Overview of TREC 2004

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

1 Introduction

The thirteenth Text REtrieval Conference, TREC 2004, was held at the National Institute of Standards and Technology (NIST) November 16–19, 2004. 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 2004 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 2004 contained seven areas of focus called "tracks". Six of the tracks had run in at least one previous TREC, while the seventh track, the terabyte track, was new in TREC 2004. The retrieval tasks performed in each of the tracks are summarized in Section 3 below.

Table 2 at the end of this paper lists the 103 groups that participated in TREC 2004. 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 toward 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. Most of the retrieval tasks in TREC 2004 are ad hoc tasks.

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. The named page finding part of the web track task is a known-item search.

In a *categorization* task, the system is responsible for assigning a document to one or more categories from among a given set of categories. The genomics track had several categorization tasks in TREC 2004, and the novelty track tasks required assigning sentences from within documents to "relevant" and "novel" categories. The web track also had a variant of a categorization task, though in this case the topics, not the documents, were to be categorized.

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 terabyte track was introduced this year to investigate both retrieval and evaluation issues associated with collections significantly larger than 2 gigabytes of text.

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 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

```
<num> Number:  656
<title> lead poisoning children
<desc>
How are young children being protected against lead poisoning from paint and
water pipes?
<narr>
Documents describing the extent of the problem, including suits against
manufacturers and product recalls, are relevant.  Descriptions of future plans
for lead poisoning abatement projects are also relevant.  Worker problems with
lead are not relevant.  Other poison hazards for children are not relevant.
```

Figure 1: A sample TREC 2004 topic from the robust track test set.

short queries; these title fields consist of up to three words that best describe the topic. The description ("desc") field is a one sentence description of the topic area. The narrative ("narr") gives a concise description of what makes a document relevant.

Participants are free to use any method they wish to create queries from the topic statements. TREC distinguishes among two major categories of query construction techniques, automatic methods and manual methods. An automatic method is a means of deriving a query from the topic statement with no manual intervention whatsoever; a manual method is anything else. The definition of manual query construction methods is very broad, ranging from simple tweaks to an automatically derived query, through manual construction of an initial query, to multiple query reformulations based on the document sets retrieved. Since these methods require radically different amounts of (human) effort, care must be taken when comparing manual results to ensure that the runs are truly comparable.

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

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 usually 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 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 [10].

2.2 Evaluation

Retrieval runs on a test collection can be evaluated in a number of ways. In TREC, ad hoc tasks are evaluated using the `trec_eval` package written by Chris Buckley of Sabir Research [1]. 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 at ten documents 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 at ten documents 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 par-

ticular 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, average precision is the area underneath a non-interpolated recall-precision curve.

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's overview paper.

3 TREC 2004 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 1 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 2004 tracks. See the track reports later in these proceedings for a more complete description of each track.

3.1 The genomics track

The genomics track was introduced as a "pre-track" in 2002. It is the first TREC track devoted to retrieval within a specific domain; one of the goals of the track is to see how exploiting domain-specific information improves retrieval effectiveness.

The 2004 genomics track contained an ad hoc retrieval task and three variants of a categorization task. The ad hoc task used a 10-year subset (1994–2003) of MEDLINE, a bibliographic database of the biomedical literature maintained by the US National Library of Medicine who donated the subset to the track. The subset used in the track contains about 4.5 million MEDLINE records (which include title and abstract as well as other bibliographic information) and is about 9GB of data. The 50 topics for the ad hoc task were derived from information needs obtained through interviews of biomedical researchers. Pools were created using one run from each of the 27 participating groups using a depth of 75. Relevance judgments were made by assessors with backgrounds in biology using a three-point scale of definitely relevant, probably relevant, and not relevant. Both definitely relevant and probably relevant were considered relevant when computing evaluation scores.

Domain knowledge was most frequently exploited by using resources such as the MeSH hierarchy (a controlled vocabulary used to index medical literature) to expand queries. Careful use of such resources appears to increase retrieval effectiveness, though some attempts to exploit such information decreased effectiveness relative to a generic baseline.

