Overview of the Ninth Text REtrieval Conference (TREC-9)

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

1 Introduction

The ninth Text REtrieval Conference (TREC-9) was held at the National Institute of Standards and Technology (NIST) on November 13–16, 2000. The conference was co-sponsored by NIST, the Information Technology Office of the Defense Advanced Research Projects Agency (DARPA/ITO), and the Advanced Research and Development Activity (ARDA) office of the Department of Defense.

TREC-9 is the latest in a series of workshops designed to foster research in text retrieval. The workshop series has four 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.

The previous eight TRECs each had an "ad hoc" main task through which eight large test collections were built [17]. In recognition that sufficient infrastructure exists to support researchers interested in this traditional retrieval task, the ad hoc main task was discontinued in TREC-9 so that more TREC resources could be focused on building evaluation infrastructure for other retrieval tasks (called "tracks"). The seven tracks included in TREC-9 were Cross-Language Retrieval, Filtering, Interactive Retrieval, Query Analysis, Question Answering, Spoken Document Retrieval, and Web Retrieval.

Table 1 lists the groups that participated in TREC-9. Sixty-nine groups including participants from 17 different countries were represented. The diversity of the participating groups ensures that TREC represents many different approaches to text retrieval.

This paper serves as an introduction to the research described in detail in the remainder of the volume. The next section provides a summary of the retrieval background knowledge that is assumed in the other papers. Section 3 presents a short description of each track—a more complete description of a track can be found in that track's overview paper in the proceedings. The final section looks forward to future TREC conferences.

2 Text Retrieval

Text retrieval, also called information retrieval or document retrieval, is concerned with locating documents that are relevant to a user's information need. Traditionally, the emphasis in text retrieval research has been to provide access to natural language texts where the set of documents to be searched is large and topically diverse. Since the documents are free text and not specially structured for access by computers, standard database technologies are not effective solutions to the problem.

The prototypical retrieval task is a researcher doing a literature search in a library. In this environment the retrieval system knows the set of documents to be searched (the library's holdings), but cannot anticipate

Table 1: Organizations participating in TREC-9

Australian National University/CSIRO	NTT DATA Corporation
Alicante University	Oregon Health Sciences University
AT&T Labs Research	Pam Wood
BBN Technologies	Queens College, CUNY
Chapman University	New Mexico State University
Chinese University of Kong Kong	RICOH Co., Ltd.
CL Research	RMIT University/CSIRO
Carnegie Mellon University (2 groups)	Rutgers University (2 groups)
Conexor Oy	Sabir Research
CWI, The Netherlands	Seoul National University
Dipartimento di Informatica, Pisa	Sheffield/Cambridge/SoftSound/ICSI
Dublin City University	Southern Methodist University
Fudan University	State University of New York at Buffalo
Fujitsu Laboratories, Ltd.	Sun Microsystems
Hummingbird Communications	Syracuse University
IBM T. J. Watson Research Center (2 groups)	Trans-EZ Inc.
IIT/AAT/NCR	TwentyOne
Imperial College	University of Alberta
Informatique-CDC	University of California, Berkeley
IRIT/SIG	University of Cambridge
Johns Hopkins University	University of Glasgow
Justsystem Corporation	University of Iowa
KAIST	University of Maryland, College Park
Katholieke Universiteit Nijmegen	University of Massachusetts
KDD R&D Laboratories/Waseda University	University of Melbourne
Korea University	Universite de Montreal
LIMSI (2 groups)	University of North Carolina, Chapel Hill
Microsoft (2 groups)	Universite de Neuchatel
MITRE	University of Padova
MNIS-TextWise Labs	University of Sheffield
MuliText Project	USC-ISI
National Taiwan University	Xerox Research Centre Europe
NeurOK, LLC	

the particular topic that will be investigated. We call this an *ad hoc* retrieval task, reflecting the arbitrary subject of the search and its short duration. Other examples of ad hoc searches are web surfers using Internet search engines, lawyers performing patent searches or looking for precedences in case law, and analysts searching archived news reports for particular events. A retrieval system's response to an ad hoc search is generally a list of documents ranked by decreasing similarity to the query.

In a document routing or *filtering* task, the topic of interest is known and stable, but the document collection is constantly changing [3]. For example, an analyst who wishes to monitor a news feed for items on a particular subject requires a solution to a filtering task. The filtering task generally requires a retrieval system to make a binary decision whether to retrieve each document in the document stream as the system sees it. The retrieval system's response in the filtering task is therefore an unordered set of documents (accumulated over time) rather than a ranked list.

