Overview of TREC 2003

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

1 Introduction

The twelfth Text REtrieval Conference, TREC 2003, was held at the National Institute of Standards and Technology (NIST) November 18–21, 2003. The conference was co-sponsored by NIST, the US Department of Defense Advanced Research and Development Activity (ARDA), and the Defense Advanced Research Projects Agency (DARPA).

TREC 2003 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 2003 contained six areas of focus called "tracks". Three of the tracks, novelty, question answering, and web, were continuations of tracks that had run in earlier TRECs. The remaining three tracks, genomics, High-Accuracy-Retrieval-from-Documents (HARD), and robust retrieval, were new tracks in 2003. The retrieval tasks performed in each of the tracks are summarized in Section 3 below.

Table 1 lists the 93 groups that participated in TREC 2003. The participating groups come from 22 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 MEDLINE record.

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 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.

Table 1: Organizations participating in TREC 2003

Ajou University	NTT Communication Science Laboratories
Axontologic, Inc.	OcE Technologies
BBN	Oregon Health and Science University
California State University San Marcos	Queens College, CUNY
Carnegie Mellon University (2 group)	RMIT University
Center for Computing Science & U. Maryland	Rutgers University (3 groups)
Chinese Academy of Sciences (2 groups)	Saarland University (2 groups)
Chinese Information Processing Center	Sabir Research, Inc.
Clairvoyance Corporation	State University of New York at Buffalo
CL Research	StreamSage, Inc.
Copernic Research	Tarragon Consulting Corporation
CSIRO	Tsinghua University (2 groups)
Dublin City University	Universitat Politecnica de Catalunya & Universitat de Girona
Erasmus MC	Université de Neuchatel
Fondazione Ugo Bordoni	University Hospital of Geneva
Fraunhofer Institute (SCAI)	University of Alaska, Fairbanks
Fudan University	University of Albany
Hummingbird	University of Amsterdam
IBM Research, Haifa	University of California, Berkeley
IBM TJ Watson Research Center (2 groups)	University of Colorado & Columbia U.
Illinois Institute of Technology	University of Edinburgh
Indiana University, Bloomington	University of Edinburgh & Stanford U.
Indian Institute of Technology Bombay	University of Glasgow
IRIT/SIG	University of Helsinki
ITC-irst	University of Illinois at Chicago
Johns Hopkins University/APL	University of Illinois at Urbana-Champaign
Kasetsart University	University of Iowa
Korea University	University of Limerick
Language Computer Corporation	University of Maryland
Lehigh University	University of Maryland Baltimore County
LexiClone, Inc.	University of Massachusetts
Macquarie University	University of Melbourne
Massachusetts Institute of Technology	University of Michigan
Meiji University	University of Pisa
Microsoft Research Asia	University of Sheffield
Microsoft Research Ltd	University of Southern California/ISI
MITRE Corp.	University of Sunderland
National Library of Medicine & U. Maryland	University of Tampere
National Research Council Canada	University of Tokyo
National Taiwan University	University of Wales, Bangor
National University of Singapore (2 groups)	University of Waterloo (2 groups)
New Mexico State University	Virginia Tech

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. TREC 2003 did not contain an explicit filtering task, though aspects of the filtering task were present in the novelty track tasks.

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. The information extraction task in the genomics track is similar to a question answering task in that the goal was to extract a short segment of a document as a description of a gene.

2.1 Test collections

Text retrieval has a long history of using retrieval experiments on test collections to advance the state of the art [4, 6, 10], 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 earliest TRECs, but it has been stable since TREC-5 (1996). 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 robust retrieval 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. For topics 301 and later, the "title" field was specially designed to allow experiments with very short queries; these 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.

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

```
<num> Number:  602
<title> Czech, Slovak sovereignty

<desc> Description:
Retrieve information regarding the process by which the Czech and Slovak
republics of Czechoslovakia established separate sovereign countries.

<narr> Narrative:
A relevant document will provide specific dates and details regarding the
separation movement.  Documents relating to normal activities of the separate
nations, both internal and external are not relevant.
```

Figure 1: A sample TREC 2003 topic from the robust retrieval track.

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 [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 [9] 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 than 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 [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 0.1 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. 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 [3]. 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.

