Overview of the Eighth Text REtrieval Conference (TREC-8)

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

1 Introduction

The eighth Text REtrieval Conference (TREC-8) was held at the National Institute of Standards and Technology (NIST) on November 16–19, 1999. The conference was co-sponsored by NIST and the Information Technology Office of the Defense Advanced Research Projects Agency (DARPA).

TREC-8 is the latest in a series of workshops designed to foster research in text retrieval. For analyses of the results of previous workshops, see Tague-Sutcliffe and Blustein [11], Harman [4], and Sparck Jones [10]. 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-8. Sixty-six groups including participants from 16 different countries 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. It concentrates on the main task, *ad hoc retrieval*, which is defined in the next section. Details regarding the test collections and evaluation methodology used in TREC follow in sections 3 and 4, while section 5 provides an overview of the ad hoc retrieval results. In addition to the main ad hoc task, TREC-8 contained seven "tracks," tasks that focus research on particular subproblems of text retrieval. Taken together, the tracks represent the bulk of the experiments performed in TREC-8. However, each track has its own overview paper included in the proceedings, so this paper presents only a short summary of each track in section 6. The final section looks forward to future TREC conferences.

2 The Ad Hoc Retrieval Task

The ad hoc retrieval task investigates the performance of systems that search a static set of documents using new questions (called *topics* in TREC). 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. NIST provides the participants approximately 2 gigabytes worth of documents and a set of 50 natural language topic statements. The participants produce a set of *queries* from the topic statements and run those queries against the documents. The output from this run is the official test result for the ad hoc task. Participants return the best 1000 documents retrieved for each topic to NIST for evaluation.

Table 1: Organizations participating in TREC-8

ACSys
AT&T Labs Research
CL Research
CLARITECH Corporation
Cambridge University
Carnegie Mellon University
Center for Information Research, Russia
City University/Microsoft
Cymfony Inc.
DSO National Laboratories, Singapore
Dartmouth College
Dublin City University
Eurospider Information Technology
Fondazione Ugo Bordoni
Fujitsu Laboratories, Ltd.
GE/Rutgers/SICS/UHelsinki/UPenn
IBM T. [redacted]. Watson Research Center (2 groups)
IIT/AAT/NCR
IRIT/SIG
Imperial College
Informatique-CDC
[redacted]ohns Hopkins University
KDD R&D Laboratories
Kasetsart University
LIMSI-CNRS (2 groups)
MIT Laboratory for Computer Science
MITRE
Management Information Technologies, Inc.
Microsoft Research Ltd
MuliText Project
NTT DATA Corporation
National Taiwan University
New Mexico State University
Oracle
Oregon Health Sciences University
Oslo College
Queens College, CUNY
RICOH Co., Ltd.
RMIT
Rutgers University (3 groups)
Sabir Research/Cornell University
Seoul National University
Sharp Laboratories of Europe Ltd.
Southern Methodist University
State University of New York at Buffalo
TextWise, Inc.
The University of Sheffield, UK
TwentyOne
USheffield/CambridgeU/SoftSound/ICSI Berkeley
Universite de Montreal
Universite de Neuchatel
University of California, Berkeley
University of Iowa
University of Maryland, College Park
University of Massachusetts
University of North Carolina (2 groups)
University of North Texas
University of Ottawa
University of Surrey
University of Twente
Xerox Research Centre Europe

Participants are free to use any method they desire to create the queries from the topic statements. TREC distinguishes among two major categories of query construction techniques, automatic methods and manual methods. An automatic method is a means of deriving a query from the topic statement with no manual intervention whatsoever; a manual method is anything else. The definition of manual query construction methods is very broad, ranging from simple tweaks to an automatically derived query, through manual construction of an initial query, to multiple query reformulations based on the document sets retrieved. Since these methods require radically different amounts of (human) effort, care must be taken when comparing manual results to ensure that the runs are truly comparable.

The right answers, called *relevance judgments*, for the ad hoc topics are not known at the time the participants produce their runs, though participants may use the documents, topics, and relevance judgments from previous TRECs to develop their systems. Participants are also free to use other sources of training data if they desire. Fifty new topics (401–450) were created for the TREC-8 ad hoc task. The set of documents used in the task was the documents contained on TREC Disks 4 and 5, excluding the *Congressional Record* subcollection. (See section 3.1 for details about this document set.) Disks 4 and 5 have now been used as the set of test documents for TRECs 6, 7, and 8 to produce a test collection with 150 topics.

Participants were allowed to submit up to five ad hoc runs to NIST. The runs could differ as the result of using different query construction techniques, or using different searching methods with the same queries.

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

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

When submitting a run, participants were required to state whether the queries were produced manually or automatically. If any run used an automatic method, participants were required to submit a run that used just the “title” and “description” fields of the topic statements (see section 3.2 for a description of the topics).

3 The Test Collections

Like most traditional retrieval collections, there are three distinct parts to the collections used in TREC: the documents, the topics, and the relevance judgments. This section describes each of these pieces for the ad hoc collection.

3.1 Documents

TREC documents are distributed on CD-ROM’s with approximately 1 GB of text on each, compressed to fit. For TREC-8, Disks 1–5 were all available as training material (see table 2) and Disks 4–5 were used for the ad hoc task. The *Congressional Record* subcollection on Disk 4 was excluded from the test document set.

Documents are tagged using SGML to allow easy parsing (see fig. 1). The documents in the different datasets have been tagged with identical major structures but they have different minor structures. The philosophy in the formatting at NIST is to leave the data as close to the original as possible. No attempt is made to correct spelling errors, sentence fragments, strange formatting around tables, or similar faults.

```
<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 1: A document extract from the *Financial Times*.

3.2 Topics

The format of the TREC topics has evolved over time as illustrated in table 3. The table shows the number of words included in the different parts of the topic statements for each TREC. The original ad hoc topics (51–150) were very detailed, containing multiple fields and lists of concepts related to the topic subject. The ad hoc topics used in TREC-3 (151–200) did not contain the concept lists and and the remaining fields were generally shorter than in 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. The TREC-4 topics (201–250) were therefore made even shorter: a single field consisting of a one sentence description of the information need. However, the one-sentence topic eliminated from the topic the statement of the criteria used to judge a document as relevant—which was one of the motivating factors for providing topic statements rather than queries. The last four sets of ad hoc topics (251–450) have therefore all had the same format as in TREC-3, consisting of a title, description, and narrative. A sample TREC-8 topic is shown in figure 2.

The different parts in the most recent TREC topics allow participants to investigate the effect of different query lengths on retrieval performance. The "titles" in topics 301–450 have been specially designed to allow experiments with very short queries. The titles consist of up to three words that best describe the topic. The description field is a one sentence description of the topic area. As in TREC-7, the description field of TREC-8 topics contains all of the words in the title field, to remove the confounding effects of word choice on length experiments. The narrative gives a concise description of what makes a document relevant.

Ad hoc topics have been constructed by the same person who performed the relevance assessments for that topic (called the *assessor*) since TREC-3. Each assessor comes to NIST with ideas for topics based on his or her own interests, and searches the ad hoc collection (looking at approximately 100 documents per topic) to estimate the likely number of relevant documents per candidate topic. Because the same topic set was also to be used in the Web track this year, once the assessors searched the ad hoc collection with a candidate topic, they also searched the small Web collection using that topic. The NIST TREC team selected the final 50 topics from among the candidates based on the estimated number of relevant documents

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

	Min	Max	Mean		Min	Max	Mean
TREC-1 (51–100)	44	250	107.4	TREC-5 (251–300)	29	213	82.7
title	1	11	3.8	title	2	10	3.8
description	5	41	17.9	description	6	40	15.7
narrative	23	209	64.5	narrative	19	168	63.2
concepts	4	111	21.2				
TREC-2 (101–150)	54	231	130.8	TREC-6 (301–350)	47	156	88.4
title	2	9	4.9	title	1	5	2.7
description	6	41	18.7	description	5	62	20.4
narrative	27	165	78.8	narrative	17	142	65.3
concepts	3	88	28.5				
TREC-3 (151–200)	49	180	103.4	TREC-7 (351–400)	31	114	57.6
title	2	20	6.5	title	1	3	2.5
description	9	42	22.3	description	5	34	14.3
narrative	26	146	74.6	narrative	14	92	40.8
TREC-4 (201–250)	8	33	16.3	TREC-8 (401–450)	23	98	51.8
				title	1	4	2.5
				description	5	32	13.8
				narrative	14	75	35.5

