Overview of TREC 2001

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

1 Introduction

The tenth Text REtrieval Conference, TREC 2001, was held at the National Institute of Standards and Technology (NIST) November 13–16, 2001. The conference was co-sponsored by NIST, the Information Technology Office of the Defense Advanced Research Projects Agency (DARPA/ITO), and the US Department of Defense Advanced Research and Development Activity (ARDA).

TREC 2001 is the latest in a series of workshops designed to foster research on technologies for information retrieval. The workshop series has four goals:

- to encourage retrieval research 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.

TREC 2001 contained six areas of focus called "tracks". These included the Cross-Language Retrieval Track, the Filtering Track, the Interactive Retrieval Track, the Question Answering Track, the Video Retrieval Track, and the Web Retrieval Track. This was the first year for the video track, which was designed to investigate content-based retrieval of digital video. The other tracks were run in previous TRECs, though the particular tasks performed in some of the tracks changed for TREC 2001.

Table 1 lists the groups that participated in TREC 2001. Eighty-seven groups submitted retrieval results, an increase of approximately 25 % over the previous year. The participating groups come from twenty-one different countries and include academic, commercial, and government institutions.

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 Information Retrieval

Information retrieval is concerned with locating information that will satisfy a user's information need. Traditionally, the emphasis has been on text retrieval: providing access to natural language texts where the set of documents to be searched is large and topically diverse. There is increasing interest, however, in finding appropriate information regardless of the medium that happens to contain that information. Thus "document" can be interpreted as any unit of information such as a web page or a video clip.

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 the particular topic that will be investigated. We call this an *ad hoc* retrieval task, reflecting the arbitrary

Table 1: Organizations participating in TREC 2001

Ajou University
Alicante University
BBN Technologies
Carnegie Mellon U. (3 groups)
Chinese Academy of Sciences
Clairvoyance Corp.
CLIPS-IMAG
CL Research
Conexor Oy
CSIRO
Dublin City University
EC Wise, Inc.
Fondazione Ugo Bordoni
Fudan University
Fujitsu
Harbin Institute of Technology
Hummingbird
IBM-Almaden
IBM-Haifa
IBM-T.J. Watson (3 groups)
Illinois Institute of Technology
Imperial College of Science, Tech. & Medicine
InsightSoft-M
IRIT/SIG
ITC-irst
Johns Hopkins University, APL
Justsystems Corp.
KAIST
Kasetsart University
Katholieke Universiteit Nijmegen
KCSL
Kent Ridge Digital Labs
Korea University
Language Computer Corp.
David Lewis
LIMSI
Microsoft Research China
Microsoft Research Ltd.
MITRE
Moscow Medical Academy
National Taiwan University
New Mexico State University
NexTrieve
NTT Communication Science Labs
Oracle
Oregon Health and Science University
Pohang University of Science and Technology
Queens College, CUNY
RICOH
Rutgers University (2 groups)
SER Technology Deutschland GmbH
Sun Microsystems Labs
Syracuse University
Tampere University of Technology
Tilburg University
University of Twente
TNO-TPD & Universite de Montreal
University of Alberta
University of Amsterdam/ILLC
U. of Amsterdam & CWI & TNO & U. Twente
University of California, Berkeley
University of Glasgow
University of Illinois, Urbana/Champaign
University of Iowa
University of Maryland (2 groups)
University of Massachusetts
University of Michigan
University of Neuchatel
University of North Carolina, Chapel Hill (2 groups)
University of North Texas
University of Padova
University of Pennsylvania
University of Pisa
University of Sheffield
University of Southern California, ISI
University of Toronto
University of Waterloo
University of York
Virginia Tech
Yonsei University

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.

A *known-item search* is similar to an ad hoc search but the target of the search is a particular document (or a small set of documents) that the searcher knows to exist in the collection and wants to find again. Once again, the retrieval system's response is usually a ranked list of documents, and the system is evaluated by the rank at which the target document is retrieved.

In a document routing or *filtering* task, the topic of interest is known and stable, but the document collection is constantly changing [1]. 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.

Information retrieval has traditionally focused on returning entire 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 wade through a list of documents looking for the specific answer. To encourage research on systems that return answers instead of document lists, TREC has had a question answering track since 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 [3, 6, 9], 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 primary TREC test collections contain 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 primary 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 ad hoc task of the 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 has been used differently in topics 451–550, the web track's ad hoc 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.

```
<num> Number:  508
<title> hair loss is a symptom of what diseases

<desc> Description:
Find diseases for which hair loss is a symptom.

<narr> Narrative:
A document is relevant if it positively connects the loss of head hair in humans with a
specific disease.  In this context, "thinning hair" and "hair loss" are synonymous.  Loss
of body and/or facial hair is irrelevant, as is hair loss caused by drug therapy.
```

Figure 1: A sample TREC 2001 topic from the web track.

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 tweaked to create the topics for the ad hoc task in the web track. Participants in the web track 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. This year's actual web queries were taken from a MSNSearch log that NIST obtained from Sue Dumais of Microsoft. 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 (TREC-9 web topics) contains the literal query that was the seed of the topic. For this year's topics 501-550, NIST corrected the spelling of the words in the MSNSearch queries, but otherwise left the queries as they were submitted, leaving other errors such as punctuation or grammatical mistakes in the title fields. The description and narrative fields use correct (American) English.

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 topic 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 [7]. 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 [11].

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, TREC uses a technique called pooling [8] 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 × *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 [14]. He also found that the TREC collections were not biased against unjudged runs. In this test, he evaluated each run that contributed to the pools using both the official set of relevant documents published for that collection and the set of relevant documents produced by removing the relevant documents uniquely retrieved by the run being evaluated. For the TREC-5 ad hoc collection, he found that using the unique relevant documents increased a run's 11 point average precision score by an average of 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 [13]. 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 [5].

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 [11].

2.2 Evaluation

Retrieval runs on a test collection can be evaluated in a number of ways. In TREC, all ad hoc tasks are evaluated using the `trec_eval` package written by Chris Buckley of Sabir Research [2]. 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.

Only two of the tasks in TREC 2001, the ad hoc task in the web track and the task in the cross-language track, were tasks that can be evaluated with `trec_eval`. The remaining tasks used other evaluation measures that are described in detail in the track overview paper for that task, and are briefly described in Appendix A. The bulk of Appendix A consists of the evaluation output for each run submitted to TREC 2001.

3 TREC 2001 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 tracks that were in each TREC, the number of groups that submitted runs to that track, 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 2001 tracks. See the track reports elsewhere in this proceedings for a more complete description of each track.

3.1 The Cross-Language (CLIR) track

The task in the CLIR track 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

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

	TREC									
Track	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
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	6
Spanish	—	—	4	10	7	—	—	—	—	—
Confusion	—	—	—	4	5	—	—	—	—	—
Database Merging	—	—	—	3	3	—	—	—	—	—
Filtering	—	—	—	4	7	10	12	14	15	19
Chinese	—	—	—	—	9	12	—	—	—	—
NLP	—	—	—	—	4	2	—	—	—	—
Speech	—	—	—	—	—	13	10	10	3	—
Cross-Language	—	—	—	—	—	13	9	13	16	10
High Precision	—	—	—	—	—	5	4	—	—	—
Very Large Corpus	—	—	—	—	—	—	7	6	—	—
Query	—	—	—	—	—	—	2	5	6	—
Question Answering	—	—	—	—	—	—	—	20	28	36
Web	—	—	—	—	—	—	—	17	23	30
Video	—	—	—	—	—	—	—	—	—	12
Total participants	22	31	33	36	38	51	56	66	69	87

able to retrieve relevant documents regardless of the language a document happens to be written in. The TREC 2001 cross-language track used Arabic documents and English or French topics. An Arabic 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 created and released by the Linguistic Data Consortium ("Arabic Newswire Part 1", catalog number LDC2001T55). It consists of 869 megabytes of news articles taken from the Agence France Presse (AFP) Arabic newswire: 383,872 articles dated from May 13, 1994 through December 20, 2000.

Twenty-five topics were created for the track using the standard topic development protocol except that topic development took place at the Linguistic Data Consortium (LDC). The assessors were fluent in both Arabic and English (for most assessors Arabic was their first language). They searched the document collection using a retrieval system developed by the LDC for the task and Arabic as the query language. Once twenty-five topics were selected from among the candidate topics, the assessor who developed the topic created the full topic statement first in English and then in Arabic. The assessors' instructions were that the Arabic version of the topic should contain the same information as the English version, but should be expressed in a way that would seem natural to a native speaker of Arabic. The entire set of 25 English topics was also translated into French by Sylvain Soliman of the Délégation Générale pour l'Armement. The three different versions of the topics were then made available to the track participants who were asked to check the topics for substantive differences among the different versions. A few changes were suggested by participants, and those changes were made to produce the final version of the topics.

