH Corporation ("CLARIT Compound Queries and Constraint-Controlled Feedback in TREC-5 Ad-Hoc Experiments" by Nataša Milić-Frayling, Xiang Tong, Chengxiang Zhai and David A. Evans) used the commercial version of the CLARIT system in a series of experiments comparing the use of different sources for manual query expansion and the use of Boolean constraints to improve input to that expansion. This particular run used manual editing of an automatically-generated initial query, manual additions of Boolean constraints to that query, and finally manual expansion of the constrained query using terms selected from terminology clusters built from the full corpus. The results were not very different from the second manual CLARIT run, which used manually selected terms from the retrieved top documents.

Brkly17 – University of California, Berkeley ("Term importance, Boolean conjunct training, negative terms, and foreign language retrieval: probabilistic algorithms for TREC-5" by Fredric C. Gey, Aitao Chen, Jianzhang He, Liangjie Xu and Jason Meggs) used manually-reformulated queries based on both examination of the top retrieved documents and expansion using the

News database of the MELVYL electronic catalog (similar to that done in TREC-3 [6]. Experiments using negatively weighted terms showed a 15% decrease in performance. The basic retrieval system is a logistic regression model [5] that combines information from six measures of document relevancy based on term matches and term distribution. The coefficients were learned from the training data.

pircsAM2 – Queens College, CUNY ("TREC-5 English and Chinese Retrieval Experiments using PIRCS" by K.L. Kwok and L. Grunfeld) repeated their manual TREC-4 experiment to contrast with the their automatic simulation of the manual work. This run is the result of both a manual term reweighting, and a manual expansion of up to three terms per topic. There was no modification of the query based on examining the results. The reweighting part contributed about a 15% improvement over no reweighting, but less than a 1% improvement over the automatic version of this. The manual expansion did not show improvements over the automatic expansion.

There has been an interesting evolution in the methods used for manual query construction over the various TRECs. The rich topics for TRECs 1 and 2 showed little difference in performance for manually-produced queries over the automatic runs. When the concept section was removed from topics starting in TREC-3, most groups tried manual runs as a way of improving their scores. Examples of this would be systems such as INQUERY and PIRCS, where manual editing, manual reweighting and minimal manual query expansion were done to produce manual versions of their automatic runs. This was even more pronounced in TREC-4, where groups generally feared poor results from the short topics.

In general these low-effort manual runs were not done in TREC-5. The *pircsAM2* run was a repeat of their TREC-4 manual reweighting experiment, but was done in TREC-5 to verify that the automatic version of this had successfully replaced the manual version (which it had). Groups seemed more comfortable with the automatic version of their initial queries, and with automatic expansion, and felt no need for manual edits.

This trend was reinforced by the change of rules allowing unrestricted interaction with the systems in TREC-5. Except for the *pircsAM2* and *LNmFull2* runs, all runs performed some type of interaction. The simplest was the *Cor5M2rf* run, where minimal manual effort was spent to determine the "relevant" documents for use in the relevance feedback rather than automatically taking the top 20 documents. This can be contrasted with the *ETHme1* run, where students spent between 30 and 40 minutes per topic producing the "perfect" query.

The manual runs for TREC-5 can be loosely classified into three categories. The first category, exemplified by *uwgcx1* and *Brkly17*, used queries completely manually generated using some type of auxiliary information resource such as online dictionaries (*uwgcx1*) or news databases (*Brkly17*). The query generated for *uwgcx1* uses Boolean-type restrictors, whereas the query generated for *Brkly17* uses natural language. In both cases, the results from initial retrievals were then further expanded by looking at the retrieved documents, similar to the ways in which users of these systems might modify queries to get more relevant documents.

The second category of manual query construction runs involves a more complex type of human-machine interaction. The *CLCLUS* run is a result of experiments examining a multi-stage process of query construction, where the goal is to investigate better sets of tools that allow users to improve their queries (see [15] for similar experiments in TREC-4). The CLARITECH group tried using different sources for suggestions of expansion terms and also various levels of user-added constraints to the expansion process.

Two other groups also investigated human-machine interaction, with an emphasis on relevance feedback. FS Consulting ("Document Retrieval Using the MPS Information Server (A Report on the TREC-5 Experiment)" by Francois Schiettecatte) started with a baseline of manual queries, and then added terms based either on the user's selection of two relevant documents, or the top two system-selected documents. The University of North Carolina ("An Investigation of Relevance Feedback using Adaptive Linear and Probabilistic Models" by Robert Sumner and W.M. Shaw) tested two different models of relevance feedback, as applied by two different users.

By far the largest category of runs, however, could be labelled as "manual exploration" runs. This was specifically stated by GE, where the goal was to ask users to pick out phrases and sentences from the retrieved documents to add to the query, in hopes that this process can be imitated by automatic methods. Similar reasons are likely to apply to the ETH run and the Lexis-Nexis run, with the Cornell run and PIRCS run being more specific versions of exploration (performance differences using "relevant" documents vs the top 20 and performance differences using manual as opposed to automatic reweighting of terms).

Table 7: Variations of the hardness measure for different TRECs.

	TREC-3	TREC-4	TREC-5
best	0.6285	0.5657	0.5199
average	0.3767	0.2814	0.2460
median	0.3692	0.2929	0.2049

This use of the manual query construction category to identify new automatic methods is particularly promising.

5.4 Comments on the TREC-5 ad hoc topics

In general the TREC-5 ad hoc topics were thought to be more difficult than the TREC-4 ad hoc topics. The Cornell paper ("Using Query Zoning and Correlation Within SMART: TREC 5" by Chris Buckley, Amit Singhal and Mandar Mitra) contains a table (Table 16) showing comparison of the average precision of the Cornell runs over the five TRECs. Of particular interest here is the fact that the TREC-5 Cornell system performed about 34% worse on the TREC-5 topics than on the TREC-4 topics. Whereas some of this difference has to do with training, most of the difference is due to "harder" topics.