```
<num> Number:  409
<title> legal, Pan Am, 103

<desc> Description:
What legal actions have resulted from the destruction
of Pan Am Flight 103 over Lockerbie, Scotland, on
December 21, 1988?
<narr> Narrative:
Documents describing any charges, claims, or fines
presented to or imposed by any court or tribunal are
relevant, but documents that discuss charges made in
diplomatic jousting are not relevant.
```

Figure 2: A sample TREC-8 topic.

in both collections and balancing the load across assessors.

3.3 Relevance assessments

Relevance judgments are of critical importance to a test collection. For each topic it is necessary to compile a list of relevant documents—as comprehensive a list as possible. All TRECs have used the pooling method [9] to assemble the relevance assessments. In this method a pool of possible relevant documents is created by taking a sample of documents selected by the various participating systems. This pool is then shown to the human assessor, who makes a binary (yes/no) relevance judgment for each document in the pool. Unjudged documents are assumed to be not relevant. The particular sampling method used in TREC is to take the top 100 documents retrieved per judged 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

Table 4: Overlap of submitted results

	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
TREC-6	3,430	1445 (42 %)	92 (06 %)
ad hoc	3100	1326	89
NLP	200	113	2
HP	130	6	1
TREC-7	7,805	1611 (21 %)	93 (06 %)
ad hoc	7700	1605	92
HP	105	6	.5
TREC-8	7,100	1736 (25 %)	94 (05 %)

most likely to be relevant returned first. Each pool is sorted by document identifier so assessors cannot tell if a document was highly ranked by some system or how many systems (or which systems) retrieved the document.

To keep the assessment task manageable, only a subset of the runs that are submitted to NIST are judged (that is, contribute to the assessment pools). When participants submit their runs, they rank the submissions in the order they prefer them to be judged. NIST ensures that the same number of runs from each participant is used when creating the pools (provided the participant has submitted that many runs), and uses the top 100 documents for every topic from every judged run. This strategy does not take advantage of recent proposals such as those by Zobel [14] or Cormack, Palmer, and Clarke [2] for finding more relevant documents in fewer total documents judged. Besides being difficult to handle logistically at NIST, there are concerns about how implementing these proposals might bias the assessments. Zobel suggests judging more documents for topics that have had many relevant documents found so far and fewer documents for topics with fewer relevant documents found so far as a way to improve the completeness of the pools. However, assessors would know that documents added later in the pools came from lower in the systems' rankings and that may affect their judgments. Cormack et al. suggest judging more documents from runs that have returned more relevant documents recently and fewer documents from runs that have returned fewer relevant document recently. But that would bias the pools towards systems that retrieve relevant documents early in their rankings. For test collections, a lack of bias in the relevance judgments is more important than the total number of relevant documents found.

3.3.1 Overlap

Table 4 summarizes the amount of overlap in the ad hoc pool for each TREC. The first data column in the table gives the maximum possible size of the pool. Since the top 100 documents from each run are judged, this number is usually 100 times the number of runs used to form the pool, though in some years track runs contributed fewer than 100 documents. The next column shows the number of documents that were actually in the pool (i.e., the number of unique documents retrieved in the top 100 across all judged runs) averaged over the number of topics. The percentage given in that column is the size of the actual pool relative to the

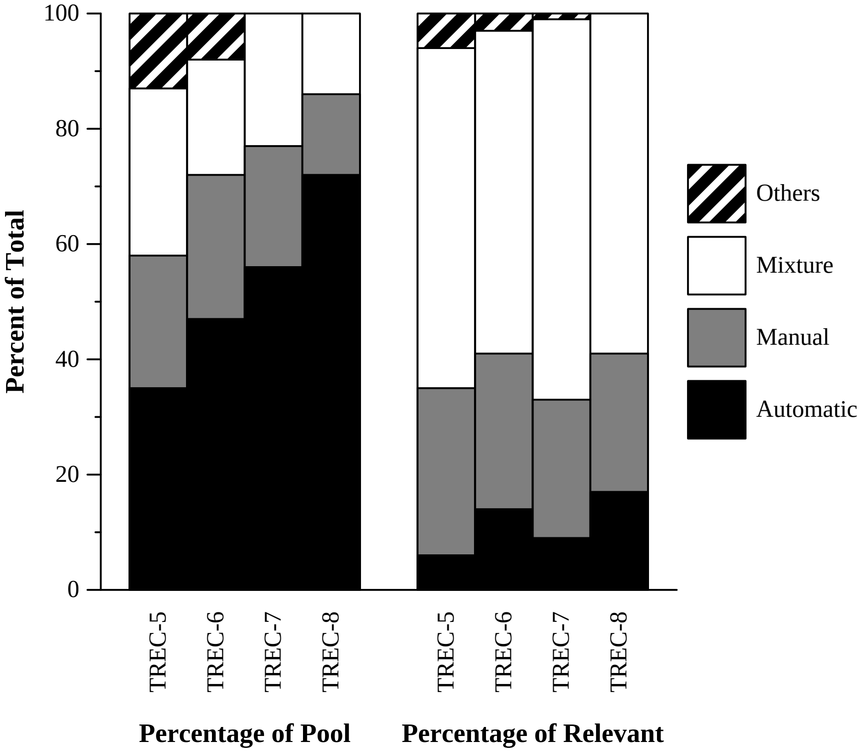


Figure 3: Mean percentage of pools and relevant documents by category. Means are computed over the 50 topics contained in the test sets.

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. For TRECs 4–7, various tracks also contributed documents to the ad hoc pool. These are broken out in the appropriate rows within table 4, where the order of the tracks is significant (a document retrieved in a track listed later is not counted for that track if the document was also retrieved by a track listed earlier). The TREC-8 ad hoc pools were created only from ad hoc runs.

The average overlap found in the pools has been stable since TREC-4 except for TREC-6. The tremendous drop in the size of the ad hoc pool for that year reflects the difference in the number of runs NIST was able to assess that year. Table 4 also shows that the average number of relevant documents per topic has remained stable after decreasing from an early high. NIST has deliberately chosen more tightly focused topics to better guarantee the completeness of the relevance assessments.

The figures for average overlap given in table 4 hide details about the source of the documents in the pool. Figure 3 shows the mean percentage of the total pool and the mean percentage of the total number of relevant documents contributed by each type of ad hoc run for TRECs 5–8. In the figure, “Automatic” designates documents that were retrieved only be automatic runs, “Manual” designates documents that were retrieved only by manual runs, “Mixture” designates documents that were retrieved by runs of different types, and “Others” designates documents that were retrieved by other tracks that contributed to the ad hoc pools. For example, for TREC-8 72 % of the pool came from the automatic runs, 14 % of the pool from manual runs, and 14 % of the pool came from runs of both types. In contrast, 17 % of the relevant documents came from automatic runs, 24 % of the relevant documents came from manual runs, and 59 % of the relevant documents came from runs of both types. For each of the years shown, the majority of the relevant documents were retrieved by multiple categories of runs. Manual runs retrieved a higher percentage of the relevant documents than they contributed to the pools.

Figure 4 gives a different view of the same issue by looking at the groups that retrieved unique relevant documents—relevant documents that were contributed to the pool by exactly one group. The figure contains a histogram of the total number of unique relevant documents found by a group over the 50 test topics for each of the last four TRECs. The totals are subdivided using the same categories as were used in figure 3.

Each of the histograms in the figure uses the same scale. A dot underneath the x-axis indicates a group is plotted there, and all groups that retrieved at least one unique relevant document are plotted. For each year, the majority of unique documents was retrieved by manual runs. The distribution of unique relevant documents found has been roughly the same over the four years.

3.3.2 Effect of pooling on evaluation

Some people object to the use of pooling to produce a test collection because unjudged documents are assumed to be not relevant. They argue that evaluation scores for methods that did not contribute to the pools will be deflated relative to methods that did contribute because the non-contributors will have highly ranked unjudged documents. Displays such as figure 4 contribute to these fears since they demonstrate that some relevant documents are found by only one group. If that group had not participated, those relevant documents would not have been judged.

Zobel demonstrated that the quality of the pools (the number and diversity of runs contributing to the pools and the depth to which those runs are judged) does affect the quality of the final collection [14]. He also found that the TREC collections were not biased against unjudged runs. In this test, he evaluated each run that contributed to the pools using both the official set of relevant documents published for that collection and the set of relevant documents produced by removing the relevant documents uniquely retrieved by the run being evaluated. For the TREC-5 ad hoc collection, he found that using the unique relevant documents increased a run's 11 point average precision score by an average of $\frac{1}{2}$ %. The maximum increase for any run was 3.5 %. The average increase for the TREC-3 ad hoc collection was somewhat higher at 2.2 %.