Forty-eight runs from ten different groups were submitted to the track. Twenty-eight of the runs were cross-language runs, including three runs that used French as the topic language. One monolingual run and one cross-language run were manual runs.

The assessment pools were created using three runs from each group (based on the group's assigned priorities) and using the top 70 documents from each run. The average size of the pools was 910 documents. Fourteen monolingual runs and sixteen cross-language runs were added to the pools. None of the runs whose topic language was French was judged. The LDC assessors judged each document in the pools using binary (relevant/not relevant) assessments.

The average number of relevant documents over the 25 topics is 164.9 with five topics having more than

300 relevant documents. The combination of fairly broad topics and ten participating groups resulted in pools that were not as complete as they ideally would be: for seven topics, more than half of the relevant documents were retrieved by only one group, and for another six topics between 40 and 50 % of the relevant documents were retrieved by only one group. The test whereby a group's unique relevant documents are removed from the relevance judgments shows that mean average precision scores decrease by an average of 8 %, with a maximum difference of 21 %. These differences are a little larger than those that have been reported for other TREC cross-language collections [13]. They suggest that experimenters who find many unjudged documents in the top-ranked list of only one of a pair of runs to be contrasted should proceed with care.

While the effectiveness of cross-language runs is traditionally reported as a percentage of monolingual effectiveness, the most effective run submitted to the track was a a cross-language run, `BBN10XLB` submitted by BBN. The difference between the effectiveness of `BBN10XLB` and the corresponding monolingual run, `BBN10MON`, is small (mean average precision scores of .4639 and .4537 respectively), but this is the second year that a cross-language run was better than all submitted monolingual runs. In the TREC-9 cross-language track the task was retrieving Chinese documents with either English or Chinese topics. Several groups, including BBN, submitted cross-language runs that were more effective than their monolingual counterparts.

3.2 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 topic. There were three tasks in the TREC 2001 filtering track, an adaptive filtering task, a batch filtering task, and a routing task. The main focus of the track was the adaptive filtering task.

In the adaptive filtering task, a system starts with a profile derived from the topic statement and a small number of examples of relevant documents, and processes documents one at a time in date order. For each document in turn, the system must make a binary decision whether to retrieve it. If the system decides to retrieve the document, it obtains the relevance judgment for that document, and can modify its profile based on the judgment if desired. The final output is the unranked set of retrieved documents for the topic.

The batch filtering task is a simpler version of the adaptive task. In this task, the system is given a topic and a (relatively large) set of training documents such that each document in the training set is labeled as relevant or not relevant. From this data, the system creates a profile and a rule for when a document should be retrieved. The rule is applied to each document in the test set of documents without further modification. Once again, the final output is an unranked set of retrieved documents.

In the *routing* task, the system again builds a profile or query from a topic statement and a training set of 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 final output for the routing task is a list of 1000 documents ranked by decreasing similarity to the query.

The TREC 2001 filtering task used the recently released "Reuters Corpus, Volume 1, English language, 1996-08-20 to 1997-08-19" collection from Reuters (`http://about.reuters.com/researchandstandards/corpus/`) as training and test data. This collection consists of approximately 810,000 news stories from August 20, 1996 through August 19, 1997. Each document is tagged with Reuters category codes. A hierarchy of the Reuters category codes is included with the corpus.

Each topic for the filtering track was a category code. The topic statement included both the category identifier (so systems could use the hierarchical information about categories if desired) and the name of category. Eighty-four topics were created for the track by selecting category codes that were assigned to at least 2 training documents, but no more than 5 % of the training documents. The training documents were the set of documents from from August, 1996.

A document was considered to be relevant to a topic if the document was assigned the topic's category code. Since all this data is part of the Reuters corpus, 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.

Two different measures were used as the primary measures for the TREC 2001 filtering tasks. The first was a linear utility function that rewarded systems two points for retrieving a relevant document and penalized systems one point for retrieving an irrelevant document. To compute average utility over the 84 topics, the utility score for each topic was first scaled between an upper and lower bound (2 × number-relevant and -100, respectively). The second measure was a version of van Rijsbergen's F measure [10]. This measure is a function of recall and precision plus a variable, β, that controls the relative importance of precision over recall. For the filtering track, β was set to .5, which emphasizes precision. If R^+ is the number of relevant documents that were retrieved, R^- the number of relevant documents that were not retrieved, and N^+ the number of non-relevant documents that were retrieved, the F score used in the track is defined as

$$\text{T10F} = \begin{cases} 0 & \text{if } R^+ = N^+ = 0 \\ \dfrac{1.25R^+}{.25R^- + N^+ + 1.25R^+} & \text{otherwise} \end{cases}$$

Routing runs were evaluated using mean average precision since routing runs produce a ranked list of documents.

Sixty-six runs from nineteen different groups were submitted to the filtering track. Of these, 30 runs were adaptive filtering runs, 18 were batch filtering runs, and 18 were routing runs.

Because of the way the topics and relevance judgments were defined, the topics used in this year's filtering track had a much larger average number of relevant documents than the collections that have been used in previous years' tracks. Indeed, some topics had considerably more than 1000 relevant documents (so routing results for those topics are likely not meaningful since routing submissions were limited to a maximum of 1000 documents per topic). When there are relatively few relevant documents, retrieving very few documents can produce a good utility score. For this year's task, however, retrieving few documents produced a poor score. Retrieving no documents produced an average scaled utility score of 0.03. Many of the adaptive filtering submissions had a scaled utility greater than 0.2, with the best adaptive filtering run, `oraAU082201` submitted by Oracle, obtaining a scaled utility score of 0.29.

3.3 The Interactive track

The interactive track was one of the first tracks to be introduced into TREC. Since its inception, the high-level goal of the track has been the investigation of searching as an interactive task by examining the process as well as the outcome.

The TREC 2001 track was the first year of a two-year plan to implement a metrics-based comparison of interactive systems as suggested by the SIGIR 2000 Workshop on Interactive Retrieval at TREC and Beyond [4]. During this first year of the plan, TREC 2001 participants performed observational studies of subjects using publicly-accessible tools and the live web to accomplish a search task. Participants devised their own objectives for the studies, though a common objective for all the studies was to suggest a testable hypothesis for use in TREC 2002. Each searcher who participated in a study performed four searches, two from a list of fully specified search problems and two from a list of partially specified search problems. The lists of search problems were defined by the track and came from four domains:

- finding consumer medical information,
- buying an item,

Table 3: Focus of particular study done in the interactive track by each participant

CSIRO:	the correlation between searching/presentation mechanisms and search tasks
Glasgow:	the extent to which implicit evidence of relevance can be substituted for explicit evidence
OHSU:	how subjects use the Web
Rutgers:	ways to obtain longer queries (line vs. box)
U. Michigan:	the effect of domain knowledge in search effectiveness
U. Toronto:	the use of categories vs. standard search statements

- travel planning, and
- collecting material for a project on a given subject.

For example, a fully specified search problem for the travel planning domain was "Identify three interesting things to do during the weekend in Kyoto, Japan." A partially specified search problem for the same domain was "Identify three interesting places to visit in ______." Track participants were required to collect demographic information about the searchers, to collect the URLs of all pages visited during all searches, and to collect whatever other data made sense for the aims of their study.

Six groups participated in the interactive track. The main focus of each group's study is given in Table 3.

3.4 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. The TREC 2001 track contained three different tasks: the main task, the list task, and the context task. All of the tasks required completely automatic processing.

The main task was the focus of the track, and was essentially the same as the task in the TREC-8 and TREC-9 QA tracks. 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 no more than 50 bytes. Unlike previous years, questions were not guaranteed to have an answer in the collection. The system could return "NIL" as one of the five choices to indicate its belief that the collection did not contain an answer to the question.

As an additional, experimental part of the main task, systems were also required to return a "final answer" for each question. The final answer was either an integer from one to five that referred to a position in the ranked list of responses for that question, or the string "UNSURE" that indicated the system did not know what the answer was.

The main task was evaluated using the mean reciprocal rank measure that was used in previous years. An individual question received a score equal to the reciprocal of the rank at which the first correct response was returned, or zero if none of the five responses contained a correct answer. The score for a run was then the mean of the individual questions' reciprocal ranks. The correctness of a response was determined by human assessors. The assessors read each response 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. The NIL response was scored the same as other responses (except that it could never be marked Not Supported). NIL was counted as correct when no correct answer was known to exist in the collection for that question.

The document collection used for all the tasks in the QA track was the set of newspaper and newswire articles on TREC disks 1–5. The questions for the main task continued a progression of using more realistic questions in each of the three runnings of the track. In TREC-8, the majority of the questions were created expressly for the track, and thus tended to be back-formulations of a statement in a document. In TREC-9, the questions were selected from an Encarta log that contained actual questions, and a raw Excite log. Since the raw Excite log did not contain many grammatically well-formed questions, NIST staff used the Excite log as a source of ideas for actual questions. All the questions were created without looking at any documents. The resulting test set of question was much more difficult than the TREC-8 set, mainly because the TREC-9 set contained many more high-level questions such as *Who is Colin Powell?*. For this year's main task, the source of questions was a set of filtered MSNSearch logs and AskJeeves logs. Raw logs were automatically filtered (at Microsoft and AskJeeves) to select queries that contained a question word (e.g., what, when, where, which, etc.) anywhere in the query; that began with modals or the verb to be (e.g., are, can, could, define, describe, does, do, etc.); or that ended with a question mark. NIST did additional human filtering on these logs, removing queries that were not in fact questions; questions that asked for a list of items; procedural questions; questions that asked for the location of something on the web (e.g., pictures of someone); yes/no questions; and questions that were obviously too current for the document collection (e.g., questions about Britney Spears, etc.). The assessors then searched the collection looking for answers for the queries that remained. NIST fixed the spelling, punctuation, and sometimes the grammar of the queries selected to be in the final question set, but except for a very few (less than 10) questions, the content of the question was precisely what was in the log. The few changes that were made were simple changes such as substituting one Greek god for another so that the question would have an answer in the collection.

The final question set for the main task consisted of 500 questions. The question set contained many more definitional questions (e.g., *What are steroids?*) than in previous years, reflecting the content of the logs. Forty-nine of the questions have no known correct answer in the document collection.

The list task required systems to assemble a set of answers as the response for a question. Each question asked for a given number of instances of a certain type. For example, one of the questions used in the track was *Name 8 Chuck Berry songs.* The response to a list question was an unordered list of [*document-id*, *answer-string*] pairs, where each pair was treated as a single instance. For the Chuck Berry question, each of the answer-strings were to contain the title of a different Chuck Berry song.

The questions were constructed by NIST assessors. The target number of instances to retrieve was selected such that the document collection contained more than the requested number of instances, but more than one document was required to meet the target. A single document could contain multiple instances, and the same instance might be repeated in multiple documents.

The assessors judged each list as a unit. A judgment of Correct/Incorrect/Not Supported was made for each individual pair in the list using the same criteria as in the main task. While judging for correctness, the assessor also marked a set of responses as distinct. The assessor arbitrarily chose any one of a set of equivalent responses to mark as the distinct one, and marked the remainder as not distinct. Incorrect responses were always marked as not distinct, but Not Supported responses could be marked distinct. The accuracy score for a list question was computed as the number of distinct instances retrieved divided by the number of requested instances. The score for a run was the average accuracy over the 25 questions.

The context task was intended to test the systems' ability to track discourse objects (the context) through a series of questions. The expected answer types for questions in the context task were the same as in the main task. However, the questions were grouped into different series, and the QA system was expected to track the discourse objects across the individual questions of a series. For example, the following three questions were one series in the test set:

1. In what country was the first telescope to use adaptive optics used?
2. Where in the country is it located?
3. What is the name of the large observatory located there?

To answer the second question, the system needs to resolve "it" to "the first telescope to use adaptive optics", and "the country" to the country that was the answer to the first question. Similarly, "there" in the third question needs to be resolved to the answer of the second question.

NIST staff created ten question series for the context task. Most series contained three or four questions, though one series contained nine questions. There were 42 questions across all series, and each question was guaranteed to have an answer in the document collection. The context task questions were judged and evaluated as in the main task; all questions were judged by the same assessor.

A total of 92 runs was submitted to the QA track, including submissions from 36 different groups. Each group submitted at least one main task run for a total of 67 main task runs. Ten groups groups submitted eighteen runs for the list task, and six groups submitted seven runs for the context task.

The main task systems can be divided into two broad categories: systems that attempt a full understanding of the question, and systems that use a more shallow data-driven approach. The data-driven approaches rely on simpler pattern matching methods using very large corpora (frequently the web) rather than sophisticated language processing. Both approaches were successful in this year's track, with both approaches equally represented in the top systems.