To further examine this issue, a measure of "hardness" was revived from earlier experimental use in TREC-2. The hardness measure is defined as an average over a given set of runs of the precision for each topic after all relevant documents have been retrieved *OR* after 100 documents have been retrieved, if more than 100 documents are relevant. This measure is therefore oriented towards high recall performance and how well systems do at finding all the relevant documents.

The hardness measure can be calculated over different sets of runs, using different types of averaging. Table 7 shows three types of averages for three TRECs. All these averages are for the category A ad hoc runs, both automatic and manual, and both full and short versions of the topic. The row labelled "best" shows the averages across all 50 topics of the best results for each topic by any run. This is therefore the highest possible performance. The row labelled "average" is the mean across all runs for all 50 topics, and the "median" row is the corresponding median performance.

As can be seen, the topics have grown progressively harder (the lower the hardness number, the harder the task). The drop between TREC-3 and TREC-4 was expected since the TREC-3 results are all based on the full topic and the TREC-4 results are based only on the short version of the topic. The TREC-5 results include both full and short versions of the topic, but compared to TREC-3, show a 35% drop in performance (using the average measure). This was unexpected, both by NIST and by the participants.

The drop in performance on the TREC-5 topics occurred not only at the high recall end of performance (as measured by the hardness measure) but also at the high precision end (at 30 documents retrieved). Appendix B, "Summary Performance Comparisons TREC-2, TREC-3, TREC-4, TREC-5" by Karen Sparck Jones, illustrates this, and also discusses some of the issues involved in these comparisons.

Investigation has been started at NIST into why the TREC-5 topics are more difficult. Table 8 shows a first attempt to isolate factors associated with the hardness measure. The topics are sorted by hardness, using the average hardness shown in Table 7. The second column contains the number of relevant documents for each topic, and the third column the length of the topic (unstemmed, without stopword removal). There appears to be little correlation between the hardness and the number of relevant documents or the topic length. A correlation coefficient using the Pearson product moment gives a correlation of 0.19 between the number of relevant documents and the hardness, and a correlation of 0.14 between the topic length and the hardness. This can be compared with a correlation of 0.20 between the topic number and the hardness, which is clearly a random correlation.

The fourth column of Table 8 shows the hardness of the topic as computed using all runs, while the fifth and sixth columns show the hardness as calculated separately for the automatic systems and the manual systems. Some of the topics show large differences for these runs, such as topic 293. One reason that this might happen is that automatic systems could not construct as accurate a query as the manual systems, as looks to be the case in topic 263. But another hypothesis is that the manual systems are able to find (and rank in the top 100) documents that the automatic systems could not. This led to an investigation of which systems found which relevant documents, and of particular note is the fact that large numbers of unique documents (those only found by one group) occur in TREC-5. The final columns of Table 8 are the percentage of the relevant documents that were unique, both for all runs, and also looking at the unique relevant documents found only by the manual runs. There is a correlation coefficient of 0.33

Table 8: Correlation between hardness and topic characteristics

Topic	# Rel	Topic Length	Hardness All	Hardness Auto	Hardness Manual	% Unique	% Manual Only
281	1	115	0.0000	0.0000	0.0000	0.0	0.0
296	1	53	0.0000	0.0000	0.0000	0.0	0.0
267	4	97	0.0000	0.0000	0.0000	50.0	0.0
293	41	78	0.0484	0.0252	0.0740	29.3	34.1
292	59	92	0.0514	0.0191	0.0871	35.6	37.9
268	45	117	0.0576	0.0424	0.0743	22.2	11.1
278	7	44	0.0609	0.0223	0.1035	14.3	28.6
275	19	100	0.0638	0.0592	0.0690	21.1	5.3
255	109	81	0.0782	0.0597	0.0986	41.3	33.0
252	37	53	0.1024	0.0946	0.1109	21.6	13.5
300	44	84	0.1032	0.0675	0.1426	20.5	25.0
290	119	58	0.1115	0.0656	0.1621	26.1	32.8
256	22	67	0.1289	0.1577	0.0972	22.7	13.6
291	407	100	0.1480	0.0969	0.2045	57.2	41.8
279	2	88	0.1557	0.1094	0.2069	0.0	0.0
299	62	66	0.1579	0.1472	0.1696	12.9	11.3
264	281	76	0.1608	0.1341	0.1903	48.0	37.1
284	70	78	0.1761	0.1603	0.1936	25.7	21.4
295	15	73	0.1803	0.1583	0.2046	20.0	13.3
287	40	134	0.1844	0.1727	0.1974	20.0	20.0
263	15	123	0.1847	0.1250	0.2506	0.0	0.0
282	131	30	0.1882	0.1709	0.2072	48.1	47.3
260	22	94	0.1923	0.1392	0.2508	4.5	13.6
294	160	77	0.1969	0.0912	0.3134	24.4	43.1
283	84	81	0.2004	0.2184	0.1806	13.1	10.7
289	141	163	0.2049	0.1647	0.2493	17.7	12.8
266	139	36	0.2125	0.1763	0.2524	27.3	52.5
254	85	83	0.2170	0.1761	0.2621	24.7	24.7
258	115	96	0.2184	0.1441	0.3003	18.3	19.1
251	579	50	0.2195	0.1000	0.3514	56.6	46.5
271	86	49	0.2326	0.1697	0.3019	9.3	9.3
298	91	47	0.2365	0.2236	0.2508	19.8	9.9
269	594	70	0.2459	0.1397	0.3631	63.1	58.9
272	36	65	0.2495	0.2309	0.2701	2.8	2.8
277	74	48	0.2497	0.2166	0.2861	21.6	12.2
297	86	142	0.2735	0.1457	0.4146	4.7	5.8
257	135	48	0.2826	0.2869	0.2779	14.8	11.9
261	87	214	0.3122	0.3057	0.3195	11.5	10.3
286	142	57	0.3462	0.3622	0.3286	16.9	13.4
270	116	156	0.3543	0.2703	0.4469	17.2	14.7
288	92	109	0.3770	0.3974	0.3546	3.3	4.3
274	119	38	0.3815	0.2772	0.4966	12.6	13.4
262	4	99	0.4508	0.3281	0.5862	0.0	0.0
259	36	63	0.4982	0.4991	0.4971	2.8	0.0
280	32	47	0.5435	0.5293	0.5593	0.0	0.0
285	261	140	0.5684	0.5566	0.5814	11.5	9.6
253	10	92	0.5869	0.5687	0.6069	0.0	0.0
273	513	69	0.6349	0.6247	0.6462	33.7	23.6
276	7	89	0.7143	0.7321	0.6946	0.0	0.0
265	147	56	0.7572	0.8047	0.7048	12.9	12.2