Text retrieval has traditionally focused on returning documents that contain answers to questions rather than returning the answers themselves. This emphasis is both a reflection of retrieval systems' heritage as library reference systems and an acknowledgement of the difficulty of question answering. However, for certain types of questions, users would much prefer the system to answer the question than be forced to

```
<num> Number:  451
<title> What is a Bengals cat?

<desc> Description:
Provide information on the Bengal cat breed.
<narr> Narrative:
Item should include any information on the Bengal cat breed, including description, origin,
characteristics, breeding program, names of breeders and catteries carrying bengals.
References which discuss bengal clubs only are not relevant.  Discussions of bengal tigers
are not relevant.
```

Figure 1: A sample TREC-9 topic from the web track.

wade through a list of documents looking for the specific answer. To encourage research on systems that return answers instead of document lists, TREC introduced a question answering task in 1999.

2.1 Test collections

Text retrieval has a long history of using retrieval experiments on test collections to advance the state of the art [5, 10, 13], and TREC continues this tradition. A test collection is an abstraction of an operational retrieval environment that provides a means for researchers to explore the relative benefits of different retrieval strategies in a laboratory setting. Test collections consist of three parts: a set of documents, a set of information needs (called *topics* in TREC), and *relevance judgments*, an indication of which documents should be retrieved in response to which topics.

2.1.1 Documents

The document set of a test collection should be a sample of the kinds of texts that will be encountered in the operational setting of interest. It is important that the document set reflect the diversity of subject matter, word choice, literary styles, document formats, etc. of the operational setting for the retrieval results to be representative of the performance in the real task. Frequently, this means the document set must be large. The TREC test collections created in previous years' ad hoc main tasks used about 2 gigabytes of text (between 500,000 and 1,000,000 documents). The document sets used in various tracks have been smaller and larger depending on the needs of the track and the availability of data.

The TREC document sets consist mostly of newspaper or newswire articles, though there are also some government documents (the *Federal Register*, patent applications) and computer science abstracts (*Computer Selects* by Ziff-Davis publishing) included. High-level structures within each document are tagged using SGML, and each document is assigned an unique identifier called the DOCNO. In keeping of the spirit of realism, the text was kept as close to the original as possible. No attempt was made to correct spelling errors, sentence fragments, strange formatting around tables, or similar faults.

2.1.2 Topics

TREC distinguishes between a statement of information need (the topic) and the data structure that is actually given to a retrieval system (the query). The TREC test collections provide topics to allow a wide range of query construction methods to be tested and also to include a clear statement of what criteria make a document relevant. The format of a topic statement has evolved since the beginning of TREC, but it has been stable for the past several years. A topic statement generally consists of four sections: an identifier, a title, a description, and a narrative. An example topic taken from this year's web track is shown in figure 1.

The different parts of the TREC topics allow researchers to investigate the effect of different query lengths on retrieval performance. The "titles" in topics 301–450 were specially designed to allow experiments with very short queries; those title fields consist of up to three words that best describe the topic. (The title field was used differently in topics 451–500, this year's web track topics, as described below.) The description field is a one sentence description of the topic area. The narrative gives a concise description of what makes a document relevant.

Participants are free to use any method they wish to create 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.

TREC topic statements are created by the same person who performs the relevance assessments for that topic (the *assessor*). Usually, each assessor comes to NIST with ideas for topics based on his or her own interests, and searches the document collection using NIST's PRISE system to estimate the likely number of relevant documents per candidate topic. The NIST TREC team selects the final set of topics from among these candidate topics based on the estimated number of relevant documents and balancing the load across assessors.

This standard procedure for topic creation was changed for topics 451–500. These topics were to be used in the web track, and participants were concerned that the queries that users type into current web search engines are quite different from standard TREC topic statements. However, if participants were given only the literal queries submitted to a web search engine, they would not know the criteria by which documents would be judged. As a compromise, standard TREC topic statements were retrofitted around actual web queries. NIST obtained the log of queries that were submitted to the Excite search engine on December 20, 1999[1]. A sample of queries that were deemed acceptable for use in a government-sponsored evaluation was given to the assessors. Each assessor selected a query from the sample and developed a description and narrative for that query. The assessors were instructed that the original query might well be ambiguous (e.g., "cats"), and they were to develop a description and narrative that were consistent with any one interpretation of the original (e.g., "Where is the musical Cats playing?"). They then searched the web document collection to estimate the likely number of relevant documents for that topic. The "title" field of topics 451–500 contains the literal query that was the seed of the topic. Unlike other TREC topics and the description and narrative fields of these topics, the title field contains all of the spelling and grammatical errors of the original Excite query.

