Overview of TREC 2005

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

1 Introduction

The fourteenth Text REtrieval Conference, TREC 2005, was held at the National Institute of Standards and Technology (NIST) 15 to 18 November 2005. The conference was co-sponsored by NIST and the US Department of Defense Advanced Research and Development Activity (ARDA). TREC 2005 had 117 participating groups from 23 different countries. Table 2 at the end of the paper lists the participating groups.

TREC 2005 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 2005 contained seven areas of focus called "tracks". Two tracks focused on improving basic retrieval effectiveness by either providing more context or by trying to reduce the number of queries that fail. Other tracks explored tasks in question answering, detecting spam in an email stream, enterprise search, search on (almost) terabyte-scale document sets, and information access within the genomics domain. The specific tasks performed in each of the tracks are summarized in Section 3 below.

This paper serves as an introduction to the research described in detail in the remainder of the proceedings. 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 MEDLINE record, a web page, or an email message.

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 TREC 2005 tracks included some sort of an ad hoc search task.

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 task in the terabyte track and the known-item task within the enterprise track are examples of known-item search tasks.

In a *categorization* task, the system is responsible for assigning a document to one or more categories from among a given set of categories. In the spam track, deciding whether a given mail message is spam is a categorization task; the genomics track had several categorization tasks in TREC 2005 as well.

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. In addition, the expert-finding task in the enterprise track is a type of question answering task in that the system response to an expert-finding search is a set of people, not documents.

2.1 Test collections

Text retrieval has a long history of using retrieval experiments on test collections to advance the state of the art [2, 6], 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. We call the result of a retrieval system executing a task on a test collection a run.

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 2 to 3 gigabytes of text and 500 000 to 1 000 000 documents). The document sets used in various tracks have been smaller and larger depending on the needs of the track

```
<num> Number:  758
<title> Embryonic stem cells
<desc> Description:  What are embryonic stem cells, and what
restrictions are placed on their use in research?
<narr> Narrative:  Explanation of the nature of embryonic stem cells is
relevant.  Their usefulness in research is relevant.  Sources for them
and restrictions on them also are relevant.
```

Figure 1: A sample TREC 2005 topic from the terabyte track test set.

and the availability of data. The terabyte track was introduced in TREC 2004 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. High-level structures within each document are tagged using SGML or XML, 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 terabyte 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 ("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 ad hoc 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 [4]. 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 [7].

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 [5] 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. Pooling is valid when enough relevant documents are found to make the resulting judgment set approximately complete and unbiased.

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 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 [10]. 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 [9]. 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 [3].

The uniquely-retrieved-relevant-documents test can fail to indicate a problem with a collection if all the runs that contribute to the pool share a common bias—preventing such a common bias is why a diverse run set is needed for pool construction. While it is not possible to prove that no common bias exists for a collection, no common bias has been demonstrated for any of the TREC collections until this year. The retrieval test collection built in the TREC 2005 HARD and robust tracks has a demonstrable bias toward documents that contain topic title words. That is, a very large fraction of the known relevant documents for that collection contain many topic title words despite the fact that documents with fewer topic title words that would have been judged relevant exist in the collection. (Details are given in the robust track overview paper later in this volume [8].)

The bias results from pools that are shallow *relative to the number of documents in the collection*. Many otherwise diverse retrieval methodologies sensibly rank documents that have lots of topic title words before documents containing fewer topic title words since topic title words are specifically chosen to be good content indicators. But a large document set will contain many documents that include topic title words. To produce an unbiased, reusable collection, traditional pooling requires sufficient room in the pools to exhaust the spate of title-word documents and allow documents that are not title-word-heavy to enter the pool. The robust track contained one run that did not concentrate on topic title words and could thus demonstrate the bias in the other runs. No such "smoking-gun" run exists for the collections built in the TREC 2004 and 2005 terabyte track, but a similar bias must surely exist in these collections. The biased collections are still useful for comparing retrieval methodologies that have a matching bias (and the results of the 2005 tracks are valid since the runs were used to build the collections), but results on these collections need to be interpreted judiciously when comparing methodologies that do not emphasize topic title words.