The genomics domain has a number of model organism database projects in which the literature regarding a specific organism (such as a mouse) is tracked and annotated with the function of genes and proteins. The classification tasks

Table 1: 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	2004
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	28
Web	—	—	—	—	—	—	—	17	23	30	23	27	18
Video	—	—	—	—	—	—	—	—	—	12	19	—	—
Novelty	—	—	—	—	—	—	—	—	—	—	13	14	14
Genomics	—	—	—	—	—	—	—	—	—	—	—	29	33
HARD	—	—	—	—	—	—	—	—	—	—	—	14	16
Robust	—	—	—	—	—	—	—	—	—	—	—	16	14
Terabyte	—	—	—	—	—	—	—	—	—	—	—	17	
Total participants	22	31	33	36	38	51	56	66	69	87	93	93	103

used in the 2004 track mimic some aspects of this curation process with the goal of eventually automating this now largely manual task. For the classification tasks, the track used the full text articles from a two-year span of three journals. This text was made available to the track through Highwire Press. The truth data for the tasks came from the actual annotation process carried out by the human annotators in the mouse genome informatics (MGI) system. Evaluation scores were computed using normalized utility measures.

As in the ad hoc task, many groups used MeSH terms as features to classify the documents. While these approaches were relatively effective, a subsequent analysis demonstrated the benefit was largely attributable to a single MeSH term: a baseline run that classified documents solely by the presence of the MeSH term *Mice* in the MEDLINE record of the document would have been the second best run submitted to the track for the triage classification task.

3.2 The HARD track

HARD stands for "High Accuracy Retrieval from Documents". The HARD track was started in TREC 2003 with the goal of improving retrieval performance, especially at the top of the ranked list, by targeting retrieval results to the specific searcher. To facilitate such targeting, the HARD track provides metadata in the topic statement. In addition, "clarification forms" provide a limited means of interaction between the system and the searcher.

The underlying task in the HARD track was an ad hoc retrieval task. The document set was a set of newswire/newspaper articles from 2003, including (English portions) of non-US papers. The collection is approximately 1500MB of text and contains approximately 650,000 articles. Topics were created at the Linguistic Data Consortium (LDC), and were originally released in standard TREC format (i.e., just title, description, and narrative fields). Once participants submitted baseline runs using the standard topics, they received the expanded version of the topics. There were 50 topics in the test set, though only 45 topics were used in the evaluation since five topics had no relevant documents.

The expanded version of the topics contained both a statement of the retrieval unit and the metadata. The retrieval

unit was always specified, and was either "passage" or "document". The "passage" specification meant retrieval systems should return pieces of documents, rather than full documents, as a response. The types of metadata in the TREC 2004 topics included familiarity, genre, geography, subject, and related text. The first three types affected the relevance of a text: a text that was on-topic but did not satisfy one of these metadata constraints was considered not relevant when using stringent relevance criteria. The subject metadata item contained the subject domain of the topic (for example, "sports", or "politics"); a document that did not meet this criterion was off-topic. The related text metadata provided some examples of relevant or on-topic text drawn from outside the test corpus. Different topics contained different kinds and amounts of metadata.

In addition to the information included in the expanded version of the topics, participants could collect information from the searcher (the assessor who created and judged the topic) using clarification forms. A clarification form was a single, self-contained HTML form created by the participating group and specific to a single topic. There were no restrictions on what type of data could be collected using a clarification form, but the searcher spent no more than three minutes filling out any one form.

Participants then made new runs using any combination of information from the expanded topics and 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. Retrieval results were evaluated both at the document level (for all 45 topics including those with retrieval unit "passage") using `trec_eval` and using passage level evaluation measures over just the 25 topics with retrieval unit "passage".

Sixteen groups submitted 135 runs to the HARD track. Most groups were able to exploit the additional information to improve effectiveness as compared to their baseline run, generally by performing some type of relevance feedback.

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 the tasks changed significantly between 2002 and 2003. This year's track used the same tasks as the 2003 track.

The basic task in the novelty track is as follows: given a topic and an ordered set of 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 and then identify which sentences contain new information.

Fifty new topics were created for the 2004 track. As in TREC 2003, 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* which contains approximately 1,033,000 documents and 3 gigabytes of text. The document set for a topic in the test set contained the 25 relevant documents selected by the assessor as well as 0 or more irrelevant documents. The documents in a set were ordered chronologically.

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

Task 1. Given the complete document set for a topic, identify all relevant and novel sentences.

Task 2. Given the relevant sentences in the complete document set, 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 documents.

Task 4. Given the relevant sentences in the complete document set, and the novel sentences in the first 5 documents, find the novel sentences in the remaining 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.