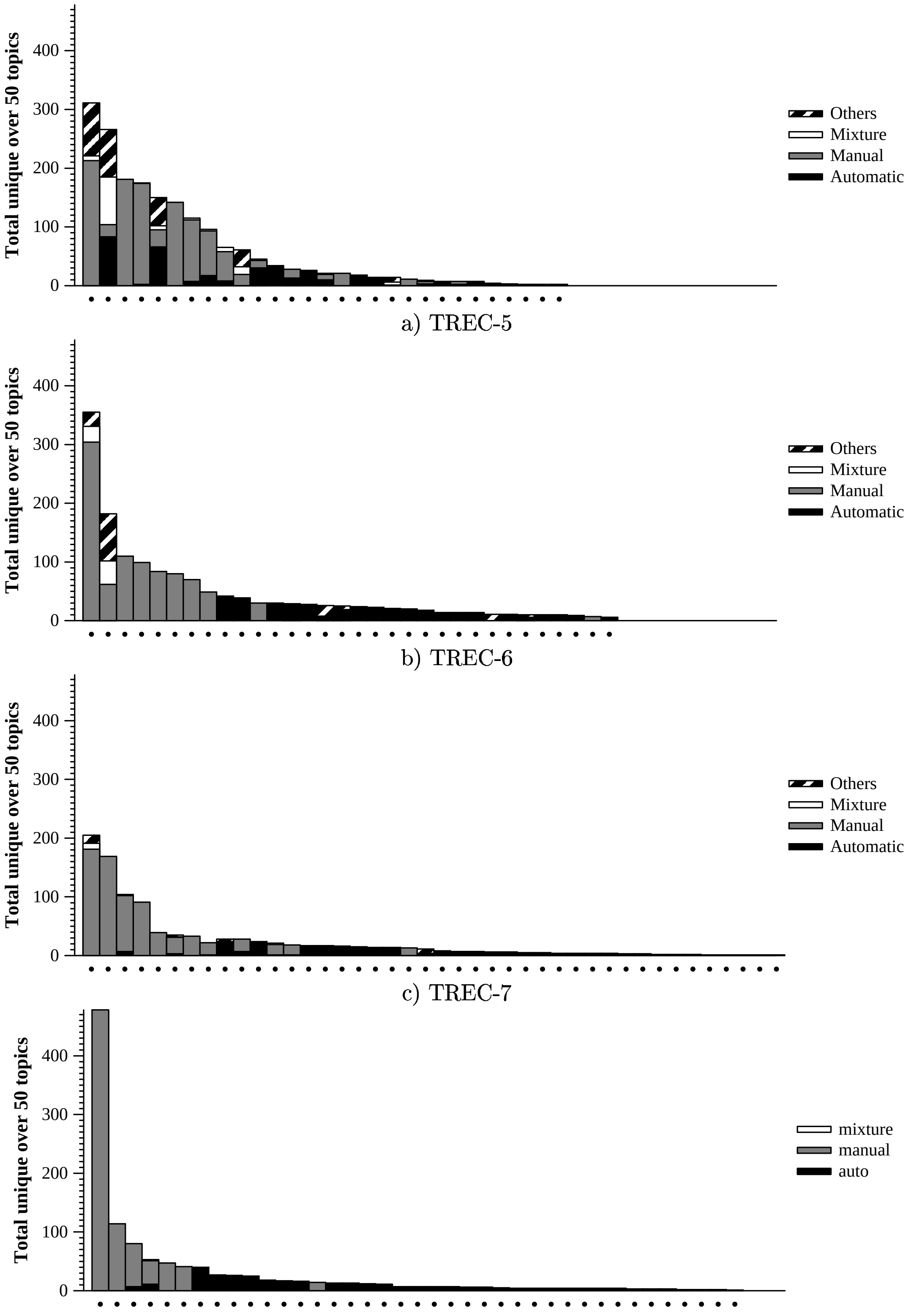
We can perform a similar check of a test collection as soon as relevance judging is complete. For each run that contributed to the pool, we compute the mean average precision (see section 4 for a definition of mean average precision) of the run using the standard relevance judgments and the set of relevance judgments produced by removing the relevant documents uniquely retrieved by that run's group. The mean percentage difference in mean average precision over the 71 runs that contributed to the ad hoc pool was 0.78 %, with a maximum difference of 9.9 %. Not surprisingly, the manual groups that had the largest number of unique relevant documents (see figure 4) also had the largest percentage differences in mean average precision. But given that the manual runs' contributions are in the pool, the difference in evaluation results for automatic runs is negligible. For automatic runs, the largest percentage difference in mean average precision scores was 3.85 %, which corresponded to an absolute difference of only .0001. Every automatic run that had a mean average precision score of at least .1 had a percentage difference of less than 1 %.

Figure 5 shows the absolute difference in mean average precision scores plotted against the number of unique relevant documents contributed by that run's group for each automatic run. The runs are sorted by increasing difference and then by number of unique relevant documents. The two obvious outliers in number of unique relevant documents (for runs `GE8ATDN1` and `iit99au1`) reflect organizations that submitted manual runs in addition to automatic runs; the vast majority of their unique relevant documents were contributed by their manual run.

While the lack of any appreciable difference in the scores of the automatic runs is not a guarantee that all relevant documents have been found, it is very strong evidence that the test collection is reliable for comparative evaluations of retrieval runs. Note that differences of less than 1 % are smaller than the differences that result from using different relevance assessors [12]. The quality of the pools is significantly enhanced by the presence of the recall-oriented manual runs. The organizers of the NTCIR (NACSIS Test Collection for evaluation of Information Retrieval systems) workshop took good advantage of this effect by performing their own manual runs to supplement their pools [6].

4 Evaluation

The entire purpose of building a test collection is to be able to compare the effectiveness of retrieval systems. Providing a common evaluation scheme is an important element of TREC.



a) TREC-5

b) TREC-6

c) TREC-7

d) TREC-8

Figure 4: Total number of unique relevant documents retrieved per TREC. Each total gives the percentages of the total that were retrieved by Automatic, Manual, Mixed, or Other runs. Groups are indicated by a dot beneath the x-axis. All groups that retrieved at least one unique relevant document are plotted.

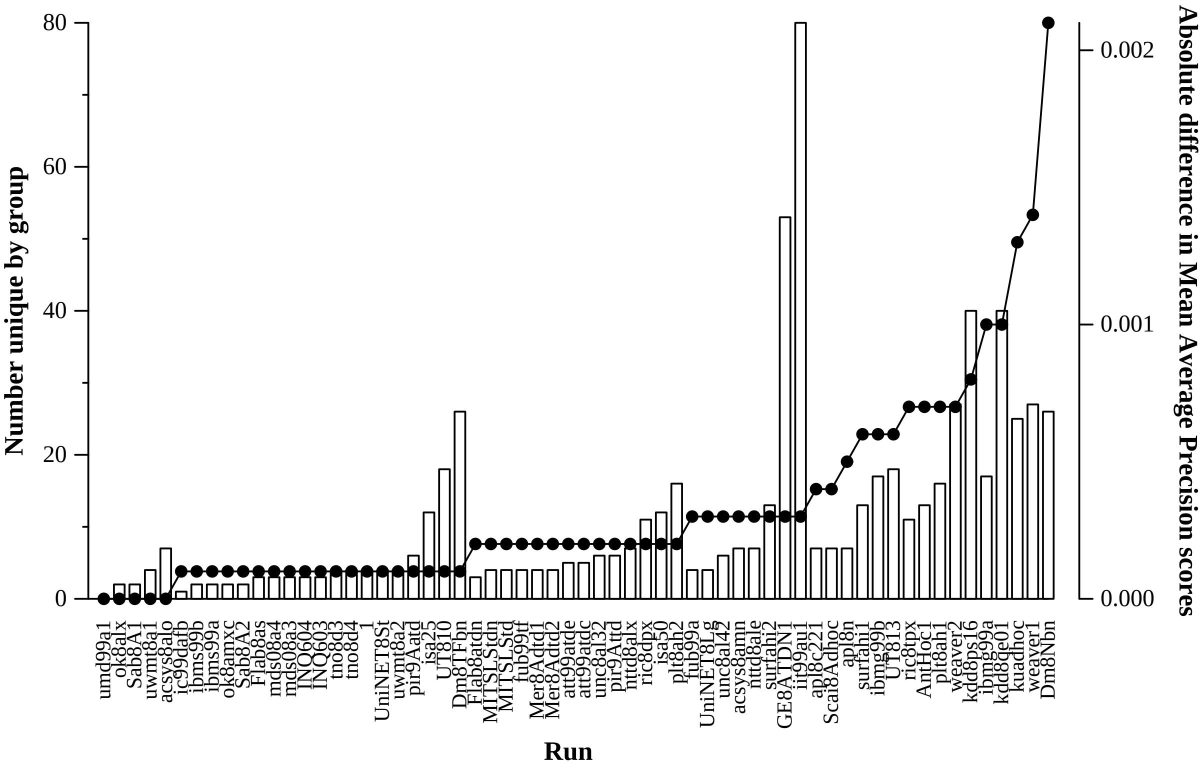


Figure 5: Absolute difference in mean average precision scores when a run is evaluated using relevance pools with and without that group's unique relevant document for TREC-8 automatic, ad hoc runs. Also plotted is the number of unique relevant documents contributed to the pools by that group. Runs are orderd by increasing absolute difference and by increasing number of unique relevant documents.

4.1 Current practice

All TREC tasks that involve returning a ranked list of documents are evaluated using the `trec_eval` package. This package, written by Chris Buckley, reports about 85 different numbers for a run. The measures reported include *recall* and *precision* at various cut-off levels plus single-valued summary measures that are derived from recall and precision. Precision is the proportion of retrieved documents that are relevant, while recall is the proportion of relevant documents that are retrieved. A cut-off level is a rank that defines the retrieved set; for example, a cut-off level of ten defines the retrieved set as the top ten documents in the ranked list. The `trec_eval` program reports the scores as averages over the set of topics where each topic is equally weighted. (The alternative is to weight each relevant document equally and thus give more weight to topics with more relevant documents. Evaluation of retrieval effectiveness historically weights topics equally since all users are assumed to be equally important.)

Precision reaches its maximal value of 1.0 when only relevant documents are retrieved, and recall reaches its maximal value (also 1.0) when all the relevant documents are retrieved. Note, however, that these theoretical maximum values are not obtainable as an average over a set of topics at a single cut-off level because different topics have different numbers of relevant documents. For example, a topic that has fewer than ten relevant documents will have a precision score less than one after ten documents are retrieved regardless of how the documents are ranked. Similarly, a topic with more than ten relevant documents must have a recall score less than one after ten documents are retrieved. At a single cut-off level, recall and precision reflect the same information, namely the number of relevant documents retrieved. At varying cut-off levels, recall and precision tend to be inversely related since retrieving more documents will usually increase recall while degrading precision and vice versa.

This overview paper generally uses two evaluation measures when discussing retrieval results, the recall-precision curve and mean (non-interpolated) average precision. A recall-precision curve plots precision as a function of recall. Since the actual recall values obtained for a topic depend on the number of relevant documents, the average recall-precision curve for a set of topics must be interpolated to a set of standard recall values. The particular interpolation method used is given in Appendix A, which also defines many of the other evaluation measures reported by `trec_eval`. Recall-precision graphs show the behavior of a retrieval run over the entire recall spectrum.

Mean average precision is the single-valued summary measure used when an entire graph is too cumbersome. The average precision for a single topic is the mean of the precision obtained after each relevant document is retrieved (using zero as the precision for relevant documents that are not retrieved). The mean average precision for a run consisting of multiple topics is the mean of the average precision scores of each of the individual topics in the run. The average precision measure has a recall component in that it reflects the performance of a retrieval run across all relevant documents, and a precision component in that it weights documents retrieved earlier more heavily than documents retrieved later. Geometrically, mean average precision is the area underneath a non-interpolated recall-precision curve.

The (reformatted) output of `trec_eval` for each submitted run is given in Appendix A. In addition to the ranked results, participants are also asked to submit data that describes their system features and timing figures to allow a primitive comparison of the amount of effort needed to produce the corresponding retrieval results. These system descriptions are not included in the printed version of the proceedings due to their size, but they are available on the TREC web site (`http://trec.nist.gov`).

5 Ad Hoc Retrieval Results

This section briefly summarizes some of the approaches used for the ad hoc task. To recap the specific task that was to be performed, the TREC-8 ad hoc task entailed using new topics 401–450 to search the documents on Disks 4 and 5 minus the *Congressional Record* documents. A run was either automatic or manual. For an automatic run, all processing was done by the machine without human intervention of any sort. All other runs were manual runs. There were 129 ad hoc runs submitted for the task: 13 manual runs and 116 automatic runs.