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 (number-retrieved-and-relevant/number-retrieved), while recall is the proportion of relevant documents that are retrieved (number-retrieved-and-relevant/number-relevant). 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 at ten documents retrieved less than 1.0 regardless of how the documents are ranked. Similarly, a topic with more than ten relevant documents must have a recall score

at ten documents retrieved less than 1.0. 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 interpolated recall-precision curve and mean average precision (non-interpolated) 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 (MAP) 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 2005 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 a smaller percentage of the tracks.

This section describes the tasks performed in the TREC 2005 tracks. See the track reports later in these proceedings for a more complete description of each track.

3.1 The enterprise track

TREC 2005 was the first year for the enterprise track, which is an outgrowth of previous years' web track tasks. The purpose of the track is to study enterprise search: satisfying a user who is searching the data of an organization to complete some task. Enterprise data generally consists of diverse types such as published

Table 1: Number of participants per track and total number of distinct participants in each TREC

	TREC													
Track	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
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	—	—	—	—	—	—	—	—	
Merging	—	—	—	3	3	—	—	—	—	—	—	—	—	
Filtering	—	—	—	4	7	10	12	14	15	19	21	—	—	—
Chinese	—	—	—	—	9	12	—	—	—	—	—	—	—	
NLP	—	—	—	—	4	2	—	—	—	—	—	—	—	
Speech	—	—	—	—	—	13	10	10	3	—	—	—	—	
XLingual	—	—	—	—	—	13	9	13	16	10	9	—	—	—
High Prec	—	—	—	—	—	5	4	—	—	—	—	—	—	
VLC	—	—	—	—	—	—	7	6	—	—	—	—	—	
Query	—	—	—	—	—	—	2	5	6	—	—	—	—	
QA	—	—	—	—	—	—	—	20	28	36	34	33	28	33
Web	—	—	—	—	—	—	—	17	23	30	23	27	18	—
Video	—	—	—	—	—	—	—	—	—	12	19	—	—	—
Novelty	—	—	—	—	—	—	—	—	—	—	13	14	14	—
Genomics	—	—	—	—	—	—	—	—	—	—	—	29	33	41
HARD	—	—	—	—	—	—	—	—	—	—	—	14	16	16
Robust	—	—	—	—	—	—	—	—	—	—	—	16	14	17
Terabyte	—	—	—	—	—	—	—	—	—	—	—	—	17	19
Enterprise	—	—	—	—	—	—	—	—	—	—	—	—	—	23
Spam	—	—	—	—	—	—	—	—	—	—	—	—	—	13
Participants	22	31	33	36	38	51	56	66	69	87	93	93	103	117

reports, intranet web sites, and email, and the goal is to have search systems deal seamlessly with the different data types.

The document set used in the track was the W3C Test collection (see `http://research.microsoft.com/users/nickcr/w3c-summary.html`). This collection, created by Nick Craswell, was created from a crawl of the World-Wide Web Consortium web site and includes email discussion lists, web pages, and the extracted text from documents in various formats (such as pdf, postscript, Word, Powerpoint, etc.). Because of the technical nature of the documents, and hence the topics that could be asked against those documents, topic development and relevance judging for the enterprise track were performed by the track participants.

The track contained three search tasks: a known-item search for a particular message in the email lists archive; an ad hoc search for the set of messages that pertain to a particular discussion covered in the email lists; and a search-for-experts task. The motivation for the expert-finding task is being able to determine who the correct contact person for a particular matter is in a large organization. For the track task, the topics were the names of W3C working groups (e.g., “Web Services Choreography”), and the correct answers were assumed to be the members of that particular working group. Systems were to return the names of the people themselves, not documents that stated the people were members of the particular working group.

Twenty-three groups participated in the enterprise track, 14 groups in the discussion search task, 9 groups in the expert-finding task, and 17 groups in the known-item search task. While groups generally attempted to exploit the thread structure and quoted material in the email tasks, the effectiveness of the searches was

generally dominated by traditional content factors. Thus, more work is needed to understand how best to support discussion search.