Fourteen groups submitted 183 runs to the novelty track, with tasks 1 and 2 having the greater participation. The inclusion of nonrelevant documents in the retrieved set appears to make task 1 much more challenging. In TREC 2003,

3	Hale Bopp comet		
	3.1	FACTOID	When was the comet discovered?
	3.2	FACTOID	How often does it approach the earth?
	3.3	LIST	In what countries was the comet visible on its last return?
	3.4	OTHER	

Figure 2: A sample QA track question series.

the best-performing systems for task 1 were roughly comparable to human performance as measured by scoring a second assessor's sentence selection against the primary assessor's choices. This year, the best systems' effectiveness was well below human performance. The particular topics used this year may also have been more difficult given that the absolute scores of TREC 2004 systems were lower than TREC 2003 scores for task 2 and task 2 is unaffected by nonrelevant documents.

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 TREC 2003 version of the track used a combined task where the test set of questions consisted of factoid, list, and definition questions. Each type of question was judged and scored separately, but the final score for a run was a weighted average of the component scores. The task in the 2004 track was similar in that the test set consisted of a mix of question types, and the final score was a weighted average of the components. The task was reorganized, however, such that the systems were to answer a series of factoid and list questions that each related to a common target, and then to respond with a list of "other" information about the target that was not covered by the previous questions in the series. This last question in the series is a more difficult variant of the definition questions in TREC 2003. This reorientation of the task requires systems to track context when answering questions, an important element of question answering that the track has not yet successfully incorporated [11].

The document set used in the track was the *AQUAINT Corpus of English News Text*. The test set consisted of 65 series of questions that together included 230 factoid questions, 56 list questions (one had to be removed from the evaluation due to no correct answers in the collection), and 65 Other questions (one had to be removed from the evaluation since it mistakenly went unjudged). Each of the questions was explicitly tagged as to what type of question it was and what series it belonged to. The target of the series was given as metadata for the whole series. An example series is given in figure 2.

The score for the factoid question component was accuracy, the percentage of factoid questions whose response was judged correct. The list and Other question components were each scored using average F, though the computation of the F score differed between the two components [12]. The final score for a run was computed as a weighted average of the three component scores: FinalScore $= .5$Accuracy $+ .25$AveListF $+ .25$AveOtherF.

Sixty-three runs from 28 different groups were submitted to the track. In general, the use of pronouns and anaphora in questions later in a series did not seem to pose a very serious challenge for the systems, in part because the target was the correct referent a large majority of the time. For most systems, the average score for the first question in a series was somewhat greater than the average score for a question that was not the first question in a series, but the difference was not great and is confounded by other effects (there are many fewer first questions to compute the average over, first questions in a series might be intrinsically easier questions, etc.).

The reorganization of the task into a set of question series had an unexpected benefit. The series proved to be an appropriate level of granularity for aggregating scores for an effective evaluation. The series is small enough to be meaningful at the task level since it represents a single user interaction, yet it is large enough to avoid the highly skewed score distributions exhibited by single questions. Computing a combined score for each series, and averaging the series scores, produces a QA task evaluation that more closely mimics classic document retrieval evaluation.

3.5 The robust track

The robust track looks to improve the consistency of retrieval technology by focusing on poorly performing topics. TREC 2004 was the second time the track was run. The initial track provided strong evidence that optimizing average effectiveness using the standard methodology and current evaluation measures further improves the effectiveness of the already-effective topics, sometimes at the expense of the poor performers. That track also showed that measuring poor performance is intrinsically difficult because there is so little signal in the sea of noise for a poorly performing topic. New measures devised for the TREC 2003 robust track do emphasize poorly performing topics, but because there is so little information, the measures are unstable.

The task in both years of the robust track was a classic ad hoc retrieval task. The TREC 2004 edition of the track used more topics than the 2003 edition in hopes of getting a more stable evaluation. In particular, the test set for 2004 consisted of 250 topics (one topic was dropped from the evaluation since it was judged to have no relevant documents). Two hundred of the topics were used in previous TREC tasks and 50 new topics were created for the track. To avoid needing new relevance judgments for the 200 old topics, an old document set was used: the set of documents on TREC disks 4 and 5 minus the *Congressional Record* documents.

The use of old topics had an additional motivation other than not needing new relevance judgments for those topics. Since the retrieval results from the previous TREC in which the topics were used are available, it is possible to select topics that are known to be challenging to a majority of retrieval systems. Fifty topics from among the 200 old topics were designated as being difficult. These topics were selected for the TREC 2003 track by choosing topics that had a low median average precision score and at least one high outlying score.

