Overview of TREC 2002

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

1 Introduction

The eleventh Text REtrieval Conference, TREC 2002, was held at the National Institute of Standards and Technology (NIST) November 19–22, 2002. The conference was co-sponsored by NIST, the Information Awareness Office of the Defense Advanced Research Projects Agency (DARPA/IAO), and the US Department of Defense Advanced Research and Development Activity (ARDA).

TREC 2002 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 2002 contained seven areas of focus called "tracks". These included the Cross-Language Retrieval Track, the Filtering Track, the Interactive Retrieval Track, the Novelty Track, the Question Answering Track, the Video Retrieval Track, and the Web Retrieval Track. This was the first year for the novelty track, which fostered research into detecting redundant information within a relevant document set. The other tracks were run in previous TRECs, though the particular tasks performed in some of the tracks changed for TREC 2002.

Table 1 lists the 93 groups that participated in TREC 2002. The participating groups come from 21 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 2002

Ajou University	Prous Science
Alicante University	Queens College, CUNY
BBN Technologies	Queensland University of Technology
Carnegie Mellon U. (3 groups)	RMIT
Chinese Academy of Sciences	Rutgers University (2 groups)
City University, London	StreamSage, Inc.
Clairvoyance Corp.	Syracuse University
CLIPS-IMAG	Tampere University of Technology
CL Research	TNO TPD, The Netherlands
Columbia University (2 groups)	Tokyo University of Science
CSIRO	Tsinghua University
CWI, The Netherlands	Université d'Angers
Dublin City University	Université de Montreal
Fudan University	University of Amsterdam (2 groups)
Hummingbird	University of Avignon
IBM-Haifa	University of Bremen
IBM-T.J. Watson (3 groups)	University of Buffalo
Illinois Institute of Technology	University of California, Berkeley
Imperial College of Science, Tech. & Medicine	University of Glasgow
Indiana University	University of Hertfordshire
InsightSoft-M	University of Illinois at Chicago
Institut EURECOM	University of Illinois at Urbana/Champaign
IRIT/SIG	University of Iowa
ITC-irst	University of Limerick
Johns Hopkins University, APL	University of Maryland, Baltimore County
Kasetsart University	University of Maryland, College Park (2 groups)
KerMIT Consortium	University of Massachusetts
Laboratory for Information Technology, Singapore	University of Melbourne
Language Computer Corp.	University of Michigan
David Lewis	University of Neuchatel
LIMSI	University of North Carolina, Chapel Hill
Massachusetts Institute of Technology	University of North Texas
Microsoft Research Asia	University of Oulu
Microsoft Research Ltd.	University of Pisa
The MITRE Corp.	University of Sheffield
Moscow Medical Academy	University of Southern California, ISI
National Institute of Informatics	University of Sunderland
National Taiwan University	University of Toronto
National University of Singapore (2 groups)	University of Twente
NTT Communication Science Labs	University of Waterloo
Oregon Health and Science University	University of York
Pohang University of Science and Technology	Yonsei University and ETRI

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 filtering 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 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:  R111
<title> Telemarketing practices U.S.

<desc> Description:
Find documents which reflect telemarketing practices in the U.S. which are intrusive or
deceptive and any efforts to control or regulate against them.
<narr> Narrative:
Telemarketing practices found to be abusive, intrusive, evasive, deceptive, fraudulent,
or in any way unwanted by persons contacted are relevant.  Only such practices in the U.S.
are relevant.  All efforts to halt these practices, including lawsuits, legislation or
regulation are also relevant.
```

Figure 1: A sample TREC 2002 topic from the filtering 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.

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 [10].

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 $X \times$ *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 [12]. 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 [11]. 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 [10].

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 three of the tasks in TREC 2002, the topic distillation task in the web track, the routing task in the filtering 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 2002.

3 TREC 2002 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 2002 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 able to retrieve relevant documents regardless of the language a document happens to be written in. The TREC 2002 cross-language track used Arabic documents and English 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 is the same document collection that was used in the TREC 2001 CLIR track. The collection 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.