2.1.3 Relevance judgments

The relevance judgments are what turns a set of documents and topics into a test collection. Given a set of relevance judgments, the retrieval task is then to retrieve all of the relevant documents and none of the irrelevant documents. TREC almost always uses binary relevance judgments—either a document is relevant to the document or it is not. To define relevance for the assessors, the assessors are told to assume that they are writing a report on the subject of the topic statement. If they would use any information contained in the document in the report, then the (entire) document should be marked relevant, otherwise it should be marked irrelevant. The assessors are instructed to judge a document as relevant regardless of the number of other documents that contain the same information.

Relevance is inherently subjective. Relevance judgments are known to differ across judges and for the same judge at different times [11]. Furthermore, a set of static, binary relevance judgments makes no provision for the fact that a real user's perception of relevance changes as he or she interacts with the retrieved documents. Despite the idiosyncratic nature of relevance, test collections are useful abstractions because the *comparative* effectiveness of different retrieval methods is stable in the face of changes to the relevance judgments [15].

The relevance judgments in early retrieval test collections were complete. That is, a relevance decision was made for every document in the collection for every topic. The size of the TREC document sets makes complete judgments utterly infeasible—with 800,000 documents, it would take over 6500 hours to judge the entire document set for one topic, assuming each document could be judged in just 30 seconds. Instead,

[1] Jack Xu of Excite released this log on his ftp site at `ftp.excite.com/pub/jack`.

TREC uses a technique called pooling [12] to create a subset of the documents (the "pool") to judge for a topic. Each document in the pool for a topic is judged for relevance by the topic author. Documents that are not in the pool are assumed to be irrelevant to that topic.

The judgment pools are created as follows. When participants submit their retrieval runs to NIST, they rank their runs in the order they prefer them to be judged. NIST chooses a number of runs to be merged into the pools, and selects that many runs from each participant respecting the preferred ordering. For each selected run, the top X documents (usually, $X = 100$) per topic are added to the topics' pools. Since the retrieval results are ranked by decreasing similarity to the query, the top documents are the documents most likely to be relevant to the topic. Many documents are retrieved in the top X for more than one run, so the pools are generally much smaller the theoretical maximum of $X \times$ *the-number-of-selected-runs* documents (usually about 1/3 the maximum size).

The use of pooling to produce a test collection has been questioned because unjudged documents are assumed to be not relevant. Critics 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.

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 [20]. 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 0.5 %. 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 %.

A similar investigation of the TREC-8 ad hoc collection showed that every automatic run that had a mean average precision score of at least .1 had a percentage difference of less than 1 % between the scores with and without that group's uniquely retrieved relevant documents [16]. That investigation also showed that the quality of the pools is significantly enhanced by the presence of recall-oriented manual runs, an effect noted by the organizers of the NTCIR (NACSIS Test Collection for evaluation of Information Retrieval systems) workshop who performed their own manual runs to supplement their pools [9].

While the lack of any appreciable difference in the scores of submitted 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. Indeed, the differences in scores resulting from incomplete pools observed here are smaller than the differences that result from using different relevance assessors [15].

2.2 Evaluation

Retrieval runs on a test collection can be evaluated in a number of ways. In TREC, all ad hoc tasks (i.e., all tasks that involve returning a ranked list of documents) are evaluated using the `trec_eval` package written by Chris Buckley of Sabir Research [4]. This package reports about 85 different numbers for a run, including *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.

Of all the numbers reported by `trec_eval`, the recall-precision curve and mean (non-interpolated) average precision are the most commonly used measures to describe TREC retrieval results. 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`).

3 TREC-9 Tracks

TREC's track structure was begun in TREC-3 (1994). The tracks serve several purposes. First, tracks act as incubators for new research areas: the first running of a track often defines what the problem *really* is, and a track creates the necessary infrastructure (test collections, evaluation methodology, etc.) to support research on its task. The tracks also demonstrate the robustness of core retrieval technology in that the same techniques are frequently appropriate for a variety of tasks. Finally, the tracks make TREC attractive to a broader community by providing tasks that match the research interests of more groups.