The retrieval results were evaluated using `trec_eval`, two measures introduced in the TREC 2003 track that emphasize poorly performing topics, and a new measure, geometric MAP, introduced in this year's track. The geometric MAP is a variant of the traditional MAP measure that uses a geometric mean rather than an arithmetic mean to average individual topic results. An analysis of the behavior of the geometric MAP measure suggests it gives appropriate emphasis to poorly performing topics while being more stable at equal topic set sizes.

The robust track received a total of 110 runs from 14 participants. All of the runs submitted to the track were automatic runs. The results indicate that the most promising approach to improving poorly performing topics is exploiting text collections other than the target collection, though the process must be carefully controlled to avoid making the results worse. The web was the collection most frequently used as an auxiliary collection.

An additional requirement in this year's track was for systems to submit a ranked list of the topics ordered by perceived difficulty. That is, the system assigned each topic a number from 1 to 250 where the topic assigned 1 was the topic the system believed it did best on, the topic assigned 2 was the topic the system believed it did next best on, etc. The purpose of the requirement was to see if systems can recognize whether a topic is difficult at run time, a first step toward doing special processing for difficult topics. While some systems were clearly better than others at predicting when a topic is difficult for that system, none of the systems were particularly good at the task. How much accuracy is required to make effective use of the predictions is still unknown.

3.6 The terabyte track

The terabyte track is a new track in 2004. The goal of the track is is to develop an evaluation methodology for terabyte-scale document collections. The track also provides an opportunity for participants to see how well their retrieval algorithms scale to much larger test sets than other TREC collections.

The document collection used in the track is the GOV2 collection, a collection of Web data crawled from Web sites in the .gov domain during early 2004. This collection contains a large proportion of the crawlable pages in .gov, including html and text, plus extracted text of pdf, word and postscript files. The collection is 426GB in size and contains approximately 25 million documents. The collection is smaller than a full terabyte due to the difficulty of obtaining and processing enough documents while allowing sufficient time for distributing the collection to participants. The collection will be expanded using data from other sources in future years. The current collection is at least an order of magnitude greater than the next-largest TREC collection.

The task in the track was a classic ad hoc retrieval task. The test set consisted of 50 topics created specifically for the track. While the document set consists of web pages, the topics were standard information-seeking requests, and

not navigational requests or topic distillation requests, for example. Systems returned the top 10,000 documents per topic so various evaluation strategies can be investigated. Participants also answered a series of questions about timing and resources required to produce the retrieval results.

Seventy runs from 17 different groups were submitted to the track. The top 85 documents per topic for two runs per group were added to the judgment pools. Initial analysis of the track results has revealed little difference in the relative effectiveness of different approaches when evaluated by MAP or by bpref, a measure created for evaluation environments where pools are known to be very incomplete [2]. There are a variety of reasons why this might be so: it may mean that current pooling practices are adequate for collections of this size, or that the runs submitted to the terabyte track happened to retrieve a sufficient set of relevant documents, or that the terabyte topics happened to be particularly narrow, and so forth. The terabyte track will continue in TREC 2005 to examine these questions.

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. Previous TREC web tracks had separately investigated topic distillation, named page finding, and home page finding tasks [4]. Since web search engines must process these types of searches (among others) without explicit knowledge of which type of search is wanted, this year's web task combined them into a single task.

For a topic distillation search a system is to return a list of entry points for good websites principally devoted to the topic. Since there are only a few good websites for any particular topic, there are only a few key ("relevant") pages for a topic distillation search. The emphasis is on returning entry pages rather than pages containing relevant information themselves since a result list of homepages provides a better overview of the coverage of a topic in the collection.

Named page and home page finding searches are similar to each other in that both are known-item tasks where the system is to return a particular page. For home page finding, the target page is the home page of the entity in the topic. For named page finding, a particular page is sought, but that page is not an entry point to a site (e.g., "1040 tax form").

For the TREC 2004 task, participants received a set of 225 title-only topics such as "West Indian manatee information" and "York county". The assessor specified which type of search was intended when the topic was created, but the test set did not include this information. Systems returned a ranked list of up to 1000 pages per topic. During judging, the assessors made binary judgments as to whether a page was appropriate with respect to the intended task. That is, the pages returned for topics whose search type was topic distillation were judged relevant if the page was a key entry page and not relevant otherwise. For the named page finding and home page finding topics, a page was judged relevant if and only if the page was the target page (or a mirror/alias of the target page). The runs were evaluated using MAP, which is equivalent to the mean reciprocal rank (MRR) measure for known-item searches.

The track used the .GOV collection created for the TREC 2002 web track and distributed by CSIRO. 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.