between the percentage of unique relevant documents for a topic and its hardness measure.

Having large numbers of unique documents for a given topic means that it is harder for the systems to retrieve all the relevant documents, and since this is what the hardness measure is based on, it is not surprising that there is a correlation between the two measures. However it is not obvious whether the large number of unique relevant documents is causing the scores to be lower, or whether the topics are so difficult that there are many relevant documents that can be found only by one system.

Figures 8 and 9 show two additional breakdowns of the sources of the relevant documents. Figure 8 shows the percentages contributed by each type of ad hoc run (and some tracks) to the judgment pool, and to the relevant documents. For example, whereas 35% of the pool for relevance judgments came from the automatic systems, only 6% of the unique relevant documents were found by these systems. The manual systems contributed 23% of the pool, and 29% of the unique relevant documents. This could mean either that these documents were simply not retrieved by the automatic systems in general, or that the automatic systems ranked them lower than nonrelevant documents.

Figure 9 gives a different view of the same issue by looking at the systems that retrieved the most unique relevant documents. It is noteworthy that most of this contribution involves manually produced results, although systems using "unusual" methods, such as the proximity ranking done by ANU and Waterloo, are finding somewhat more unique relevant documents than manual versions of more traditional systems such as ETH and Berkeley. Further investigation into the patterns of unique relevant documents in past TRECs is needed to help resolve these questions.

Obviously there are other factors as to why the TREC-5 topics seem to be harder. One such factor lies in the construction of the topics themselves, but this is difficult to measure objectively. A brief examination of the topics for TRECs 3, 4 and 5 suggests that the TREC-3 topics were more simplistic. People familiar with the workings of search engines are likely to be able predict which of these topics will be "easy", and which will not. The topics in TREC-4 and 5 appear to be more complex, i.e., they are asking for very specific information about multiple concepts rather than general information about a single area of interest.

These observations are consistent with the evolution of topic construction at NIST. The TREC-3 ad hoc topics (151–200) were the first topics "built" by the relevance assessors; earlier topics were constructed outside of NIST. The instructions for TREC-3 asked the relevance assessors to bring in some initial areas of interest (called "seeds"), and then they searched the document collections in order to create the topics. Because this was a somewhat artificial way of building topics, the TREC-4 ad hoc topics (201–250) were completely written before any interaction with the documents, and topics were selected based on the likely number of relevant documents that would be found by the systems (to eliminate very broad or very narrow topics). This same procedure was used in TREC-5, but by then the assessors were more experienced at building topics and may have built more difficult ones.

As a final comment, it is likely that the observed difficulty of the TREC-5 topics is due to a combination of factors. Even though the topics themselves are more complex, it is not easy to predict which topics will be difficult. There are interactions between the subject domains of the topics and the documents being searched, between the topics and the relevance assessments, and finally between the topics and the methodologies of query expansion and ranking of documents being used by the systems. More knowledge is needed about what makes some topics more difficult than others, especially if it can lead to what types of tools or systems are more appropriate for certain types of topics.

5.5 TREC-5 routing results

The routing evaluation used a specifically selected subset of the training topics against a new set of test documents. In TREC-4 there was difficulty obtaining new data for testing; the outcome was performing routing tests using the *Federal Register* (with new data) for 25 of the topics, and using training data and "net trash" for testing the other 25 topics. This situation was clearly not ideal and for TREC-5 NIST held back decisions on the routing topics until a new data source could be found.

When the FBIS data described earlier became available, it was decided to pick topics that had many relevant documents in the *Associated Press* data, on the assumption that the FBIS data would be similar to AP. Because of delays in getting and processing the data, this assumption could not be checked out, and problems arose that will be discussed later.

There were a total of 26 sets of results for routing evaluation, with 23 of them based on runs for the full data set. Of the 23 systems using the full data set,

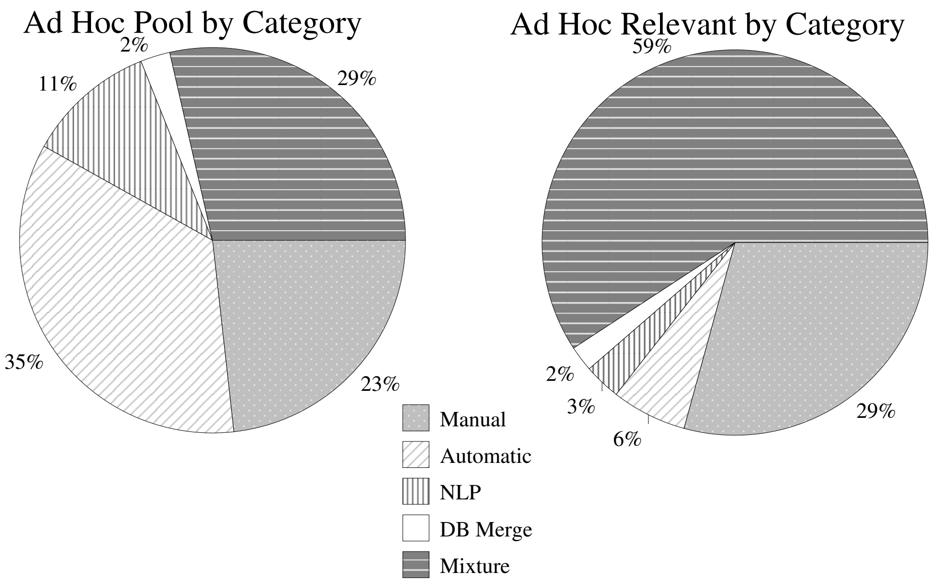

Figure 8: Distribution of categories in judged and relevant document pools.

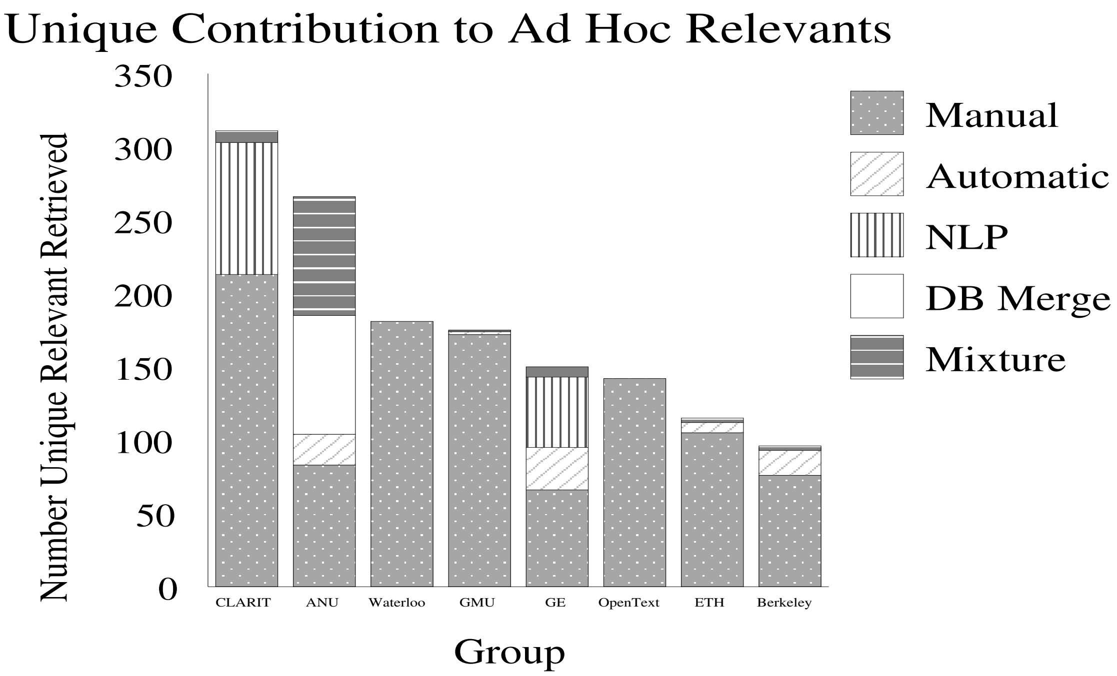

Figure 9: Percentage of unique relevant documents by category for groups retrieving many unique relevant documents.

21 used automatic construction of queries, and 2 used manual construction. There were 3 sets of category B routing results, all using automatic construction of queries.

Figure 10 shows the recall/precision curves for the eight TREC-5 groups with the highest non-interpolated average precision for the routing queries. The runs are ranked by the average precision based on the 39 topics that had more than five relevant documents.[3] A short 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.

city96r2 – City University, London ("Okapi at TREC-5" by M.M. Beaulieu, M. Gatford, Xiangji Huang, S.E. Robertson, S. Walker, and P. Williams) expanded on their query term selection method used in TREC-3 [18] (the predecessor of the Dynamic Feedback Optimization (DFO) algorithm developed by Cornell [3]). To combat overfitting of the training data, they divided the data into 3 partitions, using one to select the initial pool of terms, a second partition to do a final selection of the terms, and the third as an evaluation test set. After their "best" method was selected, the final queries were built using half the training data to select the initial pool, and the second half to do the final selection. Additionally they experimented with merging results from queries that used different term selection methods.

Cor5R1cc – Cornell ("Using Query Zoning and Correlation Within SMART: TREC 5" by Chris Buckley, Amit Singhal and Mandar Mitra) made three major revisions in their routing runs for TREC-5. The first change was the use of a "query zone" to subset the training documents into 5000 high-ranking non-relevant ones, in addition to the relevant ones, for use in expansion and reweighting (for further work, see [20]). They also used a complex process to expand and reweight the query terms, including a Rocchio method with positive and negative term weighting based more heavily on the training documents rather than on the topic. As a final change, they examined promising co-occurrence terms, with a total of 100 single terms, 10 phrases, and 50 co-occurring word pairs being added to the final query. The DFO algorithm was applied for fine tuning of all weights. These experiments yielded a total of 23% improvement over their TREC-4 algorithms.