Fifty 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 fifty 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

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

Track	TREC 1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
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	6
Spanish	—	—	4	10	7	—	—	—	—	—	—
Confusion	—	—	—	4	5	—	—	—	—	—	—
Database Merging	—	—	—	3	3	—	—	—	—	—	—
Filtering	—	—	—	4	7	10	12	14	15	19	21
Chinese	—	—	—	—	9	12	—	—	—	—	—
NLP	—	—	—	—	4	2	—	—	—	—	—
Speech	—	—	—	—	—	13	10	10	3	—	—
Cross-Language	—	—	—	—	—	13	9	13	16	10	9
High Precision	—	—	—	—	—	5	4	—	—	—	—
Very Large Corpus	—	—	—	—	—	—	7	6	—	—	—
Query	—	—	—	—	—	—	2	5	6	—	—
Question Answering	—	—	—	—	—	—	—	20	28	36	34
Web	—	—	—	—	—	—	—	17	23	30	23
Video	—	—	—	—	—	—	—	—	—	12	19
Novelty	—	—	—	—	—	—	—	—	—	—	13
Total participants	22	31	33	36	38	51	56	66	69	87	93

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 English and Arabic versions of the topics were 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-one runs from nine different groups were submitted to the track. Twenty-three of the runs were cross-language runs and eighteen were monolingual runs. One monolingual run was a manual run. Two groups submitted only monolingual runs, one group submitted only cross-language runs, and the remaining six groups submitted at least one run of each type.

The assessment pools were created using all submitted runs and using the top 100 documents from each run. The average size of the pools was 769 documents. The LDC assessors judged each document in the pools using binary (relevant/not relevant) assessments.

There was some concern over the test collection built in the TREC 2001 CLIR track in that the judgment pools were not as complete as they ideally would be. The 2001 collection contained 25 topics. For 13 of the topics, at least 40 % of the known relevant documents for that topic were retrieved by one group. Further, mean average precision scores decreased by an average of 8 %, with a maximum difference of 21 %, when runs were evaluated without using the group's unique relevant documents. This year's collection has no similar concerns. The average number of relevant documents over the 50 topics is 118.2, with a minimum of 3 relevant documents and a maximum of 523 relevant documents. Only 5 topics had at least 40 % of the relevant documents retrieved by one group. Changes in mean average precision scores when unique relevant documents are removed were similar to the TREC ad hoc collections: an average decrease of a little less than 2 % with a maximum change of 5.7 %.

The average size of the TREC 2001 pools was larger then the average size of the 2002 pools (164.9 vs. 118.2) even though the 2002 pools used more runs and went deeper into the ranked list. Thus, the results produced by different systems are clearly more similar to one another in 2002 than in 2001. But why this should be so is unclear. It could be that the systems are converging to a single effective strategy. The track made a standard set of resources such as stemmers and bi-lingual dictionaries available to participants; common resources are likely to reduce differences among systems. It may also be that the topic set in 2002 was intrinsically easier than the 2001 set, though the effectiveness of the best automatic systems is slightly lower in 2002 than in 2001 which would suggest the opposite conclusion.

As has become common in the CLIR track, the most effective runs as measured by mean average precision (MAP) were cross-language runs, not monolingual runs. The best cross-language run was from the University of Massachusetts, run UMassX6n, with a MAP score of .3996, while the best monolingual run was from the University of Neuchatel, run UniNE3, with a MAP score of .3807. These two runs were the top two runs as measured by precision at document cut-off level 10 as well, but in this case their order was reversed: `UMassX6n` had a P(10) score of.488 and `UniNE3` a score of .516. The University of Massachusetts submitted one monolingual run (MAP: .3619, P(10): .432) though there is not a corresponding cross-language run from the University of Neuchatel. Thus it is not possible to tell from this data whether monolingual access is better for high precision searches in general.