Table 2 lists the different tasks that were in each TREC, the number of groups that submitted runs to that task, and the total number of groups that participated in each TREC. The tasks within the tracks offered for a given TREC have diverged as TREC has progressed. This has helped fuel the growth in the number of participants, but has also created a smaller common base of experience among participants since each participant tends to submit runs to fewer tracks.

This section describes the tasks performed in the TREC-9 tracks. See the track reports elsewhere in this proceedings for a more complete description of each track.

3.1 The Web track

The purpose of the web track was to build a test collection that more closely mimics the retrieval environment of the World Wide Web. In creating such a collection, a variety of web retrieval strategies were also investigated. The task in the track was a traditional ad hoc retrieval task where the documents were a collection of web pages.

The web track was coordinated by David Hawking and his colleagues at CSIRO and the Australian National University. They obtained a snapshot of the web from 1997 from the Internet Archive, and produced several subsets of that spidering. A 10 gigabyte subset known as WT10g was used for the main task in the web track [1]. There was also a separate large task in the web track that used the 100 gigabyte VLC2/WT100g collection and a set of 10,000 queries selected from Electronic Monk and AltaVista query logs. See the web track report in these proceedings for more details about the large web task.

Table 2: Number of participants per task and total number of distinct participants in each TREC.

Task	TREC								
	1	2	3	4	5	6	7	8	9
Ad Hoc	18	24	26	23	28	31	42	41	—
Routing	16	25	25	15	16	21	—	—	—
Interactive	—	—	3	11	2	9	8	7	6
Spanish	—	—	4	10	7	—	—	—	—
Confusion	—	—	—	4	5	—	—	—	—
Database Merging	—	—	—	3	3	—	—	—	—
Filtering	—	—	—	4	7	10	12	14	15
Chinese	—	—	—	—	9	12	—	—	—
NLP	—	—	—	—	4	2	—	—	—
Speech	—	—	—	—	—	13	10	10	3
Cross-Language	—	—	—	—	—	13	9	13	16
High Precision	—	—	—	—	—	5	4	—	—
Very Large Corpus	—	—	—	—	—	—	7	6	—
Query	—	—	—	—	—	—	2	5	6
Question Answering	—	—	—	—	—	—	—	20	28
Web	—	—	—	—	—	—	—	17	23
Total participants	22	31	33	36	38	51	56	66	69

The topics used in the main web task were TREC topics 451–500. As described earlier, these topics were created especially for the track. Three-way relevance judgments (not relevant, relevant, and highly relevant) were used. In addition, assessors were asked to select the best document from among all the documents in the pool for each topic. While the official results of the task were scored by conflating the relevant and highly relevant categories, the additional information collected during assessing can be used to develop other evaluation schemes for web retrieval.

Twenty-three groups submitted 105 runs to the main task of the web track. Twelve of the runs used manual query construction techniques, 40 of the runs were automatic runs that used only the original Excite query, and the remaining 53 runs were automatic runs that used some other part of the topic in addition to the Excite query. Runs using the description field of the topic were always more effective than the corresponding run using only the original Excite query. (Remember that the description field corrected the spelling errors of the original query. Seven topics had some sort of error in the original Excite query, five of which were serious. Examples of serious errors are the one-word queries “nativityscenes” and “angioplast7”.)

The order of systems ranked by average effectiveness differed depending on whether the evaluation used both relevant and highly relevant documents or highly relevant documents only. This finding implies that retrieving highly relevant documents is a different task from retrieving generally relevant documents, at least to the extent that different retrieval techniques should be used. More research is needed to determine precisely which techniques work better for which task and why.

The motivation for distinguishing between highly relevant and generally relevant documents is a widespread belief that web users will be better served by systems that retrieve highly relevant documents. Taking this reasoning a step further, some have argued that web search engines should actually be evaluated on their ability to retrieve the very best page. However, the results of the web track demonstrate that using the best page as the single relevant document is too unstable to be a reliable evaluation strategy.

Once the web track assessing was complete, NIST gave the set of relevant documents (both highly relevant and generally relevant) to two additional assessors and asked them to select the best page. In all cases the definition of best was left up to the assessor. There was significant disagreement among the assessors as to the best page for a topic: all three assessors disagreed with one another for 17 of the 50 topics, and of the 13 topics for which all three assessors picked the same page, 5 topics had three or fewer pages in its relevant set. Furthermore, unlike traditional retrieval system evaluation [15], best document evaluation is not robust against changes caused by using different assessors to select best documents. The Kendall tau correlation between system rankings produced by using different assessors averaged only about .75 when using best document evaluation as compared to over .9 for traditional evaluation.