3.5 The Video track

TREC 2001 was the first year for the video track, a track designed to promote progress in content-based retrieval from digital video. As befits the first running of a track, the video track included a broad range of tasks and search strategies so the track could document the current state-of-the-art in video processing.

The video collection used in the track consisted of approximately eleven hours of MPEG-1 recordings. It contains video from "NIST Digital Video Collection, Volume 1" (`http://www.nist.gov/srd/nistsd26.htm`), from the Open Video Project (`http://www.open-video.org/`), and stockshots from the BBC. Much of the video includes a sound track, though the amount and quality of the audio varies. A transcript of the soundtrack was available for the portion of the collection drawn from the NIST Digital Video Collection, and there were keyword descriptions of the BBC stockshots. The collection consists mostly of documentaries, covering a variety of subjects. The collection also contains a variety of different production techniques.

The track included three different tasks, a shot boundary task, a known-item search task, and an ad hoc search task. The two search tasks could use either automatic or manual processing, while the shot boundary task was restricted to automatic processing.

The goal in the shot boundary task was to (automatically) identify the shot boundaries in a given video clip. Since the retrieval objects in the search tasks were shots, shot boundary detection is a fundamental component of the other tasks. It is also the video processing task that has received the most attention, so it made a good basic task for the track. The shot boundary task was performed on a 5 hour subset of the whole collection. System output was evaluated using automatic comparison to a set of reference shot boundaries created manually at NIST.

Topics for the two search tasks were contributed by the participants. The final set contained 74 topics, which was the union of the topics contributed from all participants. A topic statement included a text description of the information wanted (e.g., "scenes that depict the lift off of the Space Shuttle") and possibly some examples (either video, image, or audio, as appropriate) of that type of information. Each topic was also tagged as to whether it was intended for interactive (manual) processing, automatic processing, and/or a known-item search.

For the known item search task, participants could submit up to 100 shots maximum per topic. System results were automatically compared to the list of shots identified during topic development to determine which shots in the system output were correct. (Since there is no official reference set of shot boundaries, matching system responses to known items is a fuzzy matching process. See the video track overview for details about the matching process.) Submissions were evaluated using recall and precision over the retrieved set of shots.

For the ad hoc search task, participants could submit up to 20 shots maximum per topic. NIST assessors reviewed each submitted clip and made a binary judgment as to whether it satisfied the topic. Submissions were evaluated using the precision of the retrieved set.

Twelve groups participated in the video track, submitting a total of 36 runs. Fifteen of the runs were boundary shot runs, and the remaining 21 runs contained search results (known-item search results, ad hoc search results, or both types of search results). The boundary shot results showed that the systems participating in TREC are almost perfect at recognizing boundaries that result from cuts, while other more gradual changes are more difficult to recognize. The search tasks, particularly the known-item task, are challenging problems for automatic systems.

3.6 The Web track

The goal in the web track is to investigate retrieval behavior when the collection to be searched is a large hyperlinked structure such as the World Wide Web. The track contained two tasks in 2001, the "ad hoc" task and the homepage finding task. The ad hoc task was a traditional ad hoc retrieval task where documents (web pages) were ranked by decreasing likelihood of meeting the information need provided in the topic statement. The homepage finding task was a known-item task where the goal was to find the homepage (or site entry page) of the site described in the topic. The homepage finding task was introduced this year to allow exploration of a retrieval task for which link-based methods are known to be beneficial.

The document collection used for both tasks was the WT10g collection (`http://www.ted.cmis.csiro.au/TRECWeb/access_to_data.html`) used in previous web tracks. This collection is a ten gigabyte subset of a snapshot of the web in 1997. The original snapshot of the web was provided by the Internet Archive. The WT10g collection is a sample of the snapshot selected in such a way as to balance a number of competing desiderata including final size, containing some content-heavy pages, having naturally defined sub-collections contained within the collection, and having a good closed set of hyperlinks (see `http://www.ted.cmis.csiro.au/TRECWeb/wt10ginfo.ps.gz`).

The topics used in the main web task were TREC topics 501–550. As described earlier, these topics were created by NIST assessors by retrofitting topic statements around MSNSearch queries. Three-way relevance judgments (not relevant, relevant, and highly relevant) were used again this year since the results of last year's track showed that the relative quality of runs differed depending on whether evaluation was based on all relevant documents are highly relevant documents only [12]. The evaluation results reported in Appendix A of the proceedings are computed using all relevant documents. Evaluation of the ad hoc task runs is based on `trec_eval`.