5.1 Automatic results

Of the 116 automatic runs submitted for the ad hoc task, 37 runs used the complete topic statement, 59 runs used only the title and description fields, and 20 used only the title field. A run that used only the title and description fields (so-called "short" runs) was required from every group that submitted any automatic run. Figure 6 shows the recall/precision curves for the eight TREC-8 groups with the best short runs as measured by mean average precision. The runs are ranked by average precision and only one run is shown per group. These graphs (and others in this section) are not intended to show specific comparison of results across sites but rather to provide a focal point for discussion of methodologies used in TREC. For more details on the various runs and procedures, please see the cited papers in this proceedings.

`pir9Attd` – Queens College, CUNY ("TREC-8 Ad-Hoc, Query and Filtering Track Experiments using PIRCS" by K.L. Kwok, L. Grunfeld, and M. Chan). The PIRCS system is a spreading activation method, which the authors show can be viewed as a combination of a probabilistic model and a simple language model. This run was produced using the same basic processing as the CUNY TREC-7 runs. The final result is produced by using a sequence of 5 different techniques that improve on the initial result. These methods include average within-document term frequency weights for query terms, a variable Zipf threshold for selecting indexing terms, collection enrichment (using documents from outside the collection in the first stage retrieval to improve the density of relevant documents), query expansion by adding highly associated terms based on a mutual information measure, and reweighting query terms based on the retrieved set.

`ok8amxc` – Okapi group ("Okapi/Keenbow at TREC-8" by S.E. Robertson and S. Walker). This run used the BM25 weighting scheme developed several years ago by this group and query expansion based on

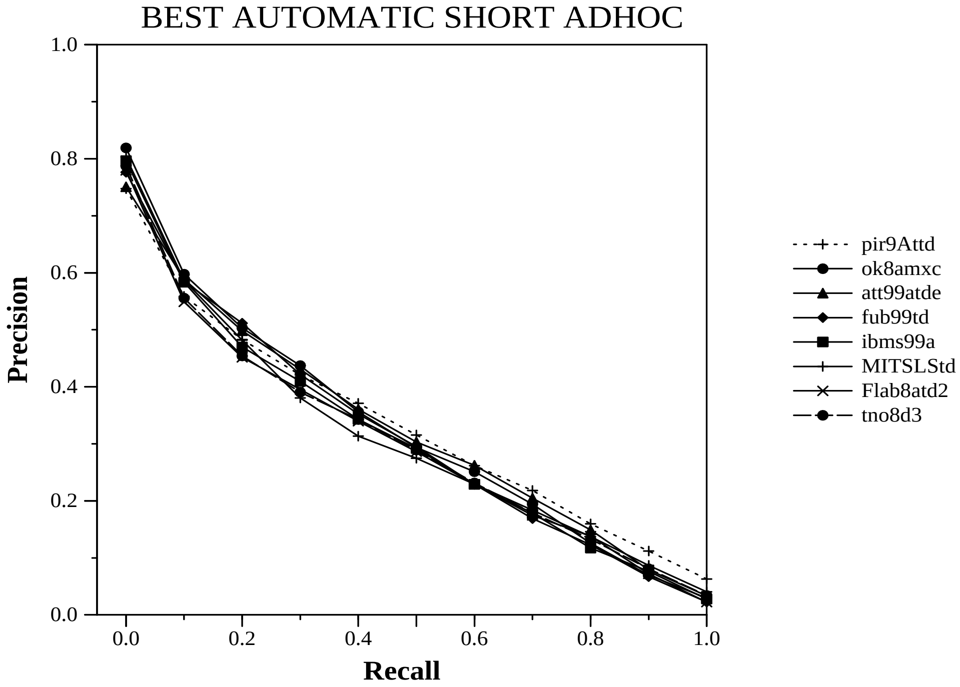


Figure 6: Recall/Precision graph for the top eight automatic short ad hoc runs.

blind feedback. The process used to select which terms should be included in the expanded query was new. Instead of simply taking the X candidate terms with the highest relevance weights, terms were added to the query if their relevance weight exceeded an absolute threshold.

att99atde – AT&T Labs–Research ("AT&T at TREC-8" by A. Singhal, S. Abney, M. Bacchiani, M. Collins, D. Hindle, and F. Pereira). AT&T also used essentially the same processing for TREC-8 as they used in TREC-7. This processing is based on a vector-space model with length-normalized, tf×idf weights, and query expansion using blind feedback and conservative collection enrichment. For this run, the blind feedback step was tweaked slightly. The initial retrieval pass produced a ranking of the top 50 documents, which was then re-ranked by promoting documents that contain multiple query terms. The top 10 documents from the new ranking were assumed to be relevant for feedback.

fub99td – Fondazione Ugo Bordoni ("TREC-8 Automatic Ad-Hoc Experiments at Fondazione Ugo Bordoni" by C. Carpineto and G. Romano). This run was produced using the Okapi formula for retrieving an initial set of documents and then expanding queries based on an information-theoretic term scoring function. The term scoring function uses the difference of the distribution of a term in the presumed-relevant set and the entire collection to compute the score. The same function was used in this group's TREC-7 work, but the ranking in the initial set of documents was much improved this year leading to much better overall retrieval.

ibms99a – IBM T.J. Watson Research Center ("Ad hoc, Cross-language and Spoken Document Information Retrieval at IBM" by M. Franz, J.S. McCarley, and R.T. Ward). This run was produced using the same processing as the IBM group's submissions to earlier TRECs. Okapi weights are used to provide an initial ranking of passages. Queries are then expanded using an LCA-like procedure based on top-ranking passages. The final document ranking is created using Okapi weights and the newly expanded query, with the score for a document computed as a function of the document's score and the score of its highest ranking passage.

MITSLStd – MIT Laboratory for Computer Science ("A Maximum Likelihood Ratio Information Retrieval Model" by K. Ng). The MIT group introduced a new probabilistic model based on the change in

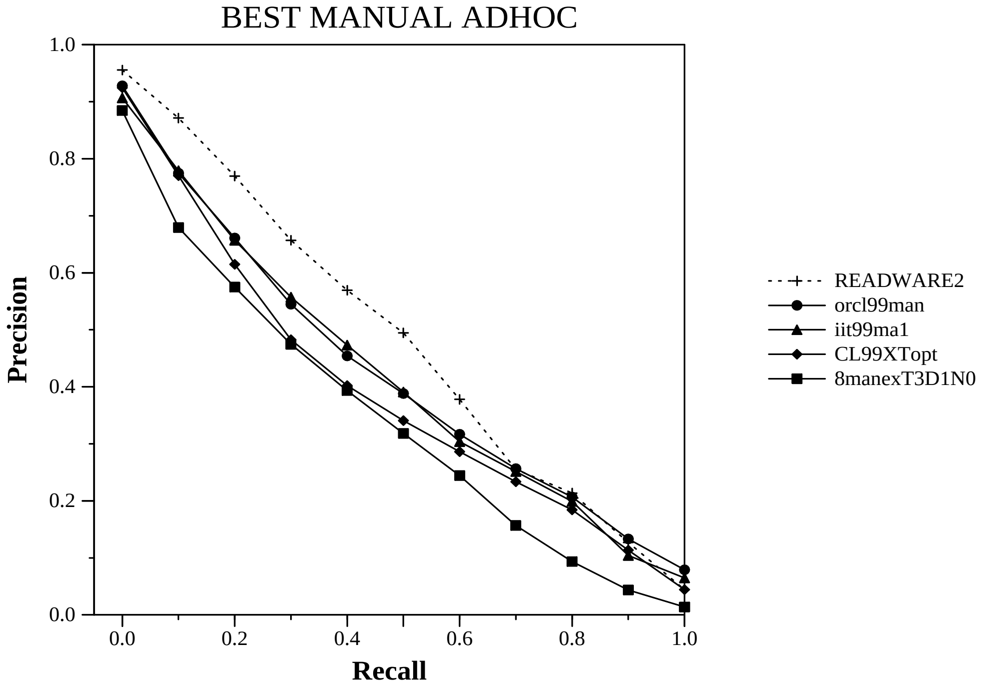


Figure 7: Recall/Precision graph for the top five manual ad hoc runs.

the likelihood of a document once the query is issued as compared to its a priori probability. The likelihoods are estimated using statistical language modeling techniques. This run was produced using the new model with an extension to incorporate blind feedback.

`Flab8atd2` – Fujitsu Laboratories ("Fujitsu Laboratories TREC8 Report—Ad hoc, Small Web, and Large Web Track" by I. Namba and N. Igata). Fujitsu Laboratories experimented with a number of techniques, many of which were either unstable (i.e., significantly improving some topics while significantly degrading an equal number of topics) or provided little benefit. This run used a modified Okapi weighting scheme, incorporated query expansion using blind feedback, increased the score of documents that contained multiple query terms, and increased the score of documents that contained good word pairs from the topic.

`tno8d3` – Twenty-One project ("Twenty-One at TREC-8: using Language Technology for Information Retrieval" by W. Kraaij, R. Pohlmann, and D. Hiemstra). The Twenty-One project uses a vector-space model where weights are based on statistical language models. Their TREC-8 processing was the same as their TREC-7 processing, which includes query expansion using blind feedback.

The immediate conclusion to be drawn from figure 6 is that there are many approaches that lead to essentially the same retrieval effectiveness. Yet while there are differences in the details of these approaches, they all share two properties that we can therefore conclude are fundamental to effective retrieval performance. Of primary importance is the use of a high-quality weighting scheme. Query expansion using terms from highly-ranked documents or documents related to highly-ranked documents is also beneficial.

5.2 Manual results

Figure 7 shows the recall/precision curves for the five TREC-8 groups with the highest mean average precision scores for manual runs. Once again, the runs are ranked by mean average precision and only one run per group is shown.

`READWARE2` – Management Information Technologies, Inc. ("High Selectivity and Accuracy with READWARE's Automated System of Knowledge Organization" by T. Adi, O.K. Ewell, and P. Adi). This run was produced by an analyst using READWARE's tools to define and refine a set of highly-specific queries for each topic. The number of queries used per topic ranged between 2 and 65 with a mean of 14. The submitted results were the union of the output of the different queries.

`orcl99man` – Oracle Corporation ("Oracle at Trec8: a Lexical Approach" by K. Mahesh, J. Kud, and P. Dixon). Oracle's interMedia Text retrieval system includes a large lexical knowledge base that provides a hierarchical classification of concepts with cross-reference links among concepts. The system provides an "ABOUT" operator that allows queries that specify a high-level concept to match documents that contain a particular expression of that concept. This run was produced by having a searcher interact with the interMedia system (using manual feedback, browsing the lexical knowledge base, etc.) to define a single (assumed-to-be-best) query for each topic. The submitted results were the ranked output produced by the final query.