3.2 The Cross-Language (CLIR) track

The CLIR task is an ad hoc retrieval task in which the documents are in one language and the topics are in a different language. The goal of the track is to facilitate research on systems that are able to retrieve relevant documents regardless of the language a document happens to be written in. The TREC-9 cross-language track used Chinese documents and English topics. A Chinese version of the topics was also developed so that cross-language retrieval performance could be compared with the equivalent monolingual performance.

The document set was approximately 250 megabytes of news articles taken from the Hong Kong Commercial Daily, the Hong Kong Daily News, and Takungpao. The documents were made available for use in TREC by Wisers, Ltd. Twenty-five topics were developed by NIST assessors. The assessment pools were created using each group's first choice cross-language run and first-choice monolingual run (if any), using the top 50 documents from each run.

Fifty-two runs from 15 different groups were submitted to the track. Thirteen of the runs were monolingual runs. Only one run (a cross-language run) was a manual run.

The effectiveness of cross-language runs is frequently reported as a percentage of monolingual effectiveness. In the CLIR track, the cross-language run submitted by the BBN group was not only better than their monolingual run (as measured by mean average precision), it was better than all the submitted monolingual runs. While BBN has since produced a monolingual run that is better than the best of their cross-language runs [19], the effectiveness of English to Chinese cross-language retrieval remains high.

3.3 The Spoken Document Retrieval (SDR) track

The SDR track fosters research on retrieval methodologies for spoken documents (i.e., recordings of speech). The task in the track is an ad hoc task in which the documents are transcriptions of audio signals.

The SDR track has run in several TRECs and the track has had the same general structure each year. Participants worked with different versions of transcripts of news broadcasts to judge the effects of errors in the transcripts on retrieval performance. The *reference* transcripts were manually produced and assumed to be perfect. (For TREC-9 the reference transcripts were a combination of human reference transcripts, closed captioning transcripts, and automatically-combined (using NIST's ROVER algorithm) automatic transcripts.) The *baseline* transcripts were produced by one automatic speech recognizer and made available to all participants. There was one TREC-9 baseline transcript, which was the "B2" transcript used in the TREC-8 track. This transcript was produced 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 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.

The document collection used in TREC-9 was the audio portion of the TDT-2 News Corpus as collected by the Linguistic Data Consortium (LDC). This corpus contains 557 hours of audio representing 1,064 news shows, which were segmented into approximately 21,500 documents. Two different versions of 50 new topics (numbers 124–173) were created for the track. The "standard" version of the topics was a one sentence description, while the "terse" version of the topics was just a few words. As in the TREC-8 track, the TREC-9 track focused on the unknown boundary condition, that is, retrieving documents from the audio stream when the system was not given the story boundaries.

Three groups participated in the track, submitting a total of 64 runs. Overall, the retrieval results were excellent: the systems could find relevant passages produced by a variety of recognizers on the full unsegmented news broadcasts, using either the terse or longer standard queries. Indeed, for each of the participants, retrieval from the transcripts created by their own recognizer was comparable to the retrieval from the human reference transcripts.

Table 3: Runsets submitted to the TREC-9 query track.

Participant	Runsets	Description
Hummingbird	hum*	7 variants of SearchServer
Microsoft	ok9u	Okapi run with no query expansion
Sabir Research	Sab*	3 variants of SMART
Sun Microsystems	SUN, SUNl	2 variants of Nova
Univ. of Massachusetts	IN7*	3 variants of INQUERY
Univ. of Melbourne	UoMd, UoMl	2 variants of MG

3.4 The Query track

The task in the query track was an ad hoc task using old TREC document and topic sets. The focus in the track was not on absolute retrieval effectiveness, but on the variability of topic performance. A variety of research (for example, see [2]) has shown that the difference in retrieval effectiveness for a given retrieval system on different topics is much greater on average than the difference in retrieval effectiveness between systems for the same topic. The development of query-specific processing strategies has been hampered as a result because the available topic sets (of size 50 for most TREC collections) are too small to isolate the effects caused by different topics. The query track was designed as a means for creating a large set of different queries for an existing TREC topic set as a first step toward query-specific processing.