The topics for the homepage finding task consisted of a single phrase such as "HKUST Computer Science Dept.". For each topic, the system was to return the home page of the entity described by the topic. For the example topic, the system should retrieve the home page for the computer science department of the Hong Kong University of Science and Technology. The home page of a related site of a different granularity, such as the home page for the entire university or a project within the computer science department was counted as incorrect for this task.

NIST assessors developed 145 topics for the homepage finding task in the following way. An assessor was given a list of 100 randomly selected pages from the WT10g collection. For each page, the assessor followed links to get to a page that he considered to be the home page of a site that contained the original page. Randomly selected pages that did not contain links or that contained links that could not be followed were skipped. Once at a homepage, the assessor created a descriptive phrase for it such that he could imagine someone using that phrase as a query. The assessors were instructed that the phrase should be short, but descriptive enough to distinguish a single site (i.e. "cs dept" alone is realistic, but is not descriptive enough for this task).

Participants returned a ranked list of 100 documents per topic. Small pools consisting of the top twelve pages from each judged run were created to check for pages that had different DOCNOs but were equivalent homepages (caused by mirroring and the like). The rank of the homepage whose rank was closest to one was used as the score for each topic. That is, if a topic had two homepages and a system retrieved the pages at ranks three and seven, then only the homepage at rank three was considered in the scoring. The main evaluation measure for the homepage finding task was the mean reciprocal rank of the homepage.

Thirty groups submitted a total 140 runs to the web track. Of those runs, 97 were ad hoc task runs and 43 were homepage finding task runs. The ad hoc task guidelines specified that at least one of the ad hoc task runs from a group should be automatic runs that used only the title portion of the topic statement ("short" runs) since such runs are the most realistic for the web environment. Seventy of the ad hoc task runs were short runs. The remaining twenty runs included two manual runs and automatic runs that used other parts of the topic statement. Since a large majority of the ad hoc runs were short runs, documents that were retrieved only by shorts runs made up 65 % of the judging pools and 26 % of the relevant documents.

Documents that were retrieved by both a short run and some other type of run made up 20 % of the pools and 61 % of the relevant documents. The two groups that did a manual run were the two groups that found the greatest numbers of unique relevant documents. Each of the two manual runs had a decrease of about 3.5 % in mean average precision when evaluated with and without its group's unique relevant documents. No automatic run had a difference greater than 2.5 % (provided the mean average precision score was at least 0.1).

The retrieval results from the track support the hypothesis that the two tasks within the track require different retrieval techniques. For the ad hoc task, content-based methods alone were sufficient for effective retrieval. The homepage finding task, however, required exploiting additional information, specifically URL text or link structure. The highest-ranked content-only homepage run had a mean reciprocal rank score that was only 30 % as good as the best homepage run: 0.774 for `tnout10epCAU` vs. 0.338 for `tnout10epC`, both submitted by the TNO/University of Twente group.

4 The Future

The final session of each TREC workshop is a planning session for future TRECs, in particular to decide on the set of tracks for the next TREC. Each of the TREC 2001 tracks will continue in TREC 2002. In addition, a new "novelty" track will also be included. The goal in the novelty track is to test systems' abilities to recognize repeated information in a document set. Participants in the track will receive a set of topics and a relatively small (less than 30), ranked set of relevant documents for each topic. Systems must process each document in the order given and flag sentences that contain relevant information that is not contained in previous documents. Evaluation of system performance will be a function of the overlap between the sentences flagged by the system and the sentences selected by human assessors.

TREC 2002 will also contain an exploratory effort or "pre-track" on the retrieval of genomic data. The pre-track will take a very broad definition of genomic data, including such things as research papers, lab reports, etc., as well as actual gene sequences. The purpose of the track is to foster collaboration between the retrieval and bioinformatics communities, as well as to provide a retrieval task on a particular kind of structured data.

Acknowledgements

Special thanks to the track coordinators who make the variety of different tasks addressed in TREC possible.

References

[1] 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.

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

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

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

[5] 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.

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

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

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

[10] C.J. van Rijsbergen. *Information Retrieval*, chapter 7. Butterworths, 2 edition, 1979.

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

[12] Ellen M. Voorhees. Evaluation by highly relevant documents. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74–82, 2001.

[13] 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`.

[14] 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.