In addition to the search task, the track also contained a classification task in which the goal was simply to label each of the 225 test topics as to what type of search was intended.

Eighteen groups submitted a total of 83 runs to the track. Nine of the runs were classification task runs. The retrieval results showed that systems are able to obtain effective overall retrieval without having to classify the queries by type. That is, groups were able to devise a single technique that performed well for home page, named page, and distillation topics. These techniques were not based solely on the text of a page, but also needed to exploit some sort of web information such as link structure or anchor text. Systems that did attempt to classify topics were generally able to do so, with most classification errors confusing named page and home page topics.

4 The Future

A significant fraction of the time of one TREC workshop is spent in planning the next TREC. A majority of the TREC 2004 tracks will continue in TREC 2005, including the genomics, HARD, QA, robust, and terabyte tracks. As described in the web track overview paper, the web track as such will end, with a new enterprise track taking its place. The goal of the enterprise track is to study enterprise search—satisfying a user who is searching the data of

an organization to accomplish some task. The novelty track will also end. Finally, a new track, the spam track, will be introduced in TREC 2005. The goal of the spam track is to provide a standard evaluation of current and proposed spam filtering approaches, thereby laying the foundation for the evaluation of more general email filtering and retrieval tasks.

Acknowledgements

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

References

[1] Chris Buckley. trec_eval IR evaluation package. Available from `http://trec.nist.gov/trec_eval/`.

[2] Chris Buckley and Ellen M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the Twenty-Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 25–32, 2004.

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

[4] Nick Craswell, David Hawking, Ross Wilkinson, and Mingfang Wu. Overview of the TREC 2003 web track. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*, pages 78–92, 2004.

[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.

[11] Ellen M. Voorhees. Overview of the TREC 2001 question answering track. In *Proceedings of the Tenth Text REtreival Conference (TREC 2001)*, pages 42–51, 2002.

[12] Ellen M. Voorhees. Overview of the TREC 2003 question answering track. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*, pages 54–68, 2004.

[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.

Table 2: Organizations participating in TREC 2004

Alias-i, Inc.	Arizona State University
California State U. San Marcos	Carnegie Mellon University
Chinese Academy of Sciences (3 groups)	Chinese University of Hong Kong
Clairvoyance Corporation	CL Research
Columbia University	ConverSpeech LLC & Stanford SGD
CSIRO	Dalhousie University
Decision Aid team-LAMSADE	Dublin City University
Etymon	Fondazione Ugo Bordoni
Fudan University (2 groups)	German University in Cairo
Hong Kong Polytechnic University	Hummingbird
IBM India Research Lab	IBM Research Lab Haifa
IBM T.J. Watson Research Center	IDA/CCS/NSA
IIT Information Retrieval Lab	Indiana University (2 groups)
IRIT/SIG	ITC-irst
Johns Hopkins University	Korea University
Language Computer Corporation	LexiClone
Macquarie University	Massachusetts Institute of Technology
Max-Planck-Institute for Computer Science	Meiji University
Microsoft Research Asia	Microsoft Research Ltd
Monash University	National Central University
National Security Agency	National Taiwan University
National University of Singapore	National U. of Singapore & Singapore-MIT Alliance
NLM-UMaryland Team	Oregon Health and Science University
PATOLIS Corporation	Peking University
Queens College, CUNY	RMIT University
Rutgers University (2 groups)	Saarland University
Sabir Research, Inc.	Shanghai JiaoTong University
SUNY at Buffalo	Tarragon Consulting Corporation
The MITRE Corporation	The Robert Gordon University
The University of Melbourne	TNO & Erasmus MC
Tsinghua University (2 groups)	UC Berkeley
U. Hospital Geneva & Swiss Federal Inst. of Tech.	Universidade de Lisboa Campo Grande
Universitat Politcnica de Catalunya	Universit Paris Sud
University of Alaska Fairbanks	University of Alberta
University of Amsterdam	University of Chicago
University of Cincinnati	University of Edinburgh
University of Edinburgh & Sydney	University of Glasgow
University of Illinois at Chicago	University of Illinois at Urbana-Champaign
University of Iowa	University of Lethbridge
University of Limerick	University of Maryland UMIACS
University of Massachusetts	University of Michigan
University of North Carolina	University of North Texas
University of Padova	University of Pisa
University of Sheffield	University of Tampere
University of Tokyo	University of Twente
University of Wales, Bangor	University of Waterloo (2 groups)
University of Wisconsin	USC-Information Sciences Institute
Virginia Tech	York University