3.2 The genomics track

The goal of genomics track is to provide a forum for evaluation of information retrieval systems in the genomics domain. It is the first TREC track devoted to retrieval within a specific domain, and thus a subgoal of the track is to explore how exploiting domain-specific information improves retrieval effectiveness. As in TREC 2004, the 2005 genomics track contained an ad hoc retrieval task and a categorization task.

The document set for the ad hoc task was the same corpus as was used in the 2004 genomics ad hoc task, a 10-year subset (1994 to 2003) of MEDLINE, the bibliographic database of biomedical literature maintained by the US National Library of Medicine. The corpus 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 topics were developed from interviews from real biologists who were asked to fill in a "generic topic template" or GTT. The GTTs were used to produced more structured topics than traditional TREC topics so systems could make better use of resources such as ontologies and databases. The 50 test topics contain ten instances for each of the following five GTTs, where the underlined portions represent the template slots:

1. Find articles describing standard methods or protocols for doing some sort of experiment or procedure.
2. Find articles describing the role of a gene involved in a given disease.
3. Find articles describing the role of a gene in a specific biological process.
4. Find articles describing interactions (e.g., promote, suppress, inhibit, etc.) between two or more genes in the function of an organ or in a disease.
5. Find articles describing one or more mutations of a given gene and its biological impact.

For example, a topic derived from the mutation GTT might be *Provide information about Mutation of Ret in thyroid function.* 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.

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 task used in the 2005 track focused on one of the tasks in this curation process, the "document triage" task. The document triage task is essentially a filtering task in which a document passes through the filter only if it should receive more careful examination with respect to a specific category. Four different categories were used in the track: Gene Ontology (GO) annotation, tumor biology, embryologic gene expression, and alleles of mutant phenotypes. The document set was the same document set used in the TREC 2004 genomics categorization task, the full text articles from a two-year span of three journals made available to the track through Highwire Press. The truth data for the task came from the actual annotation process carried out by the human annotators in the mouse genome informatics (MGI) system.

The genomics track had 41 participants, with 32 groups participating in the ad hoc search task and 19 participating in the categorization task. Retrieval effectiveness was roughly equivalent across the different topic types in the ad hoc search task. In contrast, system effectiveness was strongly dependent on the specific category in the triage task.

3.3 The HARD track

The goal of the "High Accuracy Retrieval from Documents" (HARD) track is improving retrieval system effectiveness by personalizing the search to the particular user. For the 2005 track, the method for obtaining information about the user was through clarification forms, a limited type of interaction between the system and the searcher.

The underlying task in the HARD track is an ad hoc retrieval task. Participants first submit baseline runs using the topic statements as is. They may then collect information from the searcher (the assessor who judged the topic) using clarification forms. A clarification form is 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. An example use of a clarification form is to ask the searcher which of a given set of terms are likely to be good search terms for the topic. Finally, participants make new runs using the information gathered from clarification forms.

The same document set, topics, and hence relevance judgments were used in both the HARD and robust tracks. The document set was the *AQUAINT Corpus of English News Text* (LDC catalog number LDC2002T31, see `www.ldc.upenn.edu`). The 50 test topics were a subset of the topics used in previous TREC robust tracks, which had been demonstrated to be difficult topics for systems when used on the TREC disks 4&5 document set. Relevance judgments were performed by NIST assessors based on pools of both HARD and robust runs.

The motivation for sharing the test collection between the two tracks was partly financial—NIST did not have the resources to create a separate collection for each track—but sharing also had technical benefits as well. One hypothesis as to why previous years' HARD tracks did not demonstrate as large a difference in effectiveness between baseline and final runs as expected was that many of the topics in those test sets did not really need clarification. Using topics that had been shown to be difficult in the past was one way of constructing a test set that had room for improvement. The design also allows direct comparison between the largely automatic methods used in the robust track with the limited searcher feedback of the HARD track.

Sixteen groups participated in the HARD track. The majority of runs that used clarification forms did improve over their corresponding baseline runs, and a few such runs showed noticeable improvement. While this supports the hypothesis that some forms of limited user interaction can be effective in improