Overview of the Seventh Text REtrieval Conference (TREC-7)

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

1 Introduction

The seventh Text REtrieval Conference (TREC-7) was held at the National Institute of Standards and Technology (NIST) on November 9–11, 1998. 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-7 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 [7], Tague-Sutcliffe and Blustein [9], and Harman [2]. In addition, the overview paper in each of the previous TREC proceedings summarizes the results of that TREC.

The TREC workshop series has the following goals:

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

Table 1 lists the groups that participated in TREC-7. Fifty-six groups including participants from 13 different countries and 19 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. It concentrates mainly on the main task, *ad hoc retrieval*, which is defined 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-7 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-7. 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 Task

The ad hoc 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

Table 1: Organizations participating in TREC-7

ACSys Cooperative Research Centre	Management Information Technologies, Inc.
AT&T Labs Research	Massachusetts Institute of Technology
Avignon CS Laboratory/Bertin	National Tsing Hua University
BBN Technologies	NEC Corp. and Tokyo Institute of Technology
Canadian Imperial Bank of Commerce	New Mexico State University
Carnegie Mellon University	NTT DATA Corporation
Commissariat à l'Energie Atomique	Okapi Group (City U./U. of Sheffield/Microsoft)
CLARITECH Corporation	Oregon Health Sciences University
Cornell University/SabIR Research, Inc.	Queens College, CUNY
Defense Evaluation and Research Agency	RMIT/Univ. of Melbourne/CSIRO
Eurospider	Rutgers University (2 groups)
Fondazione Ugo Bordoni	Seoul National University
FS Consulting, Inc.	Swiss Federal Institute of Technology (ETH)
Fujitsu Laboratories, Ltd.	TextWise, Inc.
GE/Rutgers/SICS/Helsinki	TNO-TPD TU-Delft
Harris Information Systems Division	TwentyOne
IBM — Almaden Research Center	Universite de Montreal
IBM T.J. Watson Research Center (2 groups)	University of California, Berkeley
Illinois Institute of Technology	University of Cambridge
Imperial College of Science, Technology and Medicine	University of Iowa
Institut de Recherche en Informatique de Toulouse	University of Maryland
The Johns Hopkins University — APL	University of Massachusetts, Amherst
Kasetsart University	University of North Carolina, Chapel Hill
KDD R&D Laboratories	Univ. of Sheffield/Cambridge/SoftSound
Keio University	University of Toronto
Lexis-Nexis	University of Waterloo
Los Alamos National Laboratory	U.S. Department of Defense

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.

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 they 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. Topics 351–400 were created for the TREC-7 ad hoc task. The set of documents used in the task was those contained on TREC Disks 4 and 5, excluding the *Congressional Record* subcollection. See Section 3.1 for details about this document set.

Participants were allowed to submit up to three 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. When submitting a run, participants were required to state whether the queries were produced manually or automatically. If an automatic method was used, participants also stated what parts of the topic statement

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

were used (see Section 3.2).

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-7, 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.

3.2 Topics

The format of the TREC topics has evolved over time as illustrated in Table 3. The table shows the number

```
<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*.

```
<num> Number:  396
<title> sick building syndrome

<desc> Description:
Identify documents that discuss sick building syndrome or building-related
illnesses.

<narr> Narrative:
A relevant document would contain any data that refers to the sick building
or building-related illnesses, including illnesses caused by asbestos, air
conditioning, pollution controls.  Work-related illnesses not caused by the
building, such as carpal tunnel syndrome, are not relevant.
```

Figure 2: A sample TREC-7 topic.

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 three sets of ad hoc topics (251–400) have therefore all had the same format as in TREC-3, consisting of a title, description, and narrative. A sample TREC-7 topic is shown in Figure 2.

The different parts in the most recent TREC topics allow participants to investigate the effect of different

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

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

query lengths on retrieval performance. The "titles" in topics 301–400 were 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. For TREC-7 (topics 351–400), the description field contains all of the words in the title field, to remove the confounding effects of word choice on length experiments as was exhibited in TREC-6 [11]. The narrative gives a concise description of what makes a document relevant.

Ad hoc participants who used automatic query construction techniques were required to use particular parts of the topics in TREC-5 and TREC-6. The TREC-7 task description had no such requirements, but participants did have to report what parts they used when they submitted their runs.

Ad hoc topics have been constructed by the same person who performed the relevance assessments for that topic 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. NIST personnel select the final 50 topics from among the candidates based on having a range of estimated number of relevant documents and balancing the load across assessors.

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

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 [8] to assemble the relevance assessments. In this method a pool of possible relevant documents is created by taking a sample of documents selected by the various participating systems. This pool is then shown to the human assessors. The particular sampling method used in TREC is to take the top 100 documents retrieved in each submitted run for a given topic and merge them into the pool for assessment. This is a valid sampling technique since all the systems used ranked retrieval methods, with those documents most likely to be relevant returned first.

3.3.1 Overlap

Table 4 summarizes the amount of overlap in the ad hoc pool for each of the seven TRECs. 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. However, high precision track runs contribute fewer 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 possible pool size. The final column gives the average number of relevant documents in the pool and the percentage of the actual pool that was relevant. Starting in TREC-4, various tracks also contributed documents to the ad hoc pool. 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.

TREC-6 is clearly an outlier in Table 4. The tremendous drop in the size of the ad hoc pool reflects the difference in the number of runs NIST was able to assess that year. The overlap for the TREC-6 runs was less than in previous years, and this coupled with less time for assessing meant that NIST could only judge one ad hoc run per group. The overlap in TREC-7 was as high as in earlier years, so NIST was able to judge two ad hoc runs per group.

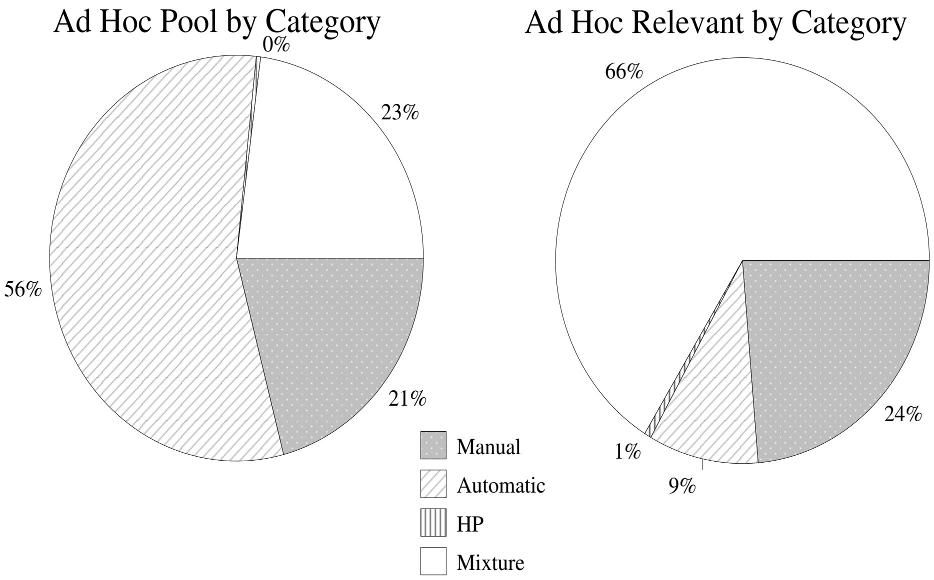

Figure 3: Distribution of categories in judged and relevant document pools.

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

3.3.2 Uniquely retrieved documents

The average overlap figures given in Table 4 hide details about the source of the documents in the pool. Figures 3 and 4 show two breakdowns of the sources of the relevant documents.

Figure 3 shows the percentages contributed by each type of ad hoc run (and high precision track) to the judgment pool, and to the relevant documents. For example, whereas 56% of the pool for relevance judgments came from the automatic systems, 9% of the relevant documents were found by only automatic systems. The manual systems contributed 21% of the pool, and 24% of the relevant documents were found only by manual systems.

Figure 4 gives a different view of the same issue by looking at the systems that retrieved the most unique relevant documents (i.e., relevant documents that were contributed to the pool by exactly one group). Almost all of the unique documents were retrieved by manual runs. Note that the pattern of the sources of unique relevant documents is very similar to the pattern found for TREC-5 [10].

4 Evaluation

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

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

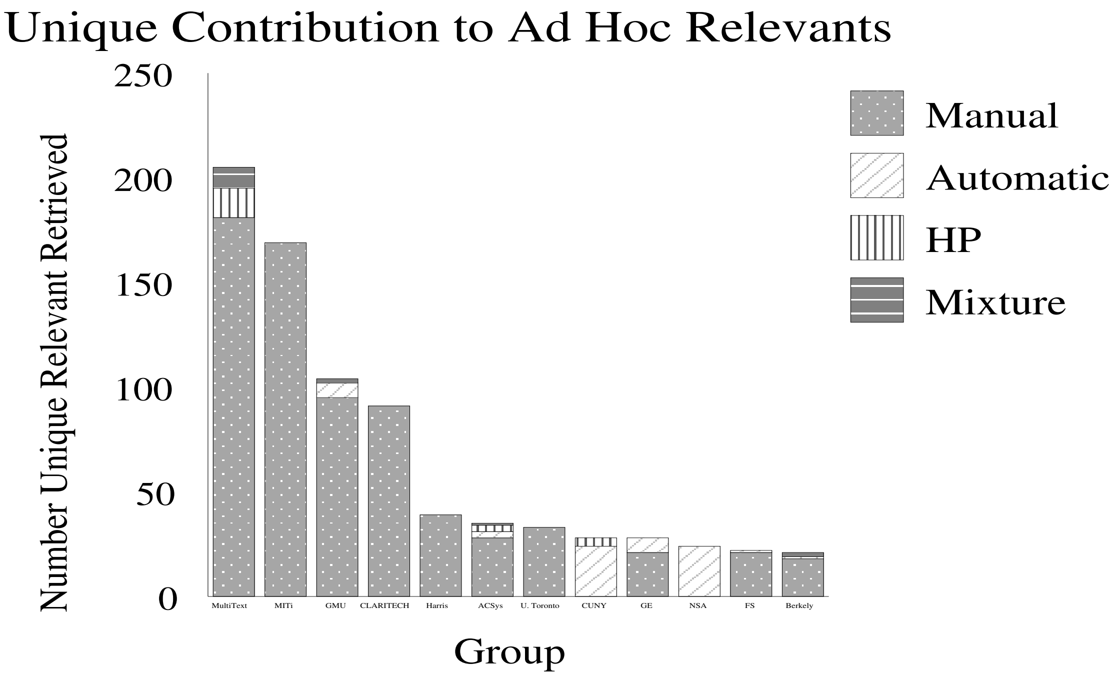

Figure 4: Percentage of unique relevant documents by category for groups retrieving more than 20 unique relevant documents.

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 as shown, for example, in Figure 5. 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.

Table 5: Kendall's tau correlations between pairs of system rankings.

	P(30)	R-Prec	Mean Ave Precision	.5 Prec	Recall R(1000)	Total Rel Ret	Rank of 1st Rel
P(10)	.8851	.8151	.7899	.7855	.7817	.7718	.6378
P(30)		.8676	.8446	.8238	.7959	.7915	.6213
R-Prec			.9245	.8654	.8342	.8320	.5896
Mean Ave Prec				.8840	.8473	.8495	.5612
Recall at .5 Prec					.7707	.7762	.5349
R(1000)						.9212	.5891
Total Rel Ret							.5880

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).

4.2 Comparison of evaluation measures

The `trec_eval` program reports so many different numbers as the evaluation of a single run because there are so many different features of a run that might be of interest. To better understand what aspect of retrieval behavior different effectiveness measures capture, NIST used the TREC-7 automatic ad hoc results to compute correlations between pairs of measures.

The correlations are given in Table 5 and were computed in the following way. Eight different measures (described below) were used in the study. Each run was evaluated using each measure, where the score for a measure was usually the average score for that measure over the 50 topics. The runs were then ranked by score for each measure. The correlation between two different measures was defined as the Kendall's tau correlation between the respective rankings. Kendall's tau computes the distance between two rankings as the minimum number of pairwise adjacent swaps to turn one ranking into the other. The distance is normalized by the number of items being ranked such that two identical rankings produce a correlation of 1.0, the correlation between a ranking and its perfect inverse is -1.0, and the expected correlation of two rankings chosen at random is 0.0.

The following measures were used in the study:

P(10): The precision after the first 10 documents are retrieved.

P(30): The precision after the first 30 documents are retrieved.

R-Prec: The precision after the first R documents are retrieved, where R is the number of relevant documents for the current topic.

Mean Ave Precision: Mean (non-interpolated) average precision as defined above.

Recall at .5 Prec: Recall at the rank where precision first dips below .5 (after at least 10 documents have been retrieved). This measure reflects the heuristic that users will keep looking at a result set while there are more relevant than non-relevant documents being retrieved.

R(1000): The recall after 1000 documents are retrieved.

Total Rel Ret: The total number of relevant documents retrieved across all 50 topics (not an average). The difference between this measure and R(1000) is in the averaging. R(1000) is averaged such that each topic is weighted equally, while the total number of relevant retrieved is dominated by topics that have many relevant documents.

Rank 1st Rel: The rank at which the first relevant document is retrieved.

The correlations between the different measures are all at least .5, showing that each pair of measures is at least somewhat correlated. This is not surprising since all the measures were designed to reflect the quality of a retrieval run. The very high correlation between R(1000) and Total Rels Ret is also not surprising, though the fact the correlation is not 1.0 demonstrates that averaging does have an effect. The weakest correlations are between the Rank 1st Rel measure and each of the others. This is an indication that the Rank 1st Rel measure is in fact a poor measure of retrieval performance. The measure is unstable both because a single topic can have an unreasonable effect on the average score, and because large differences in a score do not reflect the importance of that difference to the user. For example, ranking the first relevant document at rank 103 vs. ranking the first relevant document at rank 957 will cause a large difference in the average score while being essentially meaningless to a user.

One of the current debates in IR is whether recall is important outside a few specific applications such as patent searching. Those who question the utility of recall argue that users never look beyond the "first screen" of results and therefore the only measure that matters is precision at some small cut-off level. Proponents of recall point out that a measure such as P(10) is too coarse-grained for system tuning, even when P(10) is the final measure of interest. The only change in a document ranking that affects P(10) is a relevant document entering or leaving the top 10, while the mean average precision measure is sensitive to the entire ranking. The correlation between P(10) and mean average precision cannot answer which is a better measure, but does show that they measure different things. (The correlation of .7899 represents 384 swaps out of a maximum possible swaps of 3655 since the rankings consist of 86 different runs.)

5 Ad Hoc Retrieval Results

One of the important goals of the TREC conferences is that the participating groups freely devise their own experiments within the TREC task. For some groups this means doing the ad hoc task with the goal of achieving high retrieval effectiveness. For other groups, however, the goals are more diverse and may mean experiments in efficiency or unusual ways of using the data.

This overview of the results discusses the effectiveness of the systems and analyzes some of the similarities and differences in the approaches that were taken. In all cases, readers are referred to the system papers in this proceedings for more details.

The TREC-7 ad hoc evaluation used new topics (topics 351–400) against the documents on Disks 4 and 5 minus the *Congressional Record* documents. There were 103 sets of official results for ad hoc evaluation in TREC-7. Of these, 86 used automatic construction of queries and 17 used manual query construction.

5.1 Automatic runs

Figure 5 shows the recall/precision curves for the eight TREC-7 groups with the highest mean average precision using automatic construction of queries. 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.

ok7ax – OKAPI group ("Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive" by S.E. Robertson, S. Walker, and M. Beaulieu) continued their experiments with the BM25 weighting technique that has been so successful. They tried some experiments with term proximity in the topic, with little success. This particular run is a weighted linear combination of runs made using title, title + description, and full versions of the topic, with pseudo-feedback expansion used in each of these runs before combining. A corrected version of the run (minus the effects of manual index terms in the LA Times) had only slightly degraded performance (see paper for corrected table).

att98atdc – AT&T Labs Research ("AT&T at TREC-7" by A. Singhal, J. Choi, D. Hindle, D. Lewis, and F. Pereira) made two major changes to their TREC-6 algorithms. The first one involved some changes in the term weighting to better accommodate a mix of single terms and phrases (a new phrase list was

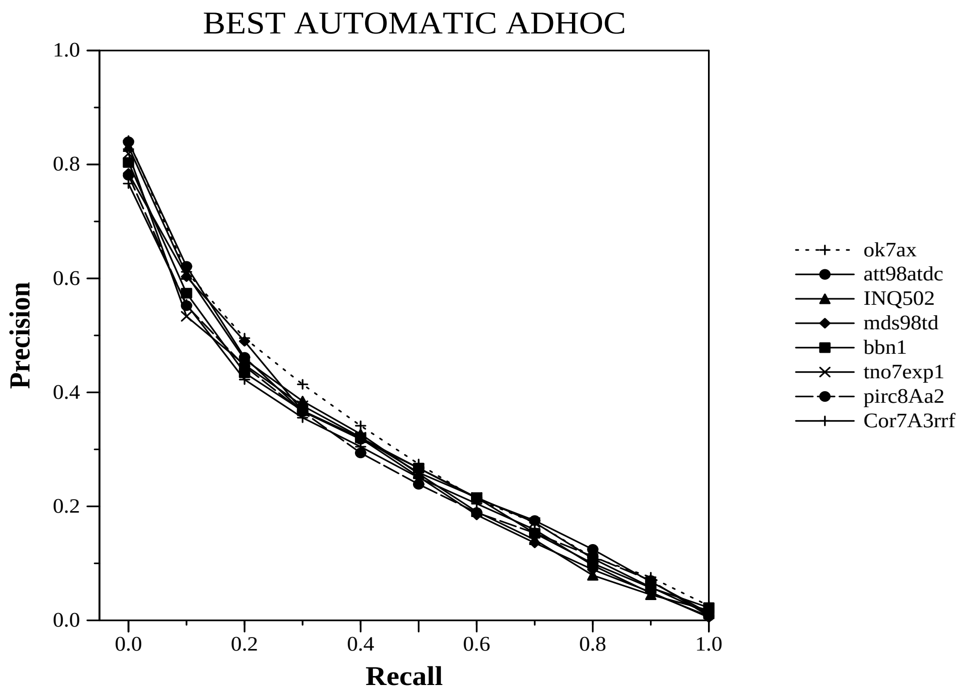

Figure 5: Recall/Precision graph for the top eight automatic ad hoc runs.

used). The second was a change in the automatic expansion methods that allows the use of a larger corpus (all 5 disks) without incurring too much query drift. Whereas the average precision increases by only 3% using this new expansion algorithm, there are fewer topics that are hurt than when using the older expansion method, therefore making it more predictable in actual operational settings.

INQ502 – University of Massachusetts ("INQUERY and TREC-7" by J. Allan, J. Callan, M. Sanderson, J. Xu, and S. Wegmann) ran both the INQUERY phrase recognition and the LCA query expansion (with passages) that had been used in TREC-6. The phrases improved performance by 3.6% in TREC-7, consistent with past work, and the LCA expansion continued to work well, improving performance by 18.5% over non-expanded queries. New for TREC-7 was the use of title terms as a filter to rerank the passages selected for expansion. This filter improved results by 2.8%, but had significant performance effects (positive and negative) on specific topics.

mds98td – MDS/CSIRO ("TREC-7 Ad Hoc, Speech, and Interactive tracks at MDS/CSIRO" by M. Fuller, M. Kaszkiel, D. Kim, C. Ng, J. Robertson, R. Wilkinson, M. Wu and J. Zobel) continued their explorations with the MG system. This year they used two-term statistical phrases, Rocchio relevance feedback, and also tested using passages vs documents. The OKAPI weighting was used for document ranking, and a cosine similarity function was used for passage matching. Their top ranked run used the title and description as input, used documents but not passages, and incorporated phrases and Rocchio expansion. They report a 4.1% improvement for use of phrases, and a 36% gain for expansion. The use of passages instead of documents gave the same results for titles but decreased performance (7.1%) for titles + descriptions.

bbn1 – BBN Technologies ("BBN at TREC-7: Using Hidden Markov Models for Information Retrieval" by D. Miller, T. Leek, and R. Schwartz) based their work on their experience with Hidden Markov models (HMMs) in speech and other language-related recognition problems. This model adapted well to the information retrieval task, working better than the basic OKAPI model without expansion, and still remaining competitive when compared to the full OKAPI results. This was the first entry of BBN to TREC, and they creatively refined the HMM model to handle phrases (bigrams) and pseudo-relevance feedback. These refinements did not improve their results as much as for other IR models; this may

Table 6: Characteristics of best automatic ad hoc runs.

Organization	Topic Parts	D + T	T only	Full Topic	Comments
Okapi group	T,D,N	0.281	0.253 (-10%)	0.284 (1%)	fused run-0.296
AT&T Labs Research	T,D	0.296	0.249 (-16%)		
U. Mass	T,D,N	0.252		0.274 (9%)	title filtered run-0.282
RMIT/UM/CSIRO	T,D	0.281	0.220 (-22%)	0.285 (1%)	
BBN	T,D,N			0.280	
TwentyOne	T,D,N			0.279	
CUNY	T,D,N	0.254	0.243 (-4%)	0.266 (5%)	with phrases-0.272
Cornell/SabIR	T,D,N	0.254*	0.239 (-6%)	0.267 (5%)	*description only

be because the basic HMM model already incorporates to some extent these techniques, or may mean that there are larger improvements to be expected after further work with this new model.

tno7exp1 – Twenty-One group ("Twenty-One at TREC-7: Ad-hoc and Cross-Language track" by D. Hiemstra and W. Kraaij) based their work on the vector-space model but developed a new weighting algorithm using a linguistically motivated probabilistic model. Their new algorithm in its basic form outperformed the basic OKAPI/SMART algorithm by 8%; incorporation of Rocchio-based feedback improved their results by 12%. All runs were made using the full topic.

pirc8Aa2 – Queens College, CUNY ("TREC-7 Ad-Hoc, High Precision and Filtering Experiments using PIRCS" by K.L. Kwok, L. Grunfeld, M. Chan and N. Dinstl) continued work with their spreading activation model using a sequence of 5 different methods to improve their basic results. These methods included the avtf weighting used in TREC-6, a variable Zipf threshold for selecting indexing terms, collection enrichment by using all 5 disks for query expansion, use of the LCA algorithm for expansion, and a reweighting of the query terms using the documents retrieved from this expansion. This group also made extensive investigations into the effects of input query length on results from different methods.

Cor7A3rrf – Cornell/SabIR Research ("SMART High Precision: TREC 6" by C. Buckley, M. Mitra, J. Walz and C. Cardie) tried many variations on their algorithms, including re-examining stemming, phrases, or alternative document clustering methods to do the final term selection from the top retrieved documents. They also investigated differential weighting for titles or for beginnings of documents. None of these variations improved performance significantly and this run uses the TREC-6 clustering approach.

Note that most of these runs use all parts of the topic (att98atdc and mds98td use only the title and description). However there is now a smaller performance difference between runs that use the full topic and runs that use only the title and description sections than was seen in earlier TRECs. This is most likely due to improved query expansion methods, but could be due to variations across topic sets. Table 6 shows the results (official and unofficial as reported in the papers) of these groups using the different topic parts. It should be noted that the improvement going to the full topic is only 1% for several groups. The decrease in performance using only the title is more marked, ranging from 4% to 22%. The TREC-7 title results should be a truer measure of the effects of using the title only than TREC-6, where the descriptions were often missing key terms. However, it is not clear how representative these titles are with respect to very short user inputs and therefore title results should best be viewed as how well these systems could perform on very short, but very good user input.

Looking at individual topic results shows a less consistent picture. Table 7 shows the number of topics that had the best performance from among a group's three runs using different input lengths. (Note that the "long" run for the Okapi system is actually their fused run, and that these are their "official" results as opposed to their corrected ones.) Not only is there a wide variation across topics, there is also a wide

Table 7: Number of topics performing best by topic length.

	Long	Desc	Title
Okapi	28	13	9
CUNY	27	10	13
Cornell	22	17	11

Table 8: More characteristics of best automatic ad hoc runs.

Organization	Model	Weighting/Similarity	Phrase Imp.	Comments
Okapi group	probabilistic	BM25	minimal*	*last reported in TREC-5
AT&T Labs Research	vector	pivot*		*byte normalization
U. Mass	inference net	belief function	3.6%	
RMIT/UM/CSIRO	vector	BM25/cosine		phrases used
BBN	HMM	probabilistic	2%	bigram phrases
TwentyOne	vector	new probabilistic		no phrases used
CUNY	spread. act.	avtf/RSV	2%	phrases used for reranking
Cornell/SabIR	vector	pivot		

variation across systems in that topics that work best at a particular length for one group did not necessarily work best at that length for the other groups.

The merged run from Okapi is taking advantage of this variation by fusing the results from the final runs for each topic length, and they gain a 4% improvement from this fusion. The group from Lexis-Nexis ("Experiments in Query Processing at LEXIS-NEXIS for TREC-7" by A.G. Rao, T. Humphrey, A. Parhizgar, C. Wilson, and D. Pliske) has also worked for several years using fusion between different methods of processing the initial topic, various term weighting algorithms, and different query expansion methods.

Table 8 shows additional characteristics of the systems. These top eight systems are derived from many models and use different term weighting algorithms and similarity measures. Of particular note here is that new models and term weighting algorithms are still being developed, and these are competitive with the more established methods. This applies both to new variations on old weighting algorithms, such as the double log tf weighting from AT&T, and to more major variations such as the new weighting algorithm from TNO, and the completely new retrieval model from BBN.

The fourth column of the table shows the widespread use of phrases in addition to single terms, but the minimal improvement from their use. The biggest improvement reported in the papers was 3.6% from UMass. Whereas most of the other groups are also using phrases, many did not bother to test for differences due to minimal results in earlier years. Cornell reported 7.7% improvement in TREC-6, but this is the improvement on top of the initial baseline, not the improvement after expansion. Private conversations with several of these groups indicate that these improvements are likely to be much less if measured after expansion. As is often the case, these minimal changes in the averages cover a wide variation in phrase performance across topics. A special run by the Okapi group (many thanks) showed less than a 1% average difference in performance, but 19 topics helped by phrases, 14 hurt, and the rest unchanged. Whereas the benefit of phrases is not proven, they are likely to remain a permanent tool in the retrieval systems in a manner similar to the earlier adoption of stemming.

It is interesting to note that many of these groups are using different phrase "gathering" techniques. The Okapi group has a manually-built phrase list with synonym classes that has slowly grown over the years based on mostly past TREC topics. The automatically-produced UMass phrase list was new for TREC-6, the Cornell list was basically unchanged from early TRECs, and the BBN list was based on a new bigram model.

Table 9 shows characteristics of the expansion tools used in these systems. The second column gives the basic expansion model, with the vector-based systems using the Rocchio expansion and other systems using

Table 9: Characterization of query expansion used in best automatic ad hoc runs.

Organization	Expansion/Feedback	Top Docs/Terms added	Disks used	Comments
Okapi group	probabilistic	Full–15/30 T+D–10/30 T only–6/20+title	1-5	
AT&T Labs Research	Rocchio	10/20+5 phrases	1-5	conservative enrichment
U. Mass	LCA	30P/50	1-5	reranking using title terms before expansion
RMIT/UM/CSIRO	Rocchio	10/40+5 phrases	?	additional experiments with passages
BBN	HMM-based	6/?	?	differential weighting on topic parts
TwentyOne	Rocchio	3/200	?	
CUNY	LCA	200P/?	1-5	
Cornell/SabIR	Rocchio	30/25	4-5	clustering, reranking

expansion models more suitable to their retrieval model. The third column shows the number of top-ranked documents (P if passages were used), and the number of terms added from these documents. It should be noted that these numbers are more similar than in earlier TRECs. The fourth column shows the source of the documents being mined for terms; note that most groups have now moved to retrieval from a wider range of documents. Of particular note is the AT&T specific investigation into "conservative enrichment" to avoid the additional noise caused by such wide searching.

Almost all groups use some type of query expansion. The group from NEC and the Tokyo Institute of Technology ("Ad Hoc Retrieval Experiments Using WordNet and Automatically Constructed Thesauri" by R. Mandala, T. Tokunaga, H. Tanaka, A. Okumura, and K. Satoh) experimented with three different thesauri, including WordNet, a simple co-occurrence-based thesaurus, and a new thesaurus that was automatically built using predicate-argument structures. The University of North Carolina ("IRIS at TREC-7" by K. Yang, K. Maglaughlin, L. Meho, and R.G. Sumner, Jr.) tried two types of relevance feedback approaches and discovered that they performed very differently, particularly when used in the various subcollections of TREC. Fujitsu Laboratories ("Fujitsu Laboratories TREC7 Report" by I. Namba, N. Igata, H. Horai, K. Nitta and K. Matsui) experimented with expansion using the top N documents, and alternatively with expansion using the top M clusters. A final example of work in query expansion is a new method based on relative entropy ("Information term selection for automatic query expansion" by C. Carpineto and G. Romano from Fondazione Ugo Bordoni, Rome and R. De Mori from the University of Avignon).

Although some type of query expansion is clearly necessary for top results, several groups did not use expansion in order to investigate some particular component of retrieval. The SPIDER system ("SPIDER Retrieval System at TREC7" by M. Braschler and M. Wechsler from Eurospider Information Technology and B. Mateev, E. Mittendorf, and P. Schäuble from the Swiss Federal Institute of Technology (ETH)) specifically explored the use of proximity and co-occurrence of terms within the description as an alternative approach to using expansion. NTT DATA ("NTT DATA at TREC-7: system approach for ad-hoc and filtering" by H. Nakajima, T. Takaki, T. Hirao, and A. Kitauchi) also used a new scoring method based on coordination and what they called the "degree of importance" for various co-occurring terms.

Two groups did experiments with various indexing methods. Johns Hopkins University ("Indexing Using Both N-Grams and Words" by J. Mayfield and P. McNamee) worked with both 5-grams and words, including adaptation of the 5-gram method to properly handle stopwords. A joint project led by Tomek Strzalkowski

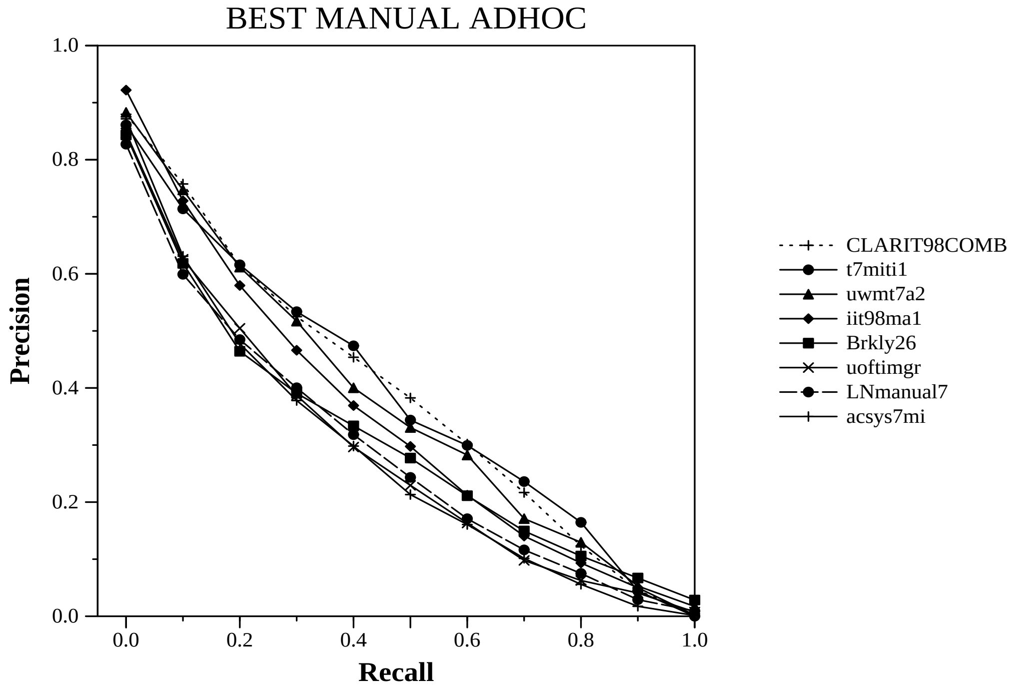

Figure 6: Recall/Precision graph for the top eight manual ad hoc runs.

continued the investigation of merging results from multiple streams of input using different indexing methods ("Natural Language Information Retrieval: TREC-7 Report" by T. Strzalkowski, G. Stein, and G. Bowden Wise from GE Research & Development, J. Perez-Carballo from Rutgers University, P. Tapananinen, T. Jarvinen, and A. Voutilainen from the University of Helsinki and J. Karlgren from the Swedish Institute of Computer Science).

5.2 TREC-6 ad hoc manual results

Figure 6 shows the recall/precision curves for the eight TREC-7 groups with the highest mean average precision scores using manual construction of queries.

CLARIT98COMB – CLARITECH Corp. ("Effectiveness of Clustering in Ad-Hoc Retrieval" by D.A. Evans, A. Huettner, X. Tong, P. Jansen, and J. Bennett) performed a user experiment measuring the difference in performance between presentation modes: a ranked list vs. a clustered set of documents. Starting from a fixed set of initial queries, users spent 30 minutes viewing documents (using the specified presentation mode), marking some as relevant and modifying the queries. The set of "known" relevant documents was used in 2 ways: as input for a Rocchio feedback run to get a 1000-document ranking for each topic and as a fixed set that were then shuffled to the top of each list.

m7miti1 – Management Information Technologies, Inc. ("Readware© Text Analysis and Retrieval in TREC-7" by T. Adi, O.K. Ewell, and P. Adi) used one analyst to manually formulate many queries per topic (an average of 18). These queries used various Readware tools and the user continued to formulate/modify these queries based on information in the retrieved documents. Since this system does no ranking of documents, the final ranked list was composed of those documents judged relevant by the searcher (5898 in all), ordered by the "complexity" of the queries that were used to retrieve the documents.

uwmt7a2 – University of Waterloo ("Deriving Very Short Queries for High Precision and Recall (MultiText Experiments for TREC-7)" by G.V. Cormack, C.R. Palmer, and M. Van Biesbrouck) investigated the building of the ideal very short query. They first spent less than 30 minutes per topic making relevance judgments, in the method used in TREC-6. These 5529 documents were then used to automatically generate a series of Boolean queries and the query yielding the highest average precision for a given topic was used. The queries were also constrained to having 3 terms or less, with the average number of terms being 1.86. The results of these "ideal" queries were what was submitted to NIST, both in the ad hoc task and in the Very Large Corpus track.

iit98ma1 – ("Use of Query Concepts and Information Extraction to Improve Information Retrieval Effectiveness" by D.O. Holmes from NCR Corporation, D.A. Grossman from U.S. Government, O. Frieder and A. Chowdhury from the Illinois Institute of Technology, and M.C. McCabe from Advanced Analytic Tools) continued work with their parallel database retrieval model. This manual run was done to test the use of phrases and proper nouns. The queries were constructed manually in 15-30 minutes per topic, using the system to retrieve documents to "mine" for good terms. The emphasis was on selecting good phrases and/or proper nouns. The full set of 50 queries consisted of 908 phrases, and 389 single terms. Of these, 541 were proper nouns including 235 names of people.

Brkly26 – University of California at Berkeley ("Manual Queries and Machine Translation in Cross-Language Retrieval and Interactive Retrieval with Cheshire II at TREC-7" by F.C. Gey, H. Jiang, A. Chen and R.R. Larson) explored the possibilities of using manually-constructed Boolean queries to improve performance. For 17 of the 50 topics they constructed these queries; note that for the rest it seemed doubtful that there would be any improvement. For all topics they did a standard automatic logistic regression ranking, but for the 17 Boolean queries they merged the results of the logistic run and the Boolean results. This method improved results for 14 of the 17 topics, 3 very dramatically.

uoftimgr – University of Toronto ("ClickIR: Text Retrieval using a Dynamic Hypertext Interface" by R.C. Bodner and M.H. Chignell) used their dynamic hypertext model to build the queries. Users took approximately 15 minutes to browse the collection, clicking on sentences that then linked to one or two new documents. These links were created automatically from terms used in previous queries from the same user. The final query that created the results sent to NIST was then assembled from these "clicked on" sentences (which were logged). The INQUERY system was used for production of the ranked list, and also to make an additional run in which "known" relevant documents were used for relevance feedback.

LNmanual7 – Lexis-Nexis ("Experiments in Query Processing at LEXIS-NEXIS for TREC-7" by A.G. Rao, T. Humphrey, A. Parhizgar, C. Wilson, and D. Pliske) experimented with human relevance feedback as opposed to automatic feedback from the top 20 documents. The users read the top 20 documents and then modified the initial automatic query, adding terms and phrases, but only if those phrases and terms were in the index. The users also had to supply the new term weighting. This hand-picking of the expansion terms improved performance by 28% over the baseline shown in the paper and 14% over the best automatic Lexis-Nexis run, with most improvements in the higher-precision areas of the graph.

acsys7mi – CSIRO, Australian National University ("ACSys TREC-7 Experiments" by D. Hawking from CSIRO Mathematics and Information Sciences and N. Craswell and P. Thistlewaite from the Australian National University) explored both manual editing of the query before feedback, additional editing of the query after viewing documents, and the use of concept scoring to better balance terms representing different concepts. The new Quokka GUI system displayed the concepts in various colors to aid in this task, and a median of 10.6 minutes was taken per topic to produce the final query used in ranking the documents. A major system error affected the results and corrected tables can be seen in the paper.

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 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.

Table 10: Number of track participants.

	TREC-6	TREC-7
CLIR	13	9
filtering	10	12
HP	5	4
interactive	9	8
query	0	2
SDR	13	10
VLC	7	6

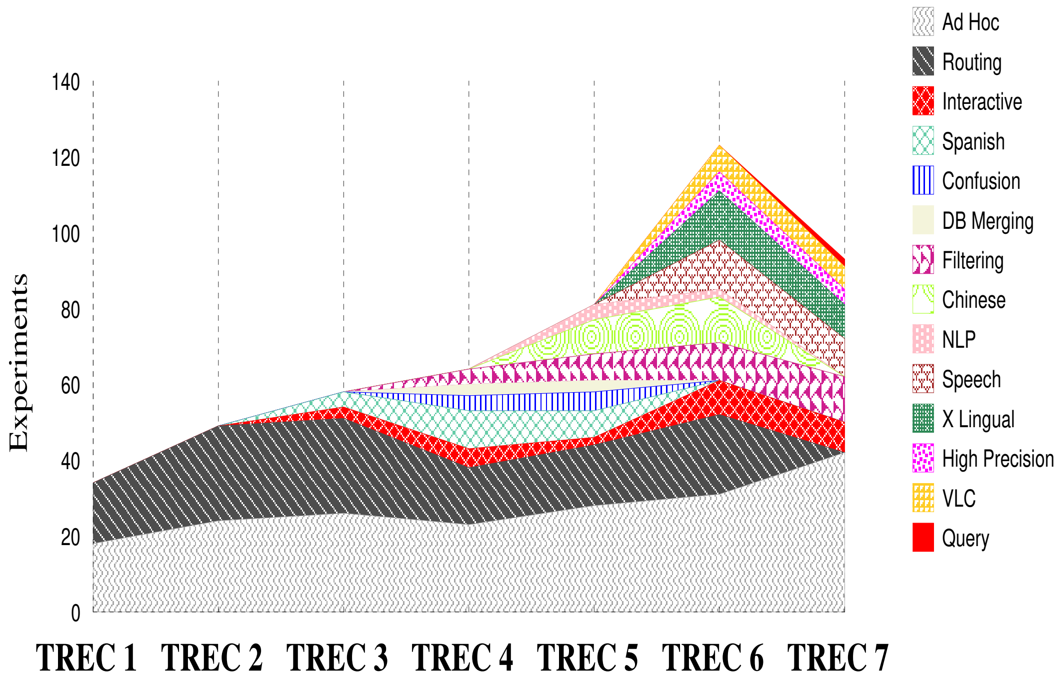

Figure 7: Number of TREC experiments by TREC task.

The tracks have had a significant impact on TREC participation. Figure 7 shows the number of experiments performed in each TREC, where the set of runs submitted for a track by one group is counted as one experiment. The number of experiments increased each year through TREC-6 then decreased in TREC-7, mostly due to the elimination of the routing main task and the Chinese track. The number of participants performing the ad hoc task continues to grow, with 42 groups taking part in TREC-7 compared to 31 in TREC-6. Table 10 gives the number of participants in each of the TREC-7 tracks for both TREC-6 and TREC-7.

The main ad hoc task provides an entry point for new participants and provides a baseline of retrieval performance. The tracks invigorate TREC by focusing research on new areas or particular aspects of text retrieval. To the extent the same retrieval techniques are used for the different tasks, the tracks also validate the findings of the ad hoc task.

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 TREC-7 tracks. The overall goals, the experimental design used, and a very brief summary of the results are listed for each track. 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 searching for documents in one language using topics in a different language. The first CLIR track was held in TREC-6 [6]. Three document sets were used in TREC-6: a set of French documents from the Swiss news agency *Schweizerische Depeschen Agentur* (SDA); a set of German documents from SDA plus a set of articles from the newspaper *New Zurich Newspaper* (NZZ); and a set of English documents from the AP newswire. All of the document sets contain news stories from approximately the same time period, but are not aligned or specially coordinated with one another. A set of 25 topics that were translated into each of the languages was also provided. Participants searched for documents in one target language using topics written in a different language.

The TREC-7 task expanded on this beginning. The document set for the TREC-7 track consisted of all the documents used in the TREC-6 track plus the Italian version of the SDA for the same time period. Participants were provided with a new set of 28 topics (with translations available in English, French, German, and Italian), and used one topic language to search the combined document set. That is, a single run retrieved documents written in different languages. Since some participants were not able to process all four languages, a second task in which English topics were run against the combined French and English document set was also run.

The TREC-7 track also defined an optional subtask. The subtask used a different document 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.

Nine groups participated in the TREC-7 CLIR track, with five groups performing the test on the full four language collection, and seven groups performing the test on the English and French collection. No runs were submitted for the optional subtask, however this subtask is planned to be repeated in TREC-8 now that groups have more experience with cross language retrieval. The results of the track demonstrate that very different approaches to cross-language retrieval can lead to comparable retrieval effectiveness.

The construction of the cross language test collection differed from the way any other TREC collection has been created. Candidate topics in the native language were created in each of four different institutions: NIST (English); EPFL Lausanne, Switzerland (French); Informationszentrum Sozialwissenschaften, Bonn, Germany (German); and CNR, Pisa, Italy (Italian). Each institution created topics that would target documents in its corresponding language. NIST selected the final set of 28 topics, balancing the set so that each institution contributed seven topics. Each of the final topics was 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. This is the first time that TREC has used multiple relevance assessors for a single topic.

For the complete overview of the track see "Cross-Language Information Retrieval (CLIR) Track Overview" by M. Braschler, J. Krause, C. Peters, and P. Schäuble.

6.2 The Filtering track

Each of the previous TRECs have had a second main task called the routing task. A routing task investigates the performance of systems that use standing queries to search new streams of documents, as news clipping services and library profiling systems do, for example. As the routing task has been defined in TREC, participants use old topics with existing relevance judgments to form routing queries, and use those queries to rank a previously unseen document collection. Real routing applications generally require a system to make a binary decision whether or not to retrieve the current document, however, not simply form a ranking of a document set. The filtering track was started in TREC-4 to address this more difficult version of the routing task.

The TREC-7 filtering track contained three tasks of increasing difficulty (and realism). For each task, topics 1–50 and the AP newswire collection on Disks 1–3 were used (with different splits into training and test sets, depending on the task). The first task was the traditional routing task. The second task was a *batch* filtering task in which systems are given topics and relevance judgments as in the routing task, and must then decide whether or not to retrieve each document in the test portion of the collection. This task is what previous filtering tracks performed. The third task, and the focus of the track, was an *adaptive* filtering task. In this task, a filtering system starts with just the 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 its query as desired.

Because filtering tasks return an unordered set of documents, not a ranking, different evaluation measures from those used for ad hoc or routing are required. 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 [5]. In TREC-7 two different utility functions were used:

$$\begin{aligned} \mathrm{F1} &= 3\mathrm{R}^{+} - 2\mathrm{N}^{+} \\ \mathrm{F3} &= 4\mathrm{R}^{+} - \mathrm{N}^{+} \end{aligned}$$

where R^{+} and N^{+} are the number of relevant and non-relevant documents retrieved, respectively. A problem with utilities as measures is that different topics have widely varying possible utility values, making it difficult to meaningfully compare scores across topics or compute an average score. An approach to scaling and normalizing utilities was introduced in this year's track [3].

Twelve groups submitted at least one TREC-7 filtering run. A total of 46 runs was submitted, consisting of 10 routing runs, 12 batch filtering runs, and 24 adaptive filtering runs. The track results demonstrated that adaptive filtering is a challenging problem for current systems. Indeed, when using the F1 utility measure to evaluate performance, the "baseline" system which retrieves no documents was the most effective system overall. Comparison with batch filtering results show that setting an appropriate threshold for when to retrieve a document is a critical, and difficult, task in adaptive filtering.

For the complete overview of the track see "The TREC-7 Filtering Track: Description and Analysis" by D.A. Hull.

6.3 The High Precision track

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

Since the track guidelines put no limits on who the user could be, an implicit assumption of the track is that the runs are performed by system experts. As such, the track provides an upper-bound on the effectiveness obtainable by the systems. The 5-minute time limit was selected so that the intrinsic effectiveness of the system, the system efficiency, and the user interface would all be tested by the task. The same 50 topics and document set as used in the TREC-7 ad hoc task were used for the HP track.

Four groups participated in the TREC-7 track, submitting a total of seven runs. One finding of the track was that retrieving 15 good documents is a simple enough task for current retrieval systems that disagreements between the searcher and the assessor regarding what constitutes a relevant document bounds performance. However, new time-based evaluation measures introduced in the track offer a possible solution.

For the complete overview of the track see "The TREC-7 High Precision Track" by C. Buckley.

6.4 The Interactive track

The interactive track is another track that was started in TREC-4. The high-level goal of the track is 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-7 track used an experimental framework designed to provide an estimate of the difference between an experimental and a control system that is uncontaminated by the differences between searchers and topics.

The experimental framework both defined a common task for participants to perform and prescribed an experimental matrix. The search task used the title and description sections plus a special "Instances" section of eight ad hoc topics; the documents searched were the *Financial Times* collection from Disk 4. The topics each described a need for information of a particular type such that multiple distinct examples or instances of that information were contained in the document collection. The searchers job was to save documents covering as many distinct answers to the question as possible in a 15-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 psychometric data from the searchers, and to report extensive data on each searcher's interactions with the search systems.

The experimental matrix defined how searchers and topics were to be divided among the experimental and control systems. (Participants were free to choose whatever systems they wanted to serve as experimental and control. That is, the track did not attempt to coordinate cross-site comparisons or test particular hypotheses.) The matrix was based on a latin square design, which provides the desired uncontaminated estimate of the difference between the systems. The minimum experiment defined by the design required eight searchers, with each searcher performing four searches with each of the two systems. The eight-searcher minimum was imposed since the results of the TREC-6 track suggested that with eight topics at least eight searchers are required to obtain statistically significant results [4].

Eight groups participated in the interactive track, performing a total of ten experiments. Since comparison of systems across sites was not supported by the experimental design, the results of the track need to be understood in the context of the particular research goals of the individual research groups.

For the complete overview of the tracks see "TREC-7 Interactive Track Report" by P. Over.

6.5 The Query track

The query track was a new track whose goal was to create a large query collection. The variability in topic performance (e.g., see the discussion of the effect of topic length on performance in Section 5.1) 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, topics 1–50.

Participants in the track created different types of queries from the topic statements and/or relevance judgments. A query of a given type was created for each of the 50 topics, forming one query set. Five 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).

Manual structured query: a manually constructed query based on the topic statement and relevant documents. The use of operators supported by the participant's system was encouraged. The TIPSTER DN2 format was used to represent the query structure.