INQ303 – University of Massachusetts at Amherst ("INQUERY at TREC-5" by James Allan, Jamie Callan, Bruce Croft, Lisa Bellesteros, John Broglio, Jinxi Xu and Hongmin Shu) used similar techniques to the new algorithms tried in TREC-4 [1]. These consisted of adding up to 250 single terms, adjacent word pairs, and nearby word pairs from the training documents. A complex selection process based on term occurrence in 200-word "best-match" passages was used. These terms and word pairs were then reweighted using the DFO algorithm for optimizing performance.

pircsg6 – Queens College, CUNY ("TREC-5 English and Chinese Retrieval Experiments using PIRCS" by K.L. Kwok and L. Grunfeld) implemented a genetic algorithm to select the best training subset as opposed to using only the short or high-ranking subdocuments for training as in TREC-4 [13]. This run is after six generations of "genetic growing", but produced results only 5% better than using the initial training subset (iteration 0).

Brkly14 – University of California, Berkeley ("Term importance, Boolean conjunct training, negative terms, and foreign language retrieval: probabilistic algorithms for TREC-5" by Fredric C. Gey, Aitao Chen, Jianzhang He, Liangjie Xu and Jason Meggs) used massive automatic query expansion using a chi-square discrimination measure similar to that used in TREC-3 [6]. There were an average of 2032 terms added to the initial query. Some experiments were also run involving maximizing the weighted contributions from the top 15 terms added by including these terms specifically in the final (50) regression equations used in retrieval.

uwgcrc0 – University of Waterloo ("Interactive Substring Retrieval (MultiText Experiments for TREC-5)" by Charles L.A. Clarke and Gordon V. Cormack) used queries that were manually built in a special query language called GCL. The routing experiments involved interactive query construction based on co-occurrence of substrings in the training documents.

[3] Of the 50 original routing topics, one was mistakenly not judged, four had no relevant documents in the test document set, and six had one or two relevant documents in the test set. The appendix includes the evaluation results over the 45 topics that had at least one relevant document.

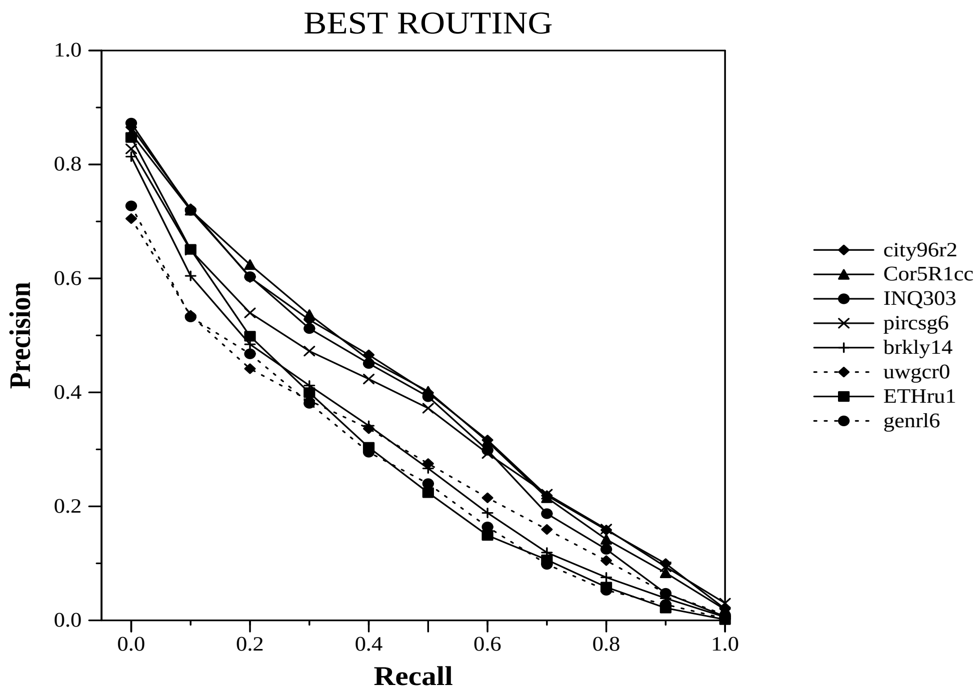

Figure 10: Recall/Precision graph for the top eight routing runs.

ETHru1 – Swiss Federal Institute of Technology (ETH) ("SPIDER Retrieval System at TREC-5" by Jean-Paul Ballerini, Marco Büchel, Ruxandra Domenig, Daniel Knaus, Bojidar Mateev, Elke Mittendorf, Peter Schäuble, Páraic Sheridan, and Martin Wechsler) investigated various feature selection methods, including the method used by OKAPI (RSV), the chi-square method, and the U method. They report that the U method worked consistently the best, likely because it is based on only positive correlations of feature occurrences. In addition to these experiments, they also tried experiments using cooccurring terms within sentences or paragraphs. These were formally motivated by the OKAPI RSV values to create independent indices of the documents, and the results were combined linearly using logistic regression for parameter discovery.

genrl6 – GE / Lockheed Martin / NYU / Rutgers ("Natural Language Information Retrieval: TREC-5 Report" by Tomek Strzalkowski, Louise Guthrie, Jussi Karlgren, Jim Leistensnider, Fang Lin, Jose Perez-Carballo, Troy Straszheim, Jin Wang and Jon Wilding) developed a new stream architecture to investigate combining of results from three different search engines using various types of data as input to those engines. This group has always concentrated on using more sophisticated natural language processing techniques to locate different types of language structures. These structures are problematic to combine for a single set of results, and the stream architecture provides the flexibility needed to investigate optimal combinations. The *genrl6* run combined four streams of input (stems, colocation, pairs and names) within the PRISE system with results from a new classification-based routing system.

It should be noted that the routing task in TREC has always served two purposes. The first is its intended purpose: to test systems in their abilities to use training data to build effective filters or profiles. The second purpose, which has become equally important in the more recent TRECs, is to serve as a learning environment for more effective retrieval techniques in general. Groups use the relevance judgments to explore the characteristics of relevant documents, such as which features are most effective to use for retrieval or how to best merge results from multiple queries. This is more profitable than simply using the previous TREC results in a retrospective manner because of the use of completely new testing data for evaluation.

A focus on using the training data as a learning environment was particularly prevalent in TREC-5.

Cornell used the relevant and non-relevant documents for investigations of Rocchio feedback algorithms, including more complex processes of expansion and weighting. Waterloo interactively searched the training data for co-occurring substrings and GE (with their partners Lockheed Martin, NYU, and Rutgers) ran major experiments in data fusion to test their new stream-based architecture. In each of these cases the experiments are assumed to lead to better ways of doing the routing task, and also to new approaches for the ad hoc task.

Three experimental themes dominated most routing experiments. The first is the discovery of optimal features (usually single terms) for use in the query or filter. City continued its experiments in repeatedly trying various combinations of terms to discover the optimal set, but for TREC-5 used subsets of the training data. Berkeley concentrated on further investigations of the use of the chi-square discrimination measure to locate large numbers of good terms, and ETH tested three different feature selection methods, including the chi-square method, the RSV (OKAPI) method, and a new method, the U measure. Xerox ("Xerox TREC-5 Site Report: Routing, Filtering, NLP, and the Spanish Tracks" by D. Hull, G. Grefenstette, B. Schulze, E. Gaussier, H. Schütze, and J. Pedersen) also investigated a new feature selection method, the binomial likelihood ratio test.

The second theme was the use of co-occurring term pairs in the training data to "expand" the query. Four groups experimented with locating and incorporating co-occurring pairs of terms, including INQUERY in both TREC-4 and TREC-5, and Cornell in TREC-5. As mentioned before, Waterloo interactively looked for word-pairs or co-occurring strings to manually add to their query. ETH used the OKAPI RSV values to formally motivate a series of experiments using co-occurring terms within different portions of the document (within sentence, within paragraph, etc.) as different methods of constructing queries. These multiple representations of the query were then linearly combined, with the parameters for that combination discovered using logistic regression on the training data.

The third theme in the routing experiments was the continuing effort to use only subsets of the training data. The number of judged documents per topic is on the order of 2000 or more, and this can be computationally difficult for complex techniques. Efficiency has motivated CUNY experiments (the PIRCS system) since TREC-3 where they tried using only the "short" documents for training. In TREC-5 this group used genetic algorithms to select the optimal set of training documents. Cornell (in TREC-5) used a new "query zone" technique to subset the training documents so that not all non-relevant documents were used for training. The goal was not just improved efficiency, but also improved effectiveness in that training was more concentrated on documents that the Cornell system is likely to retrieve.

There is another issue that suggests the use of subsets: the problem of overfitting the queries/methods to the training data. This was specifically emphasized in the City system, where they used different subsets of the training data for locating features, and used combinations of runs for their final results. Xerox used subsets to reduce overfitting, with their subsets based on finding documents within a "local zone" to the query (a predecessor to the query zoning technique used by Cornell). The Xerox paper provides more discussion of the overfitting problem and suggests some additional techniques to avoid it.

As in the ad hoc task, there is a heavy adoption rate across groups for successful techniques. For the ad hoc task these techniques revolve around better ways of handling the initial topic, or use of the top X documents for relevance feedback. Because of the existence of training data in routing, the routing experiments have generally not used the topic itself heavily, but constructed queries mainly based on the training data. The success of these techniques therefore revolves around how well the test data matches the training data, and also on how tuned the techniques are to the particular training data.

Figure 11 shows a comparison of routing results from several of the best performing systems in TRECs 3, 4, and 5. In general, the routing results show little improvement from TREC-3, although most of the systems have clearly developed superior methodologies. This is likely due to poorer matches between the training and test data for TRECs 4 and 5 than for TREC-3. The TREC-3 test data (Disk 3) was extremely similar to the training data, in that Disk 3 had similar content to that of Disks 1 and 2 (see Table 2 for details). In TREC-4 the training data was Disks 1, 2, and 3, but the test data was additional *Federal Register* material for 25 topics, and lots of "net trash" for the other 25 topics. The mixture of two distinctly different types of data, and the use of the very long *Federal Register* documents, made the routing task much more difficult.

TREC-5 used AP documents as training data, with FBIS material for test data. Whereas the types of documents are similar, the domains of the documents do not always match. So for some topics there is a