Six groups participated in the TREC-9 query track, with each group running each of 43 different querysets using one or more variants of their retrieval system. A *queryset* consists of one query for each of 50 topics (TREC topics 51-100) where each query is from the same category of queries. Three different query categories were used.

1. Short: 2-4 words selected by reading the topic statement.
2. Sentence: a sentence—normally less than one line—developed after reading the topic statement and possibly some relevant documents.
3. Sentence-Rel: a sentence developed after reading a handful of relevant documents. The topic statement was *not* used for this category of query.

Relevant documents from TREC disk 2 were used to construct the queries. Twenty-one of the querysets were developed for the TREC-8 query track, and the remaining 22 querysets were developed for the TREC-9 track.

A *runset* is the result of running one version of a retrieval system on all 43 querysets against the documents on TREC disk 1. Eighteen runsets were submitted, for a total of 774 (18×43) runs submitted to the track. Table 3 gives a short description of each runset, while Table 4 gives the average, minimum, and maximum mean average precision score computed over the 43 different querysets. As can be seen from wide range of scores for each system, the individual query formulations again had a pronounced effect on retrieval performance.

3.5 The Question Answering (QA) track

The purpose of the question answering track was to encourage research into systems that return actual answers, as opposed to ranked lists of documents, in response to a question. Participants received a set of fact-based, short-answer questions and searched a large document set to extract (or construct) an answer to each question. Participants returned a ranked list of five [*document-id*, *answer-string*] pairs per question such that each answer string was believed to contain an answer to the question and the document supported that answer. Answer strings were limited to either 50 bytes or 250 bytes depending on the run type. Each question was guaranteed to have an answer in the collection. An individual question 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

Table 4: Average, minimum, and maximum Mean Average Precision scores for each query track system computed over the 43 different querysets submitted to the track. Each queryset contains queries created for TREC topics 51–100.

	Average	Minimum	Maximum		Average	Minimum	Maximum
IN7a	0.1799	0.1017	0.2534	UoMl	0.1539	0.0812	0.2099
IN7e	0.2288	0.1456	0.3168	hum4	0.1713	0.0907	0.2580
IN7p	0.1848	0.1095	0.2593	humA	0.1741	0.0912	0.2482
SUN	0.0572	0.0062	0.1247	humB	0.1732	0.0908	0.2420
Sunl	0.0677	0.0268	0.1152	humD	0.1771	0.0897	0.2508
Saba	0.1924	0.1089	0.2665	humI	0.1736	0.0910	0.2418
Sabe	0.2516	0.1481	0.3202	humK	0.1713	0.0863	0.2425
Sabm	0.2321	0.1347	0.3057	humV	0.1648	0.0788	0.2453
UoMd	0.1612	0.1010	0.2091	ok9u	0.1917	0.0963	0.2608

contained a correct answer. The score for a submission was then the mean of the individual questions' reciprocal ranks. Question answering systems were given no credit for retrieving multiple (different) correct answers, or for recognizing that they did not know the answer.

There were several changes between the TREC-9 track and the initial running of the track in TREC-8. In TREC-9, the document set was larger, consisting of all the news articles on TREC disks 1–5. The test set of questions was also much larger, consisting of 693 questions rather than 200. A more substantial difference was the way in which the questions were created. Instead of using questions created especially for the track, which tended to be back-formulations of a sentence in some document in the collection, questions were selected from query logs. Some questions were taken from a log of questions that had been submitted to Encarta and were made available to NIST by Microsoft. Other questions were created by NIST staff using the Excite log from which the web track queries were selected for suggestions. In all cases, the questions were created without reference to the document set. Once the questions were created, NIST assessors searched the document set to find which questions had answers in the test document set. Five hundred questions were selected from among the candidate questions that had an answer in the document set. In a separate pass, NIST assessors were given a subset of the questions (but not their answers) and were asked to create equivalent, re-worded questions. For example, the question "How tall is the Empire State building?" might be re-worded as "How high is the Empire State building?", "What is the height of the Empire State building?", "The Empire State Building is how tall?", etc. A total of 193 question variants were added to the set of 500 to make the final question set of 693 questions.

Human assessors read each string and decided whether the answer string contained an answer to the question. If not, the response was judged as incorrect. If so, the assessor decided whether the answer was supported by the document returned with the string. If the answer was not supported by that document, the response was judged as "Not Supported". If it was supported, the response was judged as correct. The official scoring for the track treated Not Supported answers as incorrect.