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. Once again there were three tasks in the TREC 2002 filtering track: an adaptive filtering task, a batch filtering task, and a routing 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 2002 filtering task used the same corpus as the TREC 2001 track, "Reuters Corpus, Volume 1, English language, 1996-08-20 to 1997-08-19" from Reuters (`http://about.reuters.com/researchandstandards/corpus/`). 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, and a hierarchy of the Reuters category codes is included with the corpus.

Two distinct types of topics were created for the track. The first set of 50 topics was created by NIST assessors using the standard topic development protocol. Once a candidate topic was provisionally accepted, the author of the topic was given five document sets of approximately 100 documents each to judge for the topic. These document sets were created at NIST by using the relevant documents found in earlier rounds as input to a small set of different feedback systems. The combined set of judged documents was used as the training data for that topic. A second set of 50 topics was created by defining a topic to be the intersection of pairs of Reuters category codes. In this case, a document is relevant to the topic if it has been assigned both of the appropriate category labels. The topic statement for an intersection topic is a simple combination of the category descriptors. A filtering track run was required to process all 100 topics.

Since filtering runs do not produce a ranked list, they cannot be evaluated using the usual IR measures. Instead, filtering runs are generally evaluated using a utility function whereby 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. Because raw utility scores do not average well, the scores for individual topics are normalized, scaled, and then averaged. Details of the TREC 2002 filtering evaluation measures are given in the filtering track overview paper. Routing runs are evaluated using mean average precision since routing runs produce a ranked list of documents.

Seventy-three runs from twenty-one different groups were submitted to the filtering track. Of these, 40 runs are adaptive filtering runs, 16 are batch filtering runs, and 17 are routing runs. The most striking features of the filtering results was the large difference in effectiveness between the assessor-created topics and the intersection topics. System effectiveness was uniformly poor for the intersection topics, including those systems that did relatively well on the human constructed topics, and even including routing runs despite earlier research that shows the initial topic statement is a minor factor in system effectiveness for routing. The intersection topics were included in this year's test set to test whether this very inexpensive topic construction method is adequate for building comparative test collections. Until the reasons for the very large difference in effectiveness are understood, we must conclude that intersection topics are not good substitutes for information need statements.

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 2002 track was the second 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]. In the first year of the plan during TREC 2001, participants performed observational studies of subjects using publicly-accessible tools and the live web to accomplish a search task. The TREC 2002 track followed the observational studies by laboratory experiments focusing on question answering using web data.

The track used an "open" version of the .GOV collection that was created for the TREC 2002 web track. The collection was open in the sense that some links to pages outside the collection could be followed. Most of the participants accessed the collection through the Panoptic search engine made available by CSIRO (see `http://www.panopticsearch.com/`).

The track defined eight different search tasks, two instances each for four general searching activities: looking for personal health information; seeing guidance on US government laws and policies; making travel plans; and gathering material for a report on a given subject. The search tasks were formulated such that the searcher was asked to find either any N short answers to the question or any N web sites that met the need specified in the task. The experimental protocol used in the track was based on the protocol developed for the TREC-9 interactive track and allows the comparison of two systems (or system variants). The protocol required a minimum of 16 searchers. Each searcher performed all eight tasks, half the task on one system and the other half on the other system. Searchers were given at least ten minutes to complete the search, and groups were required to report the results obtained after ten minutes.

Six groups participated in the interactive track. Each group examined their own set of hypotheses as suggested by their TREC 2001 observational studies. See the site reports in these proceedings for details of the individual experiments.

3.4 The Novelty track

The novelty track is a new track in TREC 2002. The goal of the track is to investigate systems' abilities to locate relevant and new (nonredundant) information within the ranked set of documents returned by a traditional document retrieval system. Similar to the question answering track, the motivation for the track is to assist the user of a retrieval system by eliminating extraneous information from the system response.

The data for the track was taken from TRECs 6–8 ad hoc collections. NIST selected 50 topics from that set and selected up to 25 relevant documents for each topic (if there were more than 25 relevant documents, the top 25 according to the document ranking used were selected; in all cases documents from the *Congressional Record* were eliminated). Each set of relevant documents was ranked at NIST using the ordering produced by an effective manual run from the appropriate TREC; participants were required to process the documents in this order. Each document was also split into sentences at NIST and sentences were assigned identifiers.

Table 3: Average F scores for baseline and system results for the Novelty track.

	Relevant	New
Second human judges	0.371	0.353
Random sentences	0.040	0.036
`thunv1`	0.235	0.217

A novelty track run consists of two ordered sets of sentence identifiers for each of the 50 topics. The first set of sentences is the set of sentences the system determined to contain relevant information. The second set of sentences (required to be a subset of the first set) is the set of sentences the system determined to contain new information, that is, relevant information that is not contained in earlier sentences.

Judgment data was created by having assessors manually perform the task. Each topic was independently judged by two different assessors so that the effects of different human opinions could be assessed. In general, the two different assessors did disagree, though much of the disagreement revolved around how much context to include in the relevant set. That is, one assessor would include a string of sequential sentences in the relevant set while the other assessor would select fewer sentences from the same general area of the document. Scoring for the track was based on the smaller relevant set (and its associated new set) because that seemed the best match for the task. Participants were told that the scoring would be based on the smaller set before runs were submitted, but, of course, they did not have access to the assessor sentence sets. One assessor disagreed with the original assessor's relevance judgments for topic 310 and could find no relevant sentences in any of the documents. We eliminated that topic from the final test set, so scores were computed over the remaining 49 topics.

The track guidelines specified sentence set recall and precision as the evaluation measures for the track. Let M be the number of matched sentences, i.e., the number of sentences selected by both the assessor and the system, A be the number of sentences selected by the assessor, and S be the number of sentences selected by the system. Then sentence set recall is M/A and precision is M/S. The F measure with recall and precision weighted equally (i.e., $\beta = 1$) was used as the final score for a topic.

Thirteen groups submitted 43 runs to the novelty track. For all runs, the F score for the relevant sentence sets was greater than the score for the new sentence sets. This might suggest that finding the relevant parts of a document is somewhat easier than finding the nonredundant parts, but is more likely to be a result of the different characteristics of the two tasks. A very small percentage of the total number of sentences were relevant (a median of 2 % across the 49 topics), whereas a very high percentage of the relevant sentences were novel (a median of 93 % across the 49 topics).

One of the requirements for a new track is to do sanity-checking of the evaluation itself. To this end, NIST computed the average F scores for the second human assessor sentence sets and for sets of sentences randomly selected from the target documents. The results are shown in Table 3, which also includes the scores for the most effective system run, run `thunv3`, for comparison. The scores for the best system falls in between the human and random performance, support for a claim that the evaluation is credible.

3.5 The Question Answering (QA) track

The question answering track addresses the problem of information overload by encouraging research into systems that return actual answers, as opposed to ranked lists of documents, in response to a question. The TREC 2002 track contained two different tasks, the main task and the list task. Both tasks were also run in TREC 2001, though there were significant differences in the task definitions between the two years.

Both tasks used a new document collection known as the AQUAINT Corpus of English News Text as the source of answers. This corpus is comprised of documents from three different sources: the AP newswire from 1998–2000, the New York Times newswire from 1998–2000, and the (English portion of the) Xinhua News Agency from 1996–2000. There are approximately 1,033,000 documents and 3 gigabytes of text in the collection. The corpus may be obtained from the Linguistic Data Consortium (`www.ldc.upenn.edu`) as catalog number LDC2002T31.

The main task was the focus of the track. As in previous years, participants received a set of fact-based, short-answer questions, and systems were to return an answer to each question along with the id of a document that supports that answer. In contrast to previous years, systems could return only one response per question, and text snippets containing the answer were not acceptable—systems were required to return nothing more or less than the answer itself. Questions were not guaranteed to have an answer in the collection. A system could return "NIL" as a response to indicate its belief that the collection did not contain an answer to the question.