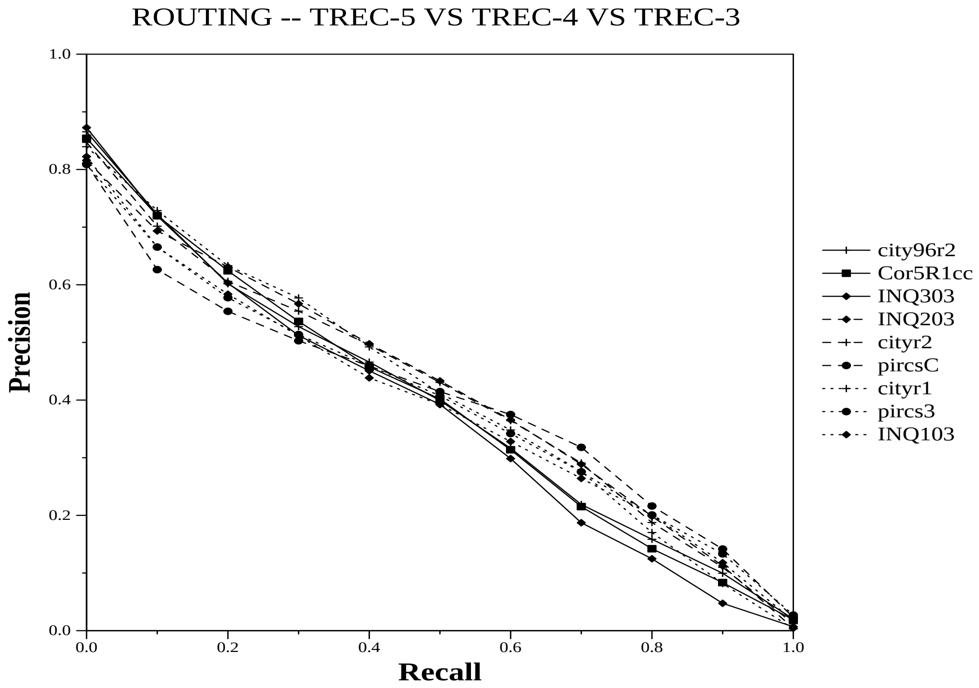

Figure 11: Recall/Precision graph for routing runs from 3 TRECs.

good match of training and test data, but for others the match is very poor, and very few relevant documents were found for those topics. Four topics had zero relevant documents in the test set, and an additional six topics had only one or two relevant documents. Even after dropping the four topics with no relevant documents from the evaluation, the results are still heavily affected by the mismatch.

This effect can be seen in Table 9 where the topic hardness measure (same definition as for the ad hoc results) is given, along with the number of relevant documents in both the test and training sets. As can be seen, there is a definite correlation between the number of relevant documents in the test set and the hardness of the topics (correlation coefficient of 0.812), and considerably less of a correlation (0.497) between the hardness and the number of relevant training documents. The table is sorted by the number of relevant test documents, and the average hardness for topics with ten or fewer relevant documents is 0.085, way below that for topics with many more relevant documents.

Note that this strong correlation between hardness and the number of relevant documents did not occur in the ad hoc task, where there is a very low correlation (less than random) between these factors. Whereas it is difficult to understand the factors that make the ad hoc topics "harder", it is obvious that one big factor in the "harder" routing topics is the degree of match between the training data and the test data. However, this is a real-world constraint, since in any operational system there will be differences in the degrees of match and systems will need to be able to recognize "mismatches" and adapt for them.

In TREC-6 an attempt will be made to have a close match between the training and test data. This should create a second comparison point for well-matched data with the improved systems, and allow further examination of the effects of the training data issues.

6 Summary

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

However, two main conclusions can be drawn from TREC-5:

- Systems seem to be adjusting to the much

Table 9: Number relevant documents in training and test sets and hardness for TREC-5 routing topics

Topic	Training Relevant	Test Relevant	Hardness
53	389	1	0.0435
207	31	1	0.1304
224	50	1	0.0000
211	190	2	0.1087
222	55	2	0.0000
243	34	2	0.0217
125	183	6	0.1739
23	141	7	0.0994
194	76	8	0.0489
44	57	10	0.0348
192	179	10	0.2783
77	179	14	0.2329
185	105	14	0.1273
173	203	15	0.0957
126	252	18	0.3913
5	40	19	0.2906
154	450	22	0.2945
1	111	30	0.2058
78	200	37	0.3161
24	145	38	0.1968
114	221	42	0.1915
58	144	45	0.3652
54	157	48	0.3333
82	217	55	0.3613
94	119	56	0.1762
123	262	57	0.2853
228	31	68	0.1988
240	168	88	0.1354
11	231	92	0.2864
95	92	93	0.1580
3	74	101	0.3852
161	133	153	0.7061
100	107	157	0.5361
6	129	158	0.3891
108	266	174	0.2839
4	41	178	0.5400
119	461	185	0.3052
221	84	193	0.2687
187	147	194	0.3483
12	272	228	0.4922
118	559	324	0.5109
202	99	583	0.5987
189	882	584	0.6135
142	847	808	0.9126
111	235	887	0.7996

shorter topics in the ad hoc task. Most of the automatic expansion methods tried in TREC-4 were used again in TREC-5, but with significant adjustments to handle the short topic. These adjustments tended to center around getting more information from the topic itself, rather than just extracting keywords. However, comparison runs using the full or long topic still produced over 20% improvement in performance in most cases. The ad hoc runs using manually-built queries (mostly) involved interactivity, since the query construction rules changed in TREC-5 to allow this. Groups either tested human-computer "teamwork", or involved users in order to better learn how to automatically build ad hoc queries.

- The routing results for TREC-5 were somewhat disappointing. Whereas there were many groups with significant improvements in performance, the overall results were not better than for TREC-4 (or for TREC-3). The problem appears to be the serious mismatch between the training and the test data, which unexpectedly has happened in both TREC-4 and TREC-5. In TREC-5 there was a domain mismatch for many of the topics, resulting in very few relevant documents. While this is not an unrealistic problem — and systems must learn to adapt to it — providing a better match between the training and test data in the future will enable a better evaluation of the new routing methods.

The six TREC-5 tracks significantly expanded the amount of research performed in TREC. Many of the tracks further explored the work initiated in the preliminary running of the track in TREC-4.

Interactive: This was the second running of the interactive track. Based on the lessons learned from the TREC-4 track on how difficult it is to fairly compare results in interactive experiments, the track concentrated on experimental design in TREC-5. Unfortunately, the final design was not decided until late in the TREC cycle, and only two groups were able to participate. The track will continue in TREC-6 using the experimental design developed in TREC-5.

Database merging: This was also the second running of the database merging track, with three groups participating in the track in TREC-5. The track has proved to be a high-overhead track (this year's task required creating 98 separate databases), and thus has not attracted much participation despite general interest in the problem. The track will likely be run again in future TRECs, but will not be run in TREC-6.