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	2003
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	—	—	—	—	—	—	—
DB 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	—	—	—	—	—
VLC	—	—	—	—	—	—	7	6	—	—	—	—
Query	—	—	—	—	—	—	2	5	6	—	—	—
QA	—	—	—	—	—	—	—	20	28	36	34	33
Web	—	—	—	—	—	—	—	17	23	30	23	27
Video	—	—	—	—	—	—	—	—	—	12	19	[a]
Novelty	—	—	—	—	—	—	—	—	—	—	13	14
Genome	—	—	—	—	—	—	—	—	—	—	—	29
HARD	—	—	—	—	—	—	—	—	—	—	—	14
Robust	—	—	—	—	—	—	—	—	—	—	—	16
Total participants	22	31	33	36	38	51	56	66	69	87	93	93

[a]The video track was spun off as a separate evaluation effort in 2003.

As TREC has expanded into tasks other than the traditional ad hoc retrieval task, new evaluation measures have had to be devised. Indeed, developing an appropriate evaluation methodology for a new task is one of the primary goals of the TREC tracks. The details of the evaluation methodology used in a track are described in the track overview paper.

3 TREC 2003 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 2003 tracks. See the track reports later in these proceedings for a more complete description of each track. Some of the descriptions given here are taken directly from the track overview papers.

3.1 The genomics track

The genomics track was a new track for TREC 2003. It is the first TREC track devoted to retrieval within a specific domain, and one of the goals of the track is to see how exploiting domain-specific information improves retrieval effectiveness. The track contained two tasks, the primary task that was an ad hoc retrieval task and the secondary task that was an information extraction task.

The scenario that motivated the primary task was that of a biological researcher or graduate student—that is, someone who already has considerable domain knowledge—confronted with the need to learn about a new gene very quickly. Since NIST assessors do not have the expertise to make judgments for the track, this first track made use of existing data that could serve as surrogate relevance judgments. The document collection consisted of approximately 526,000 MEDLINE records that were indexed between April 1, 2002 and April 1, 2003, and were donated to the track by the U.S. National Library of Medicine. A topic consisted of a gene name and an organism, and was to be interpreted as a request for the basic biology of the gene and its protein products in the designated organism. This is the information given by the Gene Reference into Function (GeneRIF) data in the LocusLink database, a database of biological information created by the National Center for Biotechnology Information. The GeneRIF data were used as the relevance judgments for the track.

An analysis of the use of GeneRIF data showed that the vast majority of GeneRIF references pointed to relevant documents, but the GeneRIF references were incomplete (i.e., there were many more relevant documents than those included as GeneRIF references). Incompleteness is not necessarily a problem for retrieval system evaluation in that *unbiased* incomplete judgments allow for fair comparisons. Unfortunately, the GeneRIF data are not unbiased relevance judgments: the Berkeley group was able to build a classifier that could distinguish documents likely to be GeneRIFs [2]. The track will need to obtain relevance judgments in some other manner in future years.

Twenty five groups submitted 49 primary task runs to the genomics track. The best performing runs did use domain-specific knowledge as part of the retrieval. Exploiting the Medical Subject Headings (MeSH) and substance name fields of the MEDLINE records and filtering for species were particularly beneficial.

Part of the GeneRIF data is a text snippet that summarizes the main point of the referred to document with respect to the gene and organism. The secondary task was an information extraction task with the goal of creating this GeneRIF annotation automatically. The test set for the secondary task consisted of 139 GeneRIFs. Effectiveness was measured as a function of the overlap between the words nominated by the system and the actual GeneRIF text.

Fourteen groups submitted 24 secondary task runs. Since the actual GeneRIF text for many of the annotations is taken directly from the title of the target document, a baseline run consisting of the title of each target document was very hard to beat. The few runs that were able to beat the baseline used classifiers to rank sentences likely to contain GeneRIF text.

3.2 The HARD track

The HARD track was another new track in TREC 2003. HARD stands for High Accuracy Retrieval from Documents, and the goal of the track was to improve retrieval performance by targeting retrieval results to the specific user. Of course, to target retrieval results in such a manner the system needs to have some knowledge about the user. The HARD track provided this information in the form of biographical data about the user, information regarding the search context, and a statement of the expected type of a result.

The underlying task in the HARD track was an ad hoc retrieval task. However, for some topics the expected type of a result was passages rather than documents. Combining document and passage retrieval into a common evaluation methodology was one of the aspects explored in the track. Another aspect was the use of "clarifying forms" to gather information about the searcher. A clarification form was a single web page that solicited information about the query from the user. Any information from the user could be collected by the form subject to the constraints that the user would spend no more than 3 minutes filling out any one form and that the form had to be entirely self-contained HTML.