The change to requiring exact answers was motivated by the belief that forcing systems to return precisely the answer is a necessary step in improving QA technology, not that it is a good idea for deployed QA systems. Whether an answer was exact was determined by the NIST assessor. Assessors judged each response by assigning it exactly one of the following judgments:

incorrect: the answer string returned by the system does not contain a correct answer or the answer is not responsive;

unsupported: the answer string contains a correct answer but the document returned does not support that answer;

non-exact: the answer string contains a correct answer and the document supports that answer, but the string contains more than just the answer or is missing bits of the answer;

correct: the answer string consists of exactly a correct answer and that answer is supported by the document returned.

Being "responsive" means such things as including units for quantitative responses (e.g., $20 instead of 20) and answering with regard to a famous entity itself rather than its replicas or imitations. Only the "correct" judgment was accepted for scoring purposes. NIL was counted as correct when no correct answer was known to exist in the collection for that question.

The test set of questions for the main task consists of 500 questions drawn from MSNSearch and AskJeeves logs. NIST fixed the spelling, punctuation, and sometimes the grammar of the questions selected to be in the final question set, but the content of the question was precisely what was in the log. (Some errors remained despite NIST's attempts to fix such mistakes; questions with errors remained in the test set.) Because it is impossible to know what kind of a response is desired for definition questions (e.g., *Who is Colin Powell? What are steroids?*) when there is no specific target user, none of this type of question was included in the test set. NIST made no other attempt to control the relative frequency of different question types. Forty-six of the questions have no known correct answer in the document collection.

Systems were required to return exactly one response per question. Within the submission file, the *questions* were ordered from most confident response to least confident response. That is, the question for which the system was most confident that it had returned a correct response was ranked first, then the question that the system was next most confident about, etc. so that the last question was the question for which the system was least confident in its response. The question ordering was done to test a system's ability to recognize whether it had found a good response since the final score assigned to a submission was based on this confidence ranking. The confidence-weighted score was inspired by the uninterpolated average precision measure for ranked retrieval output and is defined as

$$\frac{1}{500}\sum_{i=1}^{500}\frac{\text{number correct in first } i \text{ ranks}}{i}.$$

This measure rewards systems for answering questions correctly early in the ranking more than it rewards for answering questions correctly later in the ranking. (This is equivalent to penalizing systems more for incorrectly answering questions early in the ranking.)

Thirty-four different groups participated in the QA track. Each participant submitted at least one main task run for a total of 67 main task runs. The confidence-weighted evaluation measure succeeded in rewarding systems that were able to reliably determine whether they had found a good response, as illustrated in table 4. The table shows the number of questions whose answer was marked correct and the confidence-weighted score

Table 4: Number of questions answered correctly and confidence-weighted score for top 5 TREC 2002 main task QA runs.

Run	Number Correct	Confidence-weighted Score
LCCmain2002	415	0.856
exactanswer	271	0.691
pris2002	290	0.610
IRST02D1	192	0.589
IBMPQSQACYC	179	0.588

for the top five main task runs, ordered by confidence-score. The pris2002 run has a lower confidence score than the exactanswer run despite answering 19 additional questions correctly.

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 *List 9 types of sweet potatoes.* 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. As in the main task, answer-strings were required to be exact.

The questions for the list task 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. Individual instances were judged as in the main task. In addition, the assessor also marked a set of instances as distinct. The assessor arbitrarily chose any one of a set of equivalent correct instances to mark as the distinct one, and marked the remainder as not distinct. The accuracy score for a list question was computed as the number of correct distinct instances retrieved divided by the number of requested instances. The score for a run was the average accuracy over the 25 questions in the test set.

Five groups submitted nine runs for the list task.

3.6 The Video track

TREC 2002 was the second year for the video track, a track designed to promote progress in content-based retrieval from digital video. This year's track contained three tasks: the shot boundary task, the feature extraction task, and the search task.

The video data for the track consisted of MPEG-1/VCD recordings from the Internet Archive (http://www.archive.org/movies) and the Open Video Project (http://www.open-video.org). The track defined a different set of files from these sources as the development sets and test sets for the different tasks. The search test collection contained approximately 40 hours of video, and the feature extraction and shot boundary test collections each contained about five hours of video. For the search and feature extraction tasks, the track also published a reference set of shot boundaries for the video collection. Runs for these two tasks returned lists of shots as defined by the reference set.