`iit99ma1` – Illinois Institute of Technology group ("IIT at TREC-8: Improving Baseline Precision" by M.C. McCabe, D.O. Holmes, K.L. Alford, A. Chowdhury, D.A. Grossman, and O. Frieder). This run is the latest in a series of manual runs the IIT group has submitted to the past several TRECs. For this run, the searcher spent approximately a half hour formulating a query using manual relevance feedback and general knowledge to select query words. The basic retrieval strategy was a vector-space model, though there was also a mechanism to remove a document from the retrieved set if it contained concepts that were included on a (topic-specific) negated concepts list.

`CL99XTopt` – CLARITECH Corporation ("CLARIT TREC-8 Manual Ad-Hoc Experiments" by D.A. Evans, J. Bennett, X. Tong, A. Huettner, C. Zhai, and E. Stoica). In previous years, the CLARITECH group showed that clustering the result set of a search helped users find relevant documents to use in a subsequent round of relevance feedback. This year, the system enabled the clustering to be computed over an extended set of index terms. A second change allowed the number of query terms added to the query during relevance feedback to be query-dependent rather than an arbitrary, fixed number. Searchers were allowed a maximum of 20 minutes wall clock time per topic. In addition to making relevance judgments, the searchers could modify the automatically-constructed query if they chose to do so. This run was the result of both the new clustering and the query-dependent number of expansion terms.

`8manexT3D1N0` – GE Research and Development group ("Natural Language Information Retrieval: TREC-8 Report" by T. Strzalkowski, J. Perez-Carballo, J. Karlgren, A. Hulth, P. Tapanainen, and T. Lahtinen). This run was the result of an investigation into how effective natural language indexing techniques are when query statements are large. The original TREC topic statement was fed to a standard retrieval system and topic-related summaries of the top 30 documents were returned. The user reviewed each summary and removed any summary that was not relevant. This was the only manual processing in the run, and users were limited to no more than 10 minutes wall-clock time to perform the review. All summaries not explicitly removed by the user were attached to the original topic statement, and the resulting new statement was submitted to the groups' NLP-based retrieval system.

Comparisons of manual runs in the ad hoc task are especially tricky because a wide range of levels of human effort are lumped together into a single category. Comparing the graphs in figures 6 and 7, at a minimum we can conclude that users who will participate in their searches can be rewarded with better search results.

5.3 Future of the ad hoc main task

The ad hoc task was one of the two tasks that were performed in TREC-1 and has been run in every TREC since then. There are several reasons for the task's primary position in TREC.

- Historically, ad hoc retrieval has been regarded as the fundamental task in text retrieval.
- Having one task that all (or most) groups perform documents the state of the art and provides a basis of comparison for each groups' results in other tracks.

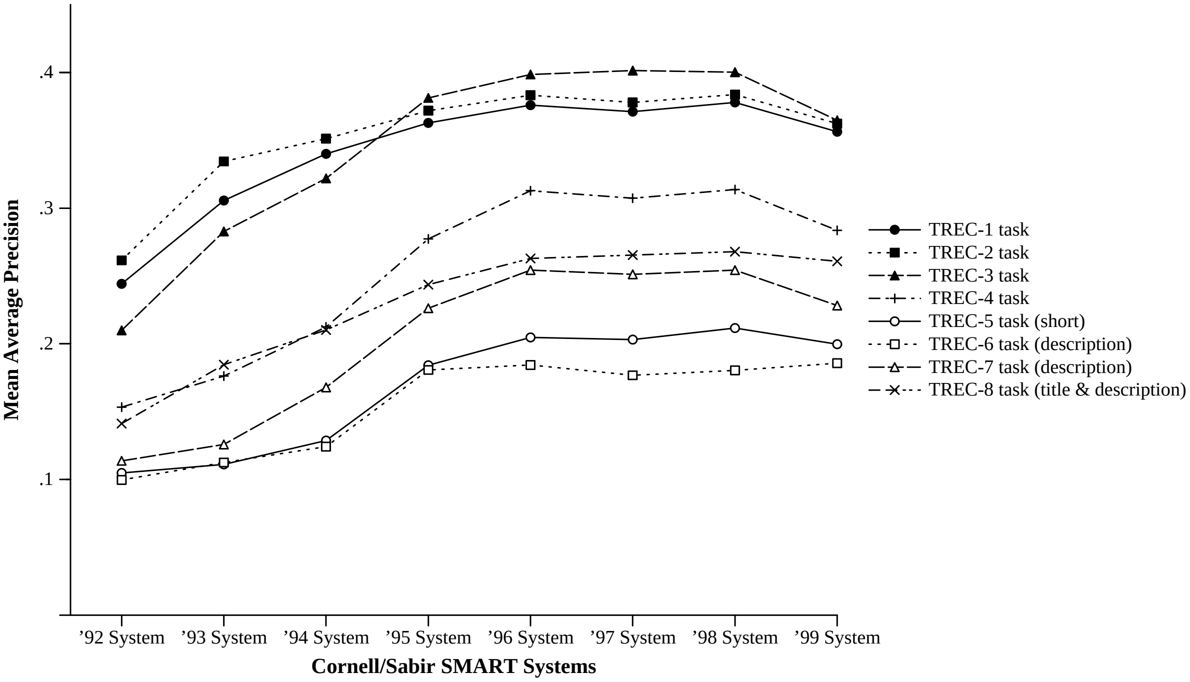


Figure 8: Mean average precision obtained by different versions of the Cornell/Sabir SMART system for the different ad hoc tasks in TREC.

- The main task is the means by which the TREC general-purpose IR test collections are built.
- The task provides a convenient starting point for new participants.

However, a number of participants also believe that the community is no longer learning enough from the results of the task to justify the (participants' and NIST) resources spent on it. Many participants' TREC-8 runs were produced using essentially the same system as previous years' runs simply to satisfy the requirement for an ad hoc run.

Figure 8 shows how the results in the ad hoc task have plateaued in recent years. The figure gives a plot of the mean average precision scores obtained by the Cornell/Sabir SMART system for each version of the system for each ad hoc task (i.e., topic and document sets) as reported by the SMART group [1]. Each line connects the scores for the same task across the system variants. The '92 System was the system that was used to produce the runs submitted to TREC-1, the '93 System was used to produce the runs submitted to TREC-2, etc. Once a new set of topics was released, the SMART group ran that set of topics on each prior version of the system. The SMART system has consistently been among the better systems in TREC (except the '99 System in which a technique designed to enhance early precision turned out to harm overall performance), so the trend is indicative of the field as a whole. Retrieval effectiveness in the ad hoc task has improved dramatically since the beginning of TREC but has now leveled off.

Because of these considerations and the fact that we now have 8 years worth of test collections, the ad hoc main task will be discontinued in future TRECs. This is not to say that we believe that the ad hoc text retrieval problem is solved. Indeed, figure 8 shows absolute performance on the task is less than ideal. Rather it is an acknowledgement that sufficient infrastructure exists so that researchers can pursue their investigations independently, and thereby free TREC resources for other tasks.

6 The Tracks

One of the goals of TREC is to provide a common task evaluation that allows cross-system comparisons, which has proven to be a key strength in TREC. A second major strength is the loose definition of the ad hoc

Table 5: Number of task participants.

	TREC-6	TREC-7	TREC-8
Ad Hoc	31	42	41
CLIR	13	9	12
Filtering	10	12	14
GIRT	—	0	2
Interactive	9	8	7
QA	—	—	20
Query	—	2	5
SDR	13	10	10
Small Web	—	—	17
Large Web	—	—	8

task, which allows a wide range of experiments. The addition of secondary tasks (called tracks) in TREC-4 combined these strengths by creating a common evaluation for retrieval subproblems.

The tracks have had a significant impact on TREC participation. Table 5 gives the number of participants in each of the TREC-8 tasks for TREC 6, 7, and 8. The total number of participating groups continues to grow each year, with 66 groups this year compared to 56 in TREC-7 and 51 in TREC-6.

Each track has a set of guidelines developed under the direction of the track coordinator. Participants are free to choose which, if any, of the tracks they will join. This section describes the tasks performed in TREC-8 tracks. See the track reports elsewhere in this proceedings for a more complete description of each track.

6.1 The Cross-Language (CLIR) track

The CLIR task focuses on retrieving documents that are written in different languages using topics that are in one language. The TREC-8 track used the same document set that was used in TREC-7: a set of French documents from the Swiss news agency *Schweizerische Depeschen Agentur* (SDA); a set of German documents from SDA plus a set of articles from the newspaper *New Zurich Newspaper* (NZZ); a set of Italian documents from SDA; and a set of English documents from the AP newswire. All of the document sets contain news stories from approximately the same time period, but are not aligned or specially coordinated with one another. Participants were provided with a new set of 28 topics, numbered 54 through 81, that had translations available in English, French, German, and Italian. Participants used one topic language to search the combined document set.

As in TREC-7, the construction of the cross-language test collection differed from the way other TREC collections are created. Candidate topics in the native language were created in each of four different institutions: NIST, USA (English); University of Zurich, Switzerland (French); Social Science Information Centre, Bonn/University of Koblenz, Germany (German); and CNR, Pisa, Italy (Italian). Each institution developed candidate topics such that a third of the candidates targeted international events, a third targeted items of interest in Europe generally, and a third targeted local items of interest. The intention was to create topics that had different distributions of relevant documents across languages. Each of the institutions contributed seven topics to the final set of 28, with representatives from each site meeting to ensure the actual question being asked in the topic was understood by all. The final topics were then translated into the three remaining languages so that the entire set of topics was available in each language. The relevance judgments for all topics for a particular document language were made at the site responsible for that language. The TREC-7 and TREC-8 cross-language collections are the only TREC collections in which multiple relevance assessors provided judgments for a single topic.