Twenty-eight groups submitted 78 runs to the track, with 34 runs using the 50-byte-limit and 44 runs using the 250-byte-limit. The best performing system, from Southern Methodist University, was able to extract a correct answer about 65 % of the time by integrating multiple natural language processing techniques with abductive reasoning [6]. While the 65 % score is a slightly worse result than the TREC-8 scores in absolute terms, it represents a very significant improvement in question answering systems. The TREC-9 task was considerably harder than the TREC-8 task because of the switch to "real" questions (which tend to be far more ambiguous than the questions constructed for the TREC-8 task). The SMU system found an answer about a third again as often as the next best system (66 % of the questions vs. 42 % of the questions).

3.6 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 task in the TREC-9 track was a question answering task. Two different types of questions were used:

- find any n Xs (for example, "Name 3 US Senators on committees regulating the nuclear industry.")
- compare two specific Xs (for example, "Do more people graduate with an MBA from Harvard Business School or MIT Sloan?")

Human searchers were given a maximum of 5 minutes to find the answer to a question and support that answer with a set of documents. A total of 8 questions was used in the track.

The document set was the set of documents used in the question answering track (the news articles from TREC disks 1–5). The track defined an experimental framework that specified a minimum number of searchers, the order in which searchers were assigned questions, and the set of data to be collected. This framework did not provide for a comparison of systems across sites, but did allow groups to estimate the effect of their own experimental manipulation free and clear of the main (additive) effects of searcher and topic.

Six groups participated in the interactive track, with some groups performing more than the minimum number of searches. Of the 829 responses submitted to TREC across all topics and groups, 309 (37 %) found no correct answer, suggesting that the 5-minute time limit made the task challenging. The percentage of no answer found was roughly the same for the two types of questions (36 % for the find any n questions and 39 % for the comparison questions).

3.7 The Filtering track

The filtering task is to retrieve just those documents in a document 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 (for TREC-9, the system also received a few (< 5) relevant documents), 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. Two other, simpler tasks were also part of the track. In the *batch* filtering task, the system is given a topic and a (relatively large) 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.

The document set for the TREC-9 filtering task was the OHSUMED test collection [8]. The test documents were the documents from 1988–1991, while a set of documents from 1987 were available for training (if the particular task allowed training). There were three topic sets: the queries from the OHSUMED test collection, a set of almost 5000 MeSH headings that were treated as topic statements, and a subset of 500 MeSH headings. Since the track used an existing test collection, no relevance judgments were made at NIST for the track.

Research into appropriate evaluation methods for filtering runs (which do not produce a ranked list and therefore cannot be evaluated by the usual IR evaluation measures) has been a major thrust of the filtering track. The earliest filtering tracks used linear utility functions as the evaluation metric. With a linear utility function, a system is rewarded some number of points for retrieving a relevant document and penalized a different number of points for retrieving an irrelevant document. Utility functions are attractive because they directly reflect the experience of a user of the filtering system. Unfortunately, there are drawbacks to the functions as evaluation measures. Utilities do not average well because the best possible score for each topic is a function of the number of relevant documents for that topic, and the worst possible score is essentially unbounded. Thus topics that have many relevant documents will dominate an average, and a single poorly performing topic can eclipse all other topics. Furthermore, it is difficult to know how to set the relative worth of relevant and irrelevant documents. For example, one of the utility functions used in the TREC-8 track rewarded systems with three points for retrieving a relevant document and penalized systems two points for retrieving an irrelevant document. This actually defines a very difficult filtering task: using this utility function, the average behavior of each of the TREC-8 filtering systems was worse than the baseline of retrieving no documents at all. The TREC-9 track used a bounded utility function as one measure and introduced a new "precision-oriented" measure, T9P. The bounded utility function rewarded systems two points for a relevant document and penalized systems one point for an irrelevant document, and in addition set a limit on the worst possible score a topic could receive. The idea of the precision-oriented measure was to penalize systems for not retrieving a sufficient number of documents, which for TREC-9 was 50 documents. Thus,

$$T9P = \frac{\text{number of relevant retrieved}}{\max(\text{number retrieved}, 50)}$$

Seventy-five runs from 14 different groups were submitted to the filtering track. Forty-one of the runs were adaptive filtering runs, 19 runs were batch filtering runs, and 15 runs were routing runs. Unlike TREC-8, several TREC-9 adaptive filtering systems obtained good average utilities while retrieving an adequate number of documents. In addition, the adaptive filtering scores were relatively close to the scores for the much easier routing task, as demonstrated by comparing the scores from the new T9P measure to the average of precision at 50 documents retrieved for the routing runs.