The goal in the shot boundary task was to (automatically) identify the shot boundaries in a given video clip. In addition to giving the location of the boundary as a time offset, systems were also required to specify whether the boundary was a cut or a gradual transition. System output was evaluated using automatic comparison to a set of reference shot boundaries created manually at NIST, using set recall and precision as the measures. Frame recall and frame precision (recall and precision of the individual frames within the shot) were also computed for each gradual transition detected by the system. Eight groups submitted 53 shot boundary runs.

There were two main motivations for the new feature extraction task. First, the ability to detect semantic

Outdoors: Segment contains a recognizably outdoor location.

Indoors: Segment contains a recognizably indoor location.

Face: Segment contains at least one human face with the nose, mouth, and both eyes visible.

People: Segment contains a group of two more humans, each of which is at least partially visible and is recognizable as a human.

Cityscape: Segment contains a recognizably city/urban/suburban setting.

Landscape: Segment contains a predominantly natural inland setting.

Text Overlay: Segment contains superimposed text large enough to be read.

Speech: A human voice uttering words is recognizable as such in this segment

Instrumental Sound: Sound produced by one or more musical instruments is recognizable as such in this segment.

Monologue: Segment contains an event in which a single person is at least partially visible and speaks for a long time without interruption by another speaker.

Figure 2: Descriptions of features to be detected in the Video track's feature extraction task.

concepts within video is seen as key to providing content-based access. The task is a first step toward building a benchmark for evaluating the effectiveness of particular feature detection methods. Second, the track implemented a plan whereby participants' extraction output for features specific to the search task were made available to other participants. This allowed the track to investigate methods for exploiting detected features in a search task.

Ten different features, shown in figure 2, were specified as test features. Shots containing the features were determined by NIST assessors (using pools of shots submitted by participants as for document relevance assessing). A shot contains a feature if at least one frame within the shot matches the feature's description, and otherwise does not contain the feature.

A feature extraction run consisted of a ranked list of up to 1000 shots ordered by likelihood that the shot contains the feature. Runs were evaluated using precision and recall, as well as uninterpolated average precision. Measures were computed for each feature individually, but not averaged across features. Eleven groups submitted 18 feature extraction runs.

The search task was a typical ad hoc retrieval task where the "documents" were video shots and the topics were multimedia statements of information need. A search task run consisted of a ranked list of the top 100 shots ordered by likelihood that the shot satisfies the topic. The track distinguished two type of runs: "manual" runs where a human formulated the query based on the topic but there was no further human intervention in the run, and "interactive" where a human formulated an initial query and then refined the query based on initial search output to form the final ranked list. Groups submitting interactive runs were required to report the amount of time the searcher spent to produce the final ranked list. Effectiveness was measured using traditional ranked retrieval measures.

The 25 test topics were created at NIST. The user model assumed during the topic creation process was that of a trained searcher trying to find material for reuse from a large video archive. Each topic statement included a brief textual description of the desired information (e.g., "Find shots containing Washington Square Park's arch in New York City.") and one or more examples of the desired information. Examples were references to video clips, still images, and audio clips. Twelve groups submitted 40 search task runs. Thirteen of the runs were interactive runs and 27 were manual runs.

3.7 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. This year's track used a new document set and defined two new tasks. The topic distillation task is an ad hoc retrieval task in which the goal is to retrieve "key resources" rather than relevant documents. The named page finding task is a known-item task where the goal is to find a particular page that has been named by the user.

The document collection used for both tasks was a new collection known as the .GOV collection (`http://www.ted.cmis.csiro.au/TRECWeb/govinfo.html`) because it is based on a crawl of .gov web sites. The crawl occurred in January, 2002 and was made to mimic the way a real search service of the .gov pages might make a crawl. The crawl was breadth-first and stopped after one million html pages had been fetched. The crawl also included approximately 250,000 other types of pages (postscript, word, and pdf files) as well as images. The documents in the collection contain both page content and the information returned by the http daemon; text extracted from the non-html pages is also included in the collection. Charlie Clark of the University of Waterloo made the crawl using a machine made available by Ed Fox of Virginia Tech. The document collection was created from the crawl by CSIRO who are also distributing the collection.