Multilingual: Seven groups submitted Spanish runs and nine groups submitted Chinese runs. As in TREC-4, the Spanish results demonstrated that many of the techniques used in English retrieval can be successfully applied to Spanish. Given the success of traditional techniques on Spanish, it was decided to discontinue the Spanish portion of the multilingual track.

This was the first year for Chinese in TREC, and most groups concentrated on segmentation issues. The Chinese track will continue in TREC-6.

Confusion: A confusion (or data corruption) track was run in TREC-4 in which characters were randomly changed to simulate the type of output one might get from an Optical Character Recognition (OCR) process. In TREC-5, the test data was actual OCR output of scanned images of the 1994 *Federal Register*. Five groups participated in the experiment designed to explore the effect different levels of OCR error has on retrieval performance.

The track introduced the known-item search as a new task for TREC. The known-item search task will be used again in the Spoken Document Retrieval (SDR) track, a new track to begin in TREC-6. The SDR track is a successor to the confusion track in that it represents a different form of "corrupted" documents. Instead of retrieving documents that are the result of OCR, systems will retrieve documents that are the result of speech recognition systems. The interaction between OCR and retrieval will continue to be explored in the new METTREC workshop.

Filtering: The TREC-5 filtering track followed the same design as the preliminary track in TREC-4, and had seven participating groups. The goal in the track was to retrieve an unranked set of documents that optimizes a pre-specified utility function. A family of three functions was used to investigate how retrieval was affected by changes in the relative worth of retrieving a relevant document versus not retrieving a nonrelevant document. The track will continue into TREC-6 with a different set of utility functions and a set of test documents that more closely matches the training data.

NLP: Four groups participated in the initial running of the natural language processing track. The track will continue in TREC-6.

In addition to the tracks above, TREC-5 had a "trial" run of the Very Large Corpus (VLC) track. The VLC track will have it first official running in TREC-6, where participants will perform ad hoc searches on approximately 20 gigabytes of text. Other new tracks in TREC-6 will be a Cross Language track, in which systems use topics in one language to retrieve documents in a second language, and a High-Precision track, where participants attempt to retrieve the best 10 documents for each topic using no more than five minutes (wall clock time) for each topic.

Acknowledgments

The authors would like to gratefully acknowledge the continued support of the TREC conferences by the Intelligent Systems Office of the Defense Advanced Research Projects Agency. Special thanks also go to the TREC program committee and the staff at NIST.

References

[1] J. Allan, L. Ballesteros, J. Callan, B. Croft, and Z. Lu. Recent experiments with INQUERY. In *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*, pp. 49–63, 1996. NIST Special Publication 500-236.

[2] Eric Brown. Fast evaluation of structured queries for information retrieval. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 30–38, 1995.

[3] C. Buckley and G. Salton. Optimization of relevance feedback weights. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 351–357, 1995.

[4] Robert Burgin. Variations in relevance judgments and the evaluation of retrieval performance. *Information Processing and Management*, 28(5):619–627, 1992.

[5] W. Cooper, A. Chen, and F. Gey. Full text retrieval based on probabilistic equations with coefficients fitted by logistic regression. In *The Second Text REtrieval Conference (TREC-2)*, pp. 57–66, 1994. NIST Special Publication 500-215.

[6] W. Cooper, A. Chen, and F. Gey. Experiments in the probabilistic retrieval of full text documents. In *Overview of the Third Text REtrieval Conference (TREC-3) [Proceedings of TREC-3.]*, pp. 127–134, 1995. NIST Special Publication 500-225.

[7] Larry Fitzpatrick and Mei Dent. Automatic feedback using past queries: Social searching? In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-97)*, pp. 306–313, 1997.

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

[9] 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)*, pp. 1–23, October 1996. NIST Special Publication 500-236.

[10] Stephen P. Harter. Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science*, 47(1):37–49, 1996.

[11] D. Hawking and P. Thistlewaite. Searching for meaning with the help of a PADRE. In *Overview of the Third Text REtrieval Conference (TREC-3) [Proceedings of TREC-3.]*, pp. 257–267, 1995. NIST Special Publication 500-225.

[12] K.L. Kwok. A new method of weighting query terms. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 187–196, 1996.

[13] K.L. Kwok and L. Grunfeld. TREC-4 ad-hoc, routing retrieval and filtering experiments using PIRCS. In *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*, pp. 145–152, 1996. NIST Special Publication 500-236.

[14] M. E. Lesk and G. Salton. Relevance assessments and retrieval system evaluation. *Information Storage and Retrieval*, 4:343–359, 1969.

[15] D. Evans N. Milić-Frayling and R. Lefferts. CLARIT TREC-4 experiments. In *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*, pp. 305–321, 1996. NIST Special Publication 500-236.

[16] A. Moffat and J. Zobel. Information systems for large document collections. In *Overview of the Third Text REtrieval Conference (TREC-3) [Proceedings of TREC-3.]*, pp. 85–93, 1995. NIST Special Publication 500-225.

[17] S.E. Robertson, S. Walker, and M.M. Hancock-Beaulieu. Large test collection experiments on an operational, interactive system: Okapi at TREC. *Information Processing and Management*, 31(3):345–360, 1995.

[18] S.E. Robertson, S. Walker, S. Jones, M.M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Overview of the Third Text REtrieval Conference (TREC-3) [Proceedings of TREC-3.]*, pp. 109–126, 1995. NIST Special Publication 500-225.

[19] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 21–29, 1996.

[20] A. Singhal, M. Mitra, and C. Buckley. Learning routing queries in a query zone. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 25–33, 1997.

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

[22] 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.

[23] 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.]*, pp. 385–398, April 1995. NIST Special Publication 500-225.

[24] J. Xu and W.B. Croft. Query expansion using local and global document analysis. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 4–11, 1996.