The document set used in the track was the set of documents from 1999 from the AQUAINT corpus plus a set of *Congressional Record* and *Federal Register* articles also from 1999. This collection consisted of approximately 372,000 documents and 1.7 GB of text. The topics were created by assessors from the Linguistics Data Consortium (LDC). The topics were patterned after standard TREC ad hoc topics, but included a set of metadata elements that described the searcher and/or the context of the search. For example, the PURPOSE metadata field explained why the user was searching for the information (its value could be one of background, details, answer, or any) and the FAMILIARITY field represented how familiar the searcher is with the general subject area of the topic (value between 1 and 5 with 1 meaning no prior knowledge and 5 meaning detailed knowledge of the subject; value could also be unknown). Biographical data such as the age, sex, and occupation of the searcher were also recorded.

Participants first ran their systems using just the standard TREC portions of the topic and no other information. They then repeated the search using any information from the metadata and/or their clarification forms. The goal was to see if the additional information helped systems to create a more effective retrieved set than the initial baseline result.

Relevance judgments were made at the LDC by the same assessor who created the topic. Two types of judgments were made, document-level judgments and passage-level judgments. Document-level judgments made without reference to the metadata are the same as standard TREC relevance judgments. Documents that are relevant in the standard TREC sense but do not meet the requirements specified by the metadata are called "SOFT-REL" documents, while relevant documents that also satisfy the metadata are called "HARD-REL". For document-level evaluation, HARD track runs were evaluated using the standard `trec_eval` measures, treating either both SOFT-REL and HARD-REL documents as relevant, or just HARD-REL documents as relevant.

Passage-level judgments were also made by the LDC assessor. If the metadata for a topic specified that the user wanted something smaller than a full document as a response, the assessor looked at each HARD-REL document in turn and marked the passages within the document that satisfied the topic. Passages were specified by an offset from the beginning of the document and a length. A single document could contain multiple relevant passages, but relevant passages never overlapped (overlapping passage were combined into a single passage if necessary). The relevance judgments were assumed to contain all the relevant passages for the topic.

The main measure used for passage-based evaluation was R-precision where R is the number of relevant passages for a topic. The passage-based evaluation treated all system responses as passages (i.e., a retrieved document was considered a single long passage). Precision was calculated on the basis of characters: the passage-based precision for a system response at rank R was the proportion of characters in the sum of the passages at ranks 1–R that were contained in a relevant passage.

Fourteen groups submitted 88 runs to the HARD track. For most groups, runs based on data obtained from clarification forms improved results as compared to the corresponding baseline run. Evaluation based on passages differs from that based on documents in that systems ranked differently when evaluated by passage-based R-precision than when evaluated by document-based R-precision.

3.3 The novelty track

The goal of the novelty track is to investigate systems' abilities to locate relevant and new (nonredundant) information within an ordered set of documents. This task models an application where the user is skimming a set of documents and the system highlights the new, on-topic information. The track was first introduced in TREC 2002, though this year's track had a number of significant changes from the initial track.

The basic task in the novelty track is as follows: given a topic and an ordered set of relevant documents segmented into sentences, return sentences that are both relevant to the topic and novel given what has already been seen. To accomplish this task, participants must first identify relevant sentences (a passage retrieval task) and then identify which sentences contain new information (a filtering task). To allow participants to focus on the filtering and passage retrieval aspects separately, four different tasks were included in the track where each task differed by the amount and kind of training data that was provided to the systems.

Fifty new topics were created for the novelty track by NIST assessors. Half of the topics focused on events and the other half focused on opinions about controversial subjects. For each topic, the assessor created a statement of information need and queried the document collection using the NIST PRISE search engine. The assessor selected 25 relevant documents and labeled the relevant and new sentences in each. The document collection used was the *AQUAINT Corpus of English News Text* assembled for the TREC 2002 question answering track. 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 choice of the collection was motivated by a desire to increase the amount of redundancy in the relevant set as compared to last year's track. The 25 relevant documents for each topic were ordered chronologically for system processing, which is easily accomplished for a newswire collection.

The four tasks in the track allowed the participants to test their approaches to novelty detection using no, partial, or complete relevance information.

Task 1. Given the set of 25 relevant documents for a topic, identify all relevant and novel sentences.

Task 2. Given the relevant sentences in all 25 documents, identify all novel sentences.