Forty-five runs from 12 different groups were submitted to the track. Eight runs were submitted for the special GIRT task (described below) and nine runs used a subset of the document collection (for example, Italian topics run against the English portion of the collection). Only one run was a manual run.

Evaluation of cross-language retrieval poses some challenges. As mentioned above, this is the only task in which multiple assessors judge the same topic. Pool creation is also affected: pools must be adequately balanced across languages to assure sufficient coverage in each language. For TREC-8, 18 runs were added to the pools. These runs included each group's first choice of runs to be judged plus runs that retrieved relatively many German or Italian documents. In addition, a monolingual Italian run was solicited by NIST to bolster the Italian pools.

We can perform the same analysis as was performed for the ad hoc collection in section 3.3.2 once the pools are judged. For the cross-language runs, the mean percentage difference in mean average precision scores computed with and without a group's unique relevant documents was 6.3 %, with a maximum percentage difference of 15.4 %. This compares with the ad hoc maximum percentage difference of less than 1 % for automatic runs. The larger difference for the cross-language test collection can be attributed to the combination of fewer runs contributing to the pools and the lack of high-recall manual runs. While the difference is clearly larger for the cross-language test collection than the ad hoc collection, a mean difference of approximately 6 % is okay for most purposes. Experimenters who find many unjudged documents in the top-ranked list of only one of a pair of runs to be contrasted may need to proceed with care.

The TREC-8 track also had an optional subtask known as the **GIRT task**. The subtask used the GIRT collection, a 31,000 document structured database (formatted as SGML fielded text data) from the field of social science, plus the NZZ articles, and a separate set of 28 topics. The rationale of the subtask was to study CLIR in a vertical domain (i.e., social science) where a German/English thesaurus is available.

6.2 The Filtering track

The tasks within the TREC-8 filtering track were the same as the TREC-7 track, though the document and topic sets differed and as did the utility functions used to evaluate the runs. The TREC-8 track used topics 351–400 and the *Financial Times* document set from Disk 4.

The filtering problem can be viewed as the inverse of the ad hoc retrieval task in that the question is assumed to be known and the document stream changes. The filtering task is to retrieve just those documents in the stream that match the user's interest as represented by the query. The main focus of the track was an *adaptive* filtering task. In this task, a filtering system starts with just a query derived from the topic statement, and processes documents one at a time in date order. If the system decides to retrieve a document, it obtains the relevance judgment for it, and can modify the query based on the judgment if desired.

For continuity with previous TRECs, two other, simpler tasks were also part of the TREC-8 track. In the *batch* filtering task, the system is given a topic and a set of known relevant documents. The system creates a query from the topic and known relevant documents, and must then decide whether or not to retrieve each document in the test portion of the collection. In the *routing* task, the system again builds a query from a topic statement and a set of relevant documents, but then uses the query to rank the test portion of the collection. Ranking the collection by similarity to the query (routing) is an easier problem than making a binary decision as to whether a document should be retrieved (batch filtering) because the latter requires a threshold that is difficult to set appropriately.

Fifty-five runs from 14 different groups were submitted to the filtering track. Thirty-three of the runs were adaptive filtering runs, 11 runs were batch filtering runs, and 11 runs were routing runs. The results of the adaptive filtering subtask demonstrate the difficulty of the problem. When evaluated using the rule that retrieving a relevant document earns a system three "points" while retrieving a nonrelevant document subtracts two "points" (the LF1 utility function below), the average behavior of each system was worse than the baseline of retrieving no documents at all. With a somewhat easier scoring metric (three points for a relevant retrieved but only 1 point subtracted for a nonrelevant retrieved) some systems performed better than the baseline on average, but very small retrieved sets were still best.

Developing appropriate measures for filtering systems continues to be an important part of the track. The main approach used in TREC is to use utility functions as measures of the quality of the retrieved set—the quality is computed as a function of the benefit of retrieving a relevant document and the cost of retrieving an irrelevant document [7]. In TREC-8 two different linear utility functions were used:

$$\begin{aligned} \mathrm{LF1} &= 3\mathrm{R}^{+} - 2\mathrm{N}^{+} \\ \mathrm{LF2} &= 3\mathrm{R}^{+} - \mathrm{N}^{+} \end{aligned}$$

where R^+ and N^+ are the number of relevant and non-relevant documents retrieved, respectively. A pair of non-linear utility functions were also defined for the TREC-8 task, but few runs that were optimized for those metrics were submitted. Many participants in the track felt that the non-linear measures did not model a user's behavior very well.

6.3 The Interactive track

The interactive track was one of the first tracks to be introduced into TREC. Since its inception, the high-level goal of the track has been the investigation of searching as an interactive task by examining the process as well as the outcome. One of the main problems with studying interactive behavior of retrieval systems is that both searchers and topics generally have a much larger effect on search results than does the retrieval system used.

The TREC-8 task was very similar to the TREC-7 task. The track used slightly modified versions of six ad hoc topics. Each of the six topics described an information need such that the document collection (the *Financial Times* collection from Disk 4) contained multiple distinct examples or instances of the requested information. The searchers' job was to save documents covering as many distinct answers to the question as possible in a 20-minute time limit. The NIST assessor for the topic made a comprehensive list of instances from the documents submitted by the track. The effectiveness of the search was evaluated by the fraction of total instances for that topic covered by the search (instance recall) and the fraction of the documents retrieved in the search that contained an instance (instance precision). Participants were also required to collect demographic and user satisfaction data from the searchers, and to report extensive data on each searcher's interactions with the search systems.

The track did not attempt to coordinate cross-site comparisons or test a particular hypothesis across sites. It did impose an experimental matrix that defined how searchers and topics were to be divided among whatever experimental and control systems the participants were testing. The matrix was based on a latin square design to provide an uncontaminated estimate of the difference between the systems. The minimum experiment defined by the design required 12 searchers so that query order would not be confounded with other effects. Each searcher performed three searches with each of the two systems, and each query was searched in each position (first through sixth) by each system.