4 The Future

The next TREC, TREC 2001, will see a few changes in the tracks that are offered. The spoken document track has met its goals and will therefore be discontinued. A new track that will focus on content-based access to digital video will continue TREC's interest in multimedia retrieval. The query track will cease to be a track and evolve into a "station." That is, the TREC web site will serve as a repository for both query statements and retrieval results produced from those queries. This will allow the research of the track to continue without the time pressures of the TREC deadlines.

The remaining tracks will continue, though the specific task involved in the track will likely change. The web track will include so-called navigational topics [14] in addition to the informational topics TREC has traditionally used. The cross-language track will focus on retrieving Arabic documents using English, French, or Arabic topics. In addition to the fact-based, short-answer questions used in the first two years of the question answering track, the 2001 track will feature a pilot study using questions that require information from multiple documents to be combined to form a correct answer. The task in the interactive track will most likely involve observing subjects using the live web to accomplish a specific task. The observations made in the track will be used to inform the definition of a task with a metrics-based evaluation plan for the following year's track, as suggested by the SIGIR Workshop on Interactive Retrieval at TREC and Beyond [7] held after SIGIR 2000. The filtering track will continue to focus on adaptive filtering using a new collection of data released by Reuters (see `http://about.reuters.com/researchandstandards/corpus/`).

References

[1] Peter Bailey, Nick Craswell, and David Hawking. Engineering a multi-purpose test collection for web retrieval experiments. *Information Processing and Management*. To appear.

[2] David Banks, Paul Over, and Nien-Fan Zhang. Blind men and elephants: Six approaches to TREC data. *Information Retrieval*, 1:7–34, 1999.

[3] Nicholas J. Belkin and W. Bruce Croft. Information filtering and information retrieval: Two sides of the same coin? *Communications of the ACM*, 35(12):29–38, December 1992.

[4] Chris Buckley. trec_eval IR evaluation package. Available from `ftp://ftp.cs.cornell.edu/pub/smart`.

[5] C. W. Cleverdon, J. Mills, and E. M. Keen. Factors determining the performance of indexing systems. Two volumes, Cranfield, England, 1968.

[6] S. Harabagiu, D. Moldovan, M. Pasca, R. Mihalcea, M. Surdeanu, R. Bunescu, R. Gîrju, V. Rus, and P. Morarescu. FALCON: Boosting knowledge for answer engines. In Voorhees and Harman [18].

[7] William Hersh and Paul Over. SIGIR workshop on interactive retrieval at TREC and beyond. *SIGIR Forum*, 34(1):24–27, Spring 2000.

[8] W.R. Hersh, C. Buckley, T.J. Leone, and D.H. Hickam. OHSUMED: An interactive retrieval evaluation and new large test collection for research. In W. Bruce Croft and C.J. van Rijsbergen, editors, *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 192–201, 1994.

[9] 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.

[10] G. Salton, editor. *The SMART Retrieval System: Experiments in Automatic Document Processing.* Prentice-Hall, Inc. Englewood Cliffs, New Jersey, 1971.

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

[12] 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.

[13] Karen Sparck Jones. *Information Retrieval Experiment.* Butterworths, London, 1981.

[14] Bob Travis and Andrei Broder. The need behind the query: Web search vs classic information retrieval. `http://www.infonortics.com/searchengines/sh01/slides-01/sh01pro.html`.

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

[16] Ellen M. Voorhees and Donna Harman. Overview of the eighth Text REtrieval Conference (TREC-8). In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, pages 1–24, 2000. NIST Special Publication 500-246. Electronic version available at `http://trec.nist.gov/pubs.html`.

[17] Ellen M. Voorhees and Donna Harman. Overview of the sixth Text REtrieval Conference (TREC-6). *Information Processing and Management*, 36(1):3–35, January 2000.

[18] E.M. Voorhees and D.K. Harman, editors. *Proceedings of the Ninth Text REtreival Conference (TREC-9)*, 2001.

[19] J. Xu and R. Weischedel. TREC-9 cross-lingual retrieval at BBN. In Voorhees and Harman [18].

[20] Justin Zobel. How reliable are the results of large-scale information retrieval experiments? 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 307–314, Melbourne, Australia, August 1998. ACM Press, New York.