Task 3. Given the relevant and novel sentences in the first 5 documents for the topic, find the relevant and novel sentences in the remaining 20 documents.

Task 4. Given the relevant sentences in all 25 documents, and the novel sentences in the first 5 documents, find the novel sentences in the remaining 20 documents.

Given the set of relevant and new sentences selected by the assessor who created the topic, the score for a novelty topic was computed as the F measure where sentence set recall and sentence set precision are equally weighted. 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 score is then computed as $F = \frac{2*P*R}{(P+R)}$.

Fourteen groups submitted 179 runs to the novelty track. All but one group submitted a run for Task 1, and most groups tried all tasks. The results showed that for the basic task in which systems were given no sentence-level training data, the best systems were more effective than human performance. That is, a second assessor who selected relevant and novel sentences based on the topic statement generally scored lower when evaluated by the author's sentences than did the systems. More data is required to determine if systems are indeed performing at the level of a human at this task.

3.4 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 track has run since TREC-8 (1999), but has expanded in both scope and difficulty since the initial tracks. The TREC 2003 track contained two tasks, the main task and the passages task. Both tasks used the *AQUAINT Corpus of English News Text* used in the novelty track as the source of answers.

In TREC 2002, the QA task was defined such that systems were required to return exact answers, text strings consisting of a complete answer and nothing else. However, pinpointing the precise extent of an answer is a more difficult problem than finding a text segment that contains an answer, and there are applications of QA technology that do not require this extra step. The passages task provided a forum for research groups interested in these applications. A passages task run consisted of exactly one response for each of a set of 413 factoid questions. A response was either a document extract (not longer than 250 characters) believed to contain an answer to the question or the string "NIL" used to indicate the system's belief that there was no correct answer in the collection. Responses were judged as either correct, unsupported, or incorrect by human assessors. The final score for a passages task run was accuracy, the percentage of responses judged correct.

Twenty-one passages task runs from eleven different groups were submitted to the QA track. As determined by comparing mean accuracy scores, the passages task was not a noticeably easier task than the exact answer task. Accuracy scores for the passages task were in general no better than accuracy scores for the factoid component in the main task that required exact answers. Two of the three groups that submitted runs for both tasks had higher accuracy scores for the exact-answer case.

The main task was a combination task consisting of three different types of questions: factoids, lists, and definitions. The goal in combining the different question types into a single task was to increase the number of systems that attempted to answer the different question types. Each question was tagged as to its type in the test set. The three question types were evaluated separately, and the final score for a main task run was a combination of the scores for the three question types.

The factoid component of the main task was identical to the passages task except responses were required to be exact answers rather than document extracts that contained an answer. A fourth value for the judgments, inexact, was added to indicate when an otherwise correct response contained too much information. As in the passages task, the score for the factoid component of the main task was accuracy.

The list component of the main task required systems to assemble an answer from information located in multiple documents. In TREC, a list question asks for different instances of a particular kind of information to be retrieved, such as *List the names of chewing gums*. List questions can be thought of as a shorthand for asking the same factoid question multiple times; the set of answers that satisfy the factoid question is the appropriate response for the list question. Unlike the previous two times the list task was run in TREC, this year's list questions did not specify a target number of instances to return. Instead, systems were expected to return all of the correct, distinct answers contained in the document collection. There were 37 list questions in the main task test set.

Within the response returned for a single question by one system, assessors judged individual items as the factoid responses were judged. In addition, the assessor marked exactly one of a set of equivalent correct items as distinct. The final answer list for a question was created by the assessor based on the answers the assessor found during question development and the set of distinct, correct answers found by the systems. This final answer list was used to compute the instance recall and instance precision of a system's response. Instance recall is the fraction of answers on the final answer list that the system returned. The corresponding instance precision measure is the fraction of instances returned by the system that are on the final answer list. Instance recall and precision were combined using the F measure with recall and precision equally weighted ($F = \frac{2*IP*IR}{(IP+IR)}$) as the final score for a list question. The score for the entire list component of the main task was the average of the F scores over the 37 questions.

Definition questions are questions such as *Who is Colin Powell?* or *What is mold?*. This was the first time definition questions were evaluated in TREC. The evaluation was based on a small pilot evaluation of definition questions that was held as part of the ARDA AQUAINT program in the fall of 2002 [12]. Evaluating systems that answer definition questions is much more difficult than evaluating systems that answer factoid questions because it is no longer useful to judge a system response as simply right or wrong. Assigning partial credit to a system response requires some mechanism for matching the concepts in the desired response to the concepts present in the system's response. The issues are similar to those that arise in the evaluation of machine translation and automatic summarization.