The goal in the topic distillation task is to assemble a short, but comprehensive, list of pages that are good information resources on a particular topic. An example use for such a search is a user assembling a bookmark list for the target topic. Examples of key resources pages include the home page of a site dedicated to the topic; the main page of a sub-site dedicated to the topic; a highly useful single document (e.g., a postscript document) dedicated to the topic; a highly useful page of links (hub page) on the topic; and a relevant service such as a search engine dedicated to the topic.

NIST assessors created 50 topics for the topic distillation task. The topics are much like regular TREC topics, except the content targets topics for which the .GOV collection contains good key resources. Assessment was performed on pooled results as in a standard ad hoc task, but a document was judged as to whether it is a good resource pages, not whether the page is relevant. The main evaluation measure used for the task was precision at cut-off level 10 to focus the systems on retrieving a concise list of good resources.

The named page task is similar to the homepage finding task from TREC 2001 except the target page could be any page in the collection rather than an entry page to a site. The topics for the name page task were created by NIST assessors who searched the document collection looking for pages that were unique and that contained content a user might want to return to. For example, one topic asked for the directions to the Berkeley National Laboratory. Topics consisted of a single phrase, such as "directions Berkeley National Laboratory" for the example above.

The goal in the task is for a system to return the one target page for each topic. For evaluation, participants returned a ranked list of 50 documents per topic, and were scored using the mean reciprocal rank of the target page across the 150 test topics. Small pools consisting of the top 10 pages from each judged run were created to check for pages that had different DOCNOs but were equivalent pages (caused by mirroring and the like). The rank of the target page whose rank was closest to one was used as the score for each topic.

Twenty-three groups submitted a total 141 runs to the web track. Of those runs, 71 were topic distillation task runs and 70 were named page finding task runs. The results of the topic distillation task suggest there is still some question as to how exactly the task should be implemented for both assessors and participants. The web track in TREC 2003 will explore these questions in depth, including adding an interactive version of the task.

4 The Future

TREC 2003 will see significant changes in the tracks to be offered: several existing tracks will be suspended and new tracks introduced. The video track will be spun off into its own evaluation program to allow the effort to expand to include other facets of video retrieval. The new TRECVID[1] workshop will meet at NIST

[1] `http://www-nlpir.nist.gov/projects/trecvid`

immediately prior to TREC 2003. Since the NTCIR[2] and CLEF[3] evaluations provide venues for cross-language retrieval research, the Cross-Language track will be discontinued in TREC. The interactive track will not be a separate track in TREC 2003, but interactive subtasks will be incorporated into other tracks. In particular, the web track will have an interactive version of the topic distillation task in 2003. Finally, the filtering track will also not run in TREC 2003. Participants interested in the filtering task are encouraged to use the filtering track mailing list (see `http://trec.nist.gov/tracks.html`) to discuss plans for future tracks.

Three new tracks will be added to TREC 2003. The Genome track will provide a forum for the evaluation of information retrieval systems in the genomics domain. This first running of the track in 2003 follows an exploratory "pre-track" that occurred during 2002. Each of the remaining two new tracks will explore different aspects of ad hoc retrieval. The task in the Robust Retrieval track will be a traditional ad hoc task, but with an emphasis on individual topic effectiveness rather than average effectiveness. The goal of this track is to improve the consistency of retrieval technology by focusing on poorly performing topics. The goal of the HARD (Highly Accurate Retrieval from Documents) track is also to improve the effectiveness of ad hoc searches, but in this case the emphasis will be on customizing retrieval for individual users by exploiting information about the search context and using very targeted interaction with the searcher.

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

[2]`http://research.nii.ac.jp/ntcir`
[3]`http://clef.iei.pi.cnr.it`

[11] 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`.

[12] 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.