Seven groups submitted interactive results. Two groups used the minimum experimental design of 12 searchers, four groups used 24 searchers, and one group used 36 searchers. Each group found little difference between their control and experimental systems. This could mean that none of the various devices implemented in the experimental systems are helpful in the instance retrieval task, or that the statistical power of the experimental design is not sufficient to detect the difference. Further study of the design of effective user studies is needed.

6.4 The Question Answering track

TREC-8 was the first time the question answering track was run. The purpose of the track was to encourage research into systems that return actual answers, as opposed to ranked lists of documents, in response to a question.

The track used the ad hoc document collection (i.e., the documents on Disks 4 and 5 minus the *Congressional Record* subcollection) and 198 fact-based, short-answer questions such as "How many calories are there in a Big Mac?" Each question was guaranteed to have at least one document in the collection that answered the question. Participants were to return a ranked list of five strings per question such that each string was believed to contain an answer to the question. Depending on the run type, answer strings were limited to either 50 or 250 bytes. Human assessors read each string and made a binary decision as to whether or not the string actually did contain an answer to the question. Individual questions received a score equal to the reciprocal of the rank at which the first correct response was returned (or 0 if none of the five responses contained a correct answer). The score for a run was the mean of the individual questions' reciprocal ranks.

Twenty groups submitted 45 runs to the track. Unsurprisingly, for every group that submitted both 50- and 250-byte runs, the 250-byte runs was better, demonstrating that the 250-byte task is easier. The submissions from AT&T Research Labs [8] suggest that existing passage-retrieval techniques can be successful for 250-byte runs, but are not suitable for 50-byte runs. For 50-byte runs, some explicit natural language processing (for example, entity-finding) appears necessary.

6.5 The Query track

The variability in topic performance makes it impossible to reach meaningful conclusions regarding query-dependent processing strategies unless there is a very large query set—much larger than the sets of 50 topics used in the TREC collections. The query track was designed as a means for creating a large set of different queries for an existing TREC topic set.

Participants in the track created different query versions for topics 51–100, possibly using relevance judgments from Disk 2. A query of a given type was created for each of the 50 topics, forming one query set. Four different query types were used:

Very short: two or three words extracted from the topic statement.

Sentence: an English sentence based on the topic statement and the relevant documents.

Manual feedback: an English sentence based on reading 5–10 relevant documents only (by someone who doesn't know the topic statement).

Weighted terms: a list of terms with weights (for example, as produced by an automatic feedback process).

Participants exchanged the query sets they created with all other participants in the track, and all participants ran all query sets. The document set used for the runs was the documents on Disk 1.

Since the track design included all groups running all query sets, a number of direct comparisons are possible. First, participants can see how effective their system is using their own queries. Second, they can see how effective their search component is when using other queries. Finally, participants can evaluate how effective their query construction strategies are by seeing how other groups fared with their queries.

Five groups participated in the track. There were a total of 23 different query sets produced and 9 different retrieval strategies used in the track. One of the main results of the track was confirmation of the wide variability in the effectiveness of different systems both within and across topics. One view of the data is shown in the graph in figure 9. To create the graph, the mean of the average precision scores for a system was computed over the 23 query sets for a given topic. The mean of those averages was then computed over the 9 different systems and plotted as the circle for the topic in the graph. The endpoints of the error bar for a topic represent the scores for the systems with the worst and best average scores. The mean of the average scores is a measure of the intrinsic difficulty of a topic, while the spread of an error bar represents how similarly the different systems performed.

6.6 The Spoken Document Retrieval (SDR) track

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

The TREC-8 track had the same general task as was used in TREC-7. A major difference between the tracks was the size of the collection used. Whereas the TREC-7 collection consisted of 87 hours of broadcast news programs, representing approximately 2900 news stories, plus 23 topics, the TREC-8 collection consisted of more than 550 hours of news broadcasts (21,500 stories) and 50 topics.

Participants worked with different versions of transcripts of the news broadcasts to judge the effects of errors in the transcripts on retrieval performance. The *reference* transcripts were based on closed captioning of the broadcasts; these transcripts were assumed to be perfect, though this assumption is less true than with the reference transcripts used in previous years. The *baseline* transcripts were produced at NIST by using NIST's installation of the BBN Rough 'N Ready BYBLOS speech recognizer. The *recognizer* transcripts were produced by the participants' own recognizer systems. The recognizer transcripts of the different

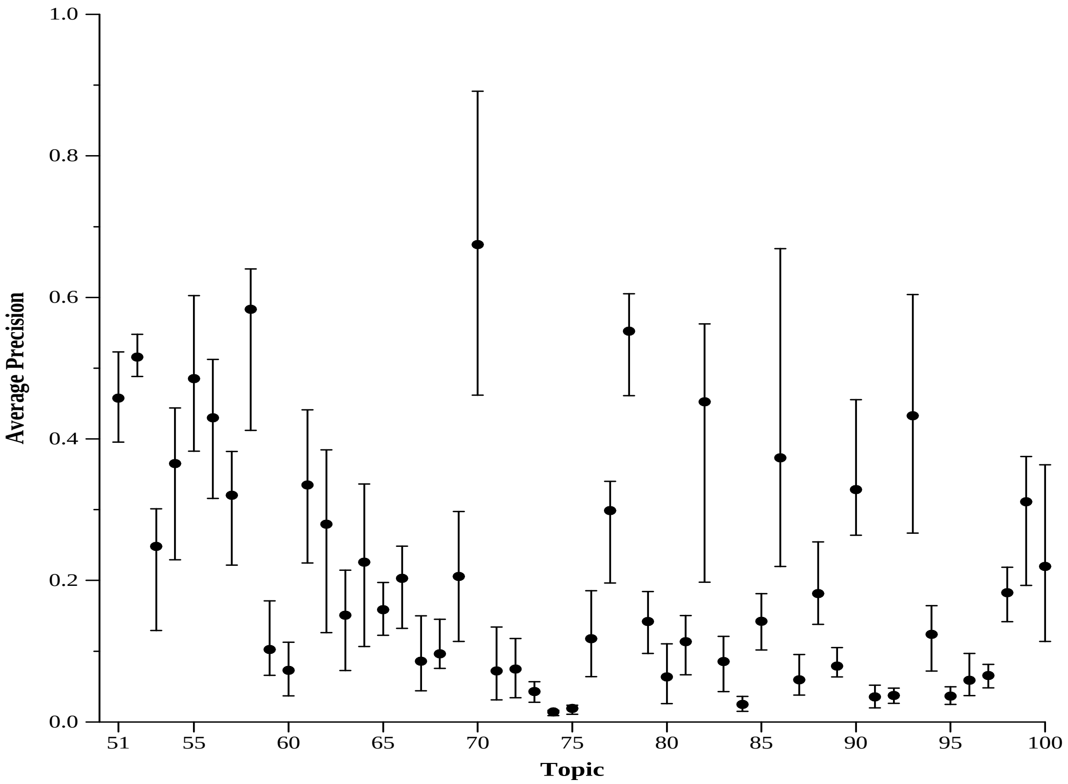


Figure 9: Average precision scores for different systems averaged over the 23 query sets per topic. The circle is the mean of the average scores computed over the 9 system variants. The error bars represent the worst and best average score for individual systems.

participants were made available to one another so that participants could perform retrieval runs against their own recognizer transcripts as well as others' recognizer transcripts (*cross-recognizer* runs). The different versions of the transcripts allowed participants to observe the effect of recognizer errors on their retrieval strategy. The different recognizer runs provide a comparison of how different recognition strategies affect retrieval.

Another difference between the TREC-7 and TREC-8 tracks was the introduction of an unknown boundary condition into the TREC-8 track. As in previous years, document boundaries were given in the reference transcripts, and these same boundaries could be used in the other versions of the transcripts as well (the known boundary case). In the unknown boundary condition, the information regarding the beginning and ending of stories was not used. Since the story boundary information is what excluded the non-news portion of the broadcasts (commercials, musical interludes, etc.) from the test collection, the unknown boundary condition entailed a much more difficult recognition task. For the unknown condition, the systems returned a ranked list of time offsets rather than a ranked list of story identifiers. During scoring, the times were mapped back to the story boundaries. Times that mapped to non-stories were assigned an invalid story identifier that was always irrelevant. Similarly, all times that mapped to a story that had already been retrieved for a topic were also assigned an invalid story identifier.

The final difference between the tracks was the provision for using a rolling language model in the participants' recognizer systems. One of the main causes of recognition errors is out-of-vocabulary words, and news stories are particularly vulnerable to this problem. To counteract this effect, participants were allowed to use a language model that (automatically) adapts to newswire texts from previous days if they desired.

Ten groups participated in the track, with six groups performing the full SDR task (recognition and retrieval) and the remaining four groups performing retrieval against the transcripts made available from NIST. In general, both speech recognition performance and retrieval performance was quite good. Retrieval performance degraded very little for transcripts with increasing word error rates, probably due to the redundancy of key words in the spoken documents. Comparisons between the same systems run when story boundaries are known and not known show that the unknown boundary condition is more difficult, though part of the difference is that the unknown boundary runs had to process commercials and other filler material that was excluded in the known boundary case. Similar comparisons show that adaptive recognition systems can be used to more effectively recognize speech data collected over time than comparable static systems.

6.7 The Web track

Like the question answering track, the web track was a new track for TREC-8. The purpose of the track was to provide the infrastructure required to reliably evaluate new search techniques and to perform repeatable experiments in the context of the World Wide Web. The track used a frozen snapshot of the web as its document collection. This collection, known as the VLC2 collection and used in last year's Very Large Collection track [5], is over 100 gigabytes and represents some 18.5 million web pages.

The track defined two subtasks, the small web and the large web tasks, based on the amount of the web data used. The small web task used a 2 gigabyte, 250,000 document subset of the VLC2 collection, while the large web task used the entire collection.

The focus of the small web task was on answering two questions:

- Do the best methods used in the TREC ad hoc task also work best on web data? and
- Can link information in web data be used to obtain more effective search rankings than can be obtained using page content alone?

The task was exactly the same as the TREC-8 ad hoc task except that the web documents were searched instead of the documents on Disks 4 and 5. The NIST relevance assessors who judged the ad hoc pools also judged the corresponding small web pools.

Seventeen groups submitted 44 runs to the small web track. Incorporating link information did not improve retrieval performance, though that may be the result of the impoverished collection of links available in a 2 gigabyte sample of the web.

Once again we can use the test described in section 3.3.2 to gauge how complete the relevance judgments are for this collection. No manual runs contributed to the pools because the track guidelines prohibited manual runs. The total number of unique relevant documents over the 50 topics found per group ranged from a high of 89 to a minimum of 5. The mean absolute difference in mean average precision scores with and without a group's unique relevant documents computed over the 27 runs that were judged was .0021, with a maximum difference of .0073. For the 21 runs whose mean average precision score was at least .1, the mean percentage difference was 1.05 % with a maximum percentage difference of 2.85 %. These differences are quite small, and suggest that the pools were adequate to build a reliable test collection.

The large web task was also a traditional ad hoc retrieval task. In this case, however, the full VLC2 collection of documents was searched using 10,000 queries extracted from logs from the Alta Vista and Electric Monk search engines. Participants submitted the top 20 documents for all 10,000 queries to the Cooperative Research Centre for Advanced Computational Systems (ACSys). ACSys selected 50 of the 10,000 queries to judge, and judged all 20 documents for each run for those 50 queries.

The large web task is a direct descendent of the Very Large Collection track of previous years, and some of the eight groups who submitted runs to the task addressed effectiveness versus efficiency issues. Other groups used the collection to investigate distributed IR algorithms or to examine whether retrieving web documents is intrinsically different from retrieving other documents such as the newspaper articles that constitute most of the TREC collections.

7 The Future

The final session of each TREC conference is a planning session for future years. As described above, the ad hoc main task will be discontinued in TREC-9. Instead, we will focus resources on building a test collection of web documents. A 10 gigabyte sample of the VLC2 corpus will be used as the document set, and 50 topic statements will be created using real web queries taken from an Excite log as topic seeds. Pooling and relevance judging will be done as for the ad hoc task in previous years.

All of the other TREC-8 tracks will continue in TREC-9. The question answering and query tracks will perform essentially the same task as in TREC-8 so that we can gain more experience with those tasks. The task in the SDR track will also be similar, though there will be more of a focus on retrieval when document boundaries are not known. The remaining tracks will change more significantly. The cross-language track will focus on retrieving Chinese documents using English topics; research on cross-language retrieval for European languages will continue in the new CLEF initiative (see `http://www.iei.pi.cnr.it/DELOS/CLEF`). The filtering track will again have adaptive filtering, batch filtering, and routing tasks, but will use medical documents to explore how the problems change in a domain-specific environment. The task in the interactive track will change from an instance retrieval task to a question-answering task.

TREC is expected to continue beyond TREC-9. The set of tracks included in a particular year will continue to vary depending on the interests of the participants and sponsors, and the suitability of the problem to the TREC environment. New tracks will be introduced as the need arises. The call for participation for a particular TREC lists the set of tracks that it will include. The call is issued in December, and is posted on the main page of the TREC web site (`http://trec.nist.gov`) while the call is active.

Acknowledgments

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

References

[1] Chris Buckley and Janet Walz. SMART in TREC 8. In Voorhees and Harman [13].

[2] Gordon V. Cormack, Christopher R. Palmer, and Charles L.A. Clarke. Efficient construction of large test collections. In Croft et al. [3], pages 282–289.

[3] W. Bruce Croft, Alistair Moffat, C.J. van Rijsbergen, Ross Wilkinson, and Justin Zobel, editors. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, August 1998. ACM Press, New York.

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

[5] David Hawking, Nick Craswell, and Paul Thistlewaite. Overview of the TREC-7 very large collection track. In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, pages 91–103, August 1999. NIST Special Publication 500-242. Electronic version available at http://trec.nist.gov/pubs.html.

[6] Noriko Kando, Kazuko Kuriyama, Toshihiko Nozue, Koji Eguchi, Hiroyuki Kato, and Souichiro Hidaka. Overview of IR tasks at the first NTCIR workshop. In *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pages 11–44, 1999.

[7] David D. Lewis. The TREC-4 filtering track. In D. K. Harman, editor, *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*, pages 165–180, October 1996. NIST Special Publication 500-236.

[8] Amit Singhal, Steve Abney, Michiel Bacciani, Michael Collins, Donald Hindle, and Fernando Pereira. AT&T at TREC-8. In Voorhees and Harman [13].

[9] 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.

[10] Karen Sparck Jones. Further reflections on TREC. *Information Processing and Management*, 36(1):37–85, 2000.

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

[12] Ellen M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, 36:697–716, 2000.

[13] E.M. Voorhees and D.K. Harman, editors. *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*. Electronic version available at http://trec.nist.gov/pubs.html, 2000.

[14] Justin Zobel. How reliable are the results of large-scale information retrieval experiments? In Croft et al. [3], pages 307–314.