The following scenario was assumed for definition questions:

> The questioner is an adult, a native speaker of English, and an "average" reader of US newspapers. In reading an article, the user has come across a term that they would like to find out more about. They may have some basic idea of what the term means either from the context of the article (for example, a bandicoot must be a type of animal) or basic background knowledge (Ulysses S. Grant was a US president). They are not experts in the domain of the target, and therefore are not seeking esoteric details (e.g., not a zoologist looking to distinguish the different species in genus Perameles).

The definition question test set contained 50 questions drawn from search engine logs; the set contained 30 questions for which the target was a (perhaps fictional) person, 10 questions for which the target was an organization, and 10 questions for which the target was some other thing.

A system response for a definition question was an unordered set of [*document-id*, *answer-string*] pairs. Each string was presumed to be a facet in the definition of the target. There were no limits placed on either the length of an individual answer string or on the number of pairs in the list, though systems were penalized for retrieving extraneous information.

Judging the quality of the systems' responses was done in two steps. In the first step, all of the answer-strings from all of the responses were presented to the assessor in a single (long) list. Using these responses and his own research done during question development, the assessor first created a list of "information nuggets" about the target. An information nugget was defined as a fact for which the assessor could make a binary decision as to whether a response contained the nugget. At the end of this step, the assessor decided which nuggets were vital—nuggets that must appear in a definition for that definition to be good. The assessor went on to the second step once the nugget list was created. In this step the assessor went through each of the system responses in turn and marked where each nugget appeared in the response. If a system returned a particular nugget more than once, it was marked only once. A single item in a system's response may match zero, one, or more than one nuggets.

Given the judgments as described above, it is straightforward to compute the nugget recall of a response: it is simply the ratio between the number of correctly retrieved nuggets to the number of nuggets on the assessor's list. But the corresponding measure of nugget precision, the ratio between the number of nuggets correctly retrieved to the total number of nuggets retrieved, is problematic since the correct value for the denominator is unknown. A trial evaluation prior to the pilot showed that assessors found enumerating *all* concepts represented in a response to be so difficult as to be unworkable. Instead, we used length as a crude approximation to precision. The length-based measure captures the intuition that users would prefer the shorter of two definitions that contain the same concepts. The final score for a definition question was the F measure where nugget recall was given five times as much emphasis as nugget precision. The score for the definition component of the main task was the average F over the 50 definition questions.

The final score for a main task run was computed as a weighted average of the three component scores:

$$\text{FinalScore} = 1/2 * \text{FactoidScore} + 1/4 * \text{ListScore} + 1/4 * \text{DefScore}.$$

Since each of the component scores ranges between 0 and 1, the final score is also in that range. The final score emphasizes the factoid component, which represented the largest number of questions and is the task people are most familiar with. The weight for the other components was made large enough to encourage participation in those subtasks.

Fifty-four main task runs from 25 different groups were submitted to the track. The results demonstrate that the list and definition tasks are challenging for systems, and that they present challenges for evaluation as well. For the definition task, the difference in evaluation scores required to have confidence in the conclusion that one run is better than another is large relative to the observed scores. This results in a fairly insensitive test since many comparisons are inconclusive. The list task scores are much more stable, but the stability is due in large part to the fact that the scores for the list task are very low.

3.5 The robust retrieval track

The robust retrieval track was another new track in TREC 2003. The goal of the track was to focus research on improving the consistency of retrieval technology by concentrating on poorly performing topics. In addition, the track brought back a classic ad hoc retrieval task to TREC.

The topic set used in the track was a set of 100 topics. Fifty of the topics were new, created by NIST assessors using the standard topic development procedure. The other 50 topics were old topics first used in the ad hoc tasks of TRECs 6–8. NIST selected these 50 topics based on the median mean average precision (MAP) score when the topic was first used: the 50 topics all had low median MAP scores with at least one run that did much better than the median to rule out flawed topics.