Automatic structured query: a query constructed automatically from the topic statement and relevance judgments. TIPSTER DN2 format used to represent the query structure.

Participants exchanged the query sets they created with all other participants in the track, and all participants ran all query sets their system could support. The document set used for the runs was the documents on Disk 2 plus the AP collection on Disk 3. The retrieval results were submitted to NIST where all runs were judged and evaluated.

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.

Unfortunately, only two groups participated in the query track, too few to make any meaningful comparisons. The track will run again in TREC-8, with the hope that heightened awareness of the problems the query track is addressing will generate participation.

For the complete overview of the track see "The TREC-7 Query Track" by C. Buckley.

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.

As the earlier confusion tracks had used, the TREC-6 SDR track used a known-item search task. The TREC-7 track implemented a full ranked retrieval task. The document collection consisted of transcripts of approximately 100 hours of broadcast news programs, representing about 3000 news stories. Participants worked with four different versions of the transcripts: the *reference* transcripts, which were hand-produced and assumed to be perfect; the first *baseline* transcripts, which were produced by a baseline speech recognition system running at about 35% word error rate; a second set of baseline transcripts, produced by the baseline recognizer running at about 50% word error rate; and the *recognizer* transcripts, which were produced by the participant's own recognizer system. Document boundaries were given in the hand-produced transcripts, and the same boundaries were used in the other versions.

NIST created a set of 23 topics, which were used to search each of the versions of the transcripts. 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. To make this comparison as complete as possible, participants were encouraged to retrieve using other groups' recognizer transcripts as well. These runs are called *cross-recognizer* runs.

Eleven groups participated in the TREC-8 SDR track. The results of this year's track display a linear correlation between the error rate of the recognition and a decrease in retrieval effectiveness, a correlation that was not present in last year's track that used a known-item search task. Not surprisingly, the correlation is stronger when recognizer error rate is computed over content-based words (e.g., named entities) rather than *all* words.

For the complete overview of the track see "1998 TREC-7 Spoken Document Retrieval Track Overview and Results" by J. Garofolo, E.M. Voorhees, C.G.P. Auzanne, V.M. Stanford, and B.A. Lund.

6.7 The Very Large Corpus (VLC) track

The VLC track explores how well retrieval algorithms scale to larger document collections. In contrast to the ad hoc task that uses a 2 GB document collection, the first running of the VLC track in TREC-6 used a 20 GB collection, while the TREC-7 track used a 100 GB document collection. The TREC-7 collection consisted of World Wide Web data that was collected by the Internet Archive (`http://www.archive.org`). The track used the TREC-7 ad hoc topics, and a set of relevance judgments produced by assessors at the Australian National University. Because of the difficulty of getting sufficient relevance judgments to accurately measure recall, the main effectiveness measure used for VLC runs was precision after 20 documents were retrieved.

To more accurately measure the effect size has on the retrieval systems used by the participants, the track provided 3 collections: the original 100 GB collections plus 1% and 10% subsamples. Participants indexed each of the three collections and ran the entire topic set on each. They then reported timing figures for each phase as well as the top 20 retrieved. The main evaluation measures were precision after 20 documents retrieved (the effectiveness measure); query response time (elapsed time as seen by the user); data structure (e.g., inverted index) building time (elapsed time as seen by the user); plus a combination timing measure that factored in the expense of the hardware used.

Seven groups participated in the VLC track, with six groups processing the entire 100 GB corpus. The track demonstrated that processing a 100 GB corpus is well within the capabilities of today's retrieval systems. Of particular note was the Multitext group that achieved sub-second query processing time while maintaining good retrieval effectiveness using hardware that cost under US$100,000.

For the complete overview of the track see "Overview of TREC-7 Very Large Collection Track" by D. Hawking, N. Craswell, and P. Thistlewaite.

7 The Future

The final session of each TREC workshop is a planning session for future TRECs—especially to decide on the set of tracks for the next TREC. Two new tracks are planned for TREC-8, the question answering track and the Web track. The question answering track is designed to encourage research on methods for *information* retrieval as opposed to document retrieval. The goal in the track will be for systems to produce short text extracts that contain the answer for each of a set of 200 questions. The goal in the Web track will be to investigate whether links can be used to enhance retrieval. The track will use a 2 GB subset of the data collected for the VLC track and a typical TREC ad hoc task. Also, participation in the query track is encouraged, since the benefits of that track increase with increased participation.

Acknowledgments

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

References

[1] D. K. Harman, editor. *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*, October 1996. NIST Special Publication 500-236.

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

[3] David A. Hull. The TREC-7 filtering track: Description and analysis. In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, August 1999. NIST Special Publication 500-242.

[4] Eric Lagergren and Paul Over. Comparing interactive information retrieval systems across sites: The TREC-6 interactive track matrix experiment. In 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*, pages 164–172, Melbourne, Australia, August 1998. ACM Press, New York.

[5] David D. Lewis. The TREC-4 filtering track. In Harman [1], pages 165–180. NIST Special Publication 500-236.

[6] Peter Schäuble and Páraic Sheridan. Cross-lanugage information retrieval (CLIR) track overview. In Voorhees and Harman [12], pages 31–43. NIST Special Publication 500-240.

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

[8] 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.

[9] 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.

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

[11] Ellen M. Voorhees and Donna Harman. Overview of the sixth Text REtrieval Conference (TREC-6). In Voorhees and Harman [12], pages 1–24. NIST Special Publication 500-240.

[12] E.M. Voorhees and D.K. Harman, editors. *Proceedings of the Sixth Text REtrieval Conference (TREC-6)*, August 1998. NIST Special Publication 500-240.