Since 50 of the topics were from previous TRECs, the track used the same document set as those years, namely the set of documents on TREC disks 4 and 5 minus the *Congressional Record* documents. No new relevance judgments were made for these topics. The 50 new topics were judged using pools created from all runs using a depth of 125 documents per topic per run. Evaluation was performed using `trec_eval` on each subset of the topics and on the combined set of 100 topics. Two new measures that focused on the poorly-performing topics were also introduced. The first of these measures was the percentage of topics that returned no relevant documents in the top ten documents retrieved. The second measure is a much more sensitive, but far less intuitive measure. If there are a total of Q topics in the test set, plot the MAP score computed over a system's worst X topics (as measured by average precision) against X for $X = 1 \ldots Q/4$. The measure is the area underneath this curve. Note that since the measure is computed over the individual system's worst X topics, different systems' scores are computed over a different set of topics in general.

The robust track received a total of 78 runs from 16 participants. All of the runs submitted to the track were automatic runs. The results of the track provide strong confirmation that average values of the traditional effectiveness measures do not reflect poorly performing topics. The new measures do emphasize systems' worst topics, but because they are defined over a subset of the topics, they are much less stable than traditional measures for a given test set size.

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. This year's track focused on finding homepages, the main entry pages to sites. There were two non-interactive tasks and one interactive task in the track.

All tasks used the .GOV collection created for the TREC 2002 web track and distributed by CSIRO (see `http://www.ted.cmis.csiro.au/TRECWeb/govinfo.html`). This collection is based on a January, 2002 crawl of .gov web sites. 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.

The two non-interactive tasks in the track were a topic distillation task and a navigational task known as the home/named page finding task. In the topic distillation task, the systems were given a broad information request and were to return a list of relevant home pages. A relevant home page was defined as the entry page to a credible site that is principally devoted to the topic. The emphasis was on returning home pages rather than pages themselves since a result list of homepages provides a better overview of the coverage of a topic in the collection. The primary effectiveness measure used was R-precision (precision after R relevant documents are retrieved) since many of the topics within the set of 50 test topics had fewer than 10 relevant home pages.

The navigational task was a known-item search task. The queries consisted of a very short description of a page such as "Tennessee Valley Authority", and the systems were to return the target page (in this case, `www.tva.gov`). The test set consisted of 300 queries, half of which had a home page as the target page. Effectiveness was measured by the mean over the 300 topics of the reciprocal of the rank at which the target page was returned.

Twenty-seven groups submitted a total of 166 runs to the non-interactive part of the web track. Ninety-three of the runs were topic distillation runs and 73 of the runs were navigational task runs. Results from both tasks showed that exploiting anchor text is an important element of effective homepage finding. Methods that exploited URL syntax and link structure had more mixed results, especially for the navigational task. Attempts to differentiate processing for named pages vs. homepages in the navigational task did not improve effectiveness.

The interactive task within the web track explored the role of the human searcher in the topic distillation task. Eight of the topics used in the non-interactive version of the task were expanded to include a search scenario to provide context for the searcher. The searchers produced a list of home pages for the topic which were then judged by the assessors along four dimensions: relevance, depth, coverage, and repetition. Each dimension was judged using a 5-point Likert scale.

Two groups participated in the task. Both groups explored whether a more structured presentation of the search results (rather than a simple ranked list) would better support a searcher in the topic distillation task. The searchers liked the structured result format better, and were somewhat more efficient with it, but there were no significant differences between the list and structured formats in the quality of the homepage lists the searchers assembled.

4 The Future

Since three of the six tracks offered in TREC 2003 were new tracks, the set of tracks to be offered in TREC 2004 will be little changed from this year. Each of the six tracks will continue in TREC 2004. In addition, a new track, currently known as the terabyte track, will be added. The main objective in the terabyte track will be to investigate ad hoc evaluation methodologies for terabyte scale collections [8]. Of course, the track will also offer participants the opportunity to see how well their retrieval methods scale to significantly larger collections.

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] G. Bhalotia, P.I. Nakov, A.S. Schwartz, and M.A. Hearst. BioText team report for the TREC 2003 genomics track. In *Proceedings of the Twelfth Text REtrievalxi Conference (TREC 2003)*, 2004.

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

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

[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] Ian Soboroff, Ellen Voorhees, and Nick Craswell. Summary of the SIGIR 2003 workshop on defining evaluation methodologies for terabyte-scale test collections. *SIGIR Forum*, 37(2):55–58, 2003.

[9] K. Sparck Jones and C. van Rijsbergen. Report on the need for and provision of an "ideal" information retrieval test collection. British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge, 1975.

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

[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. Evaluating answers to definition questions. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003), Volume 2*, pages 109–111, May 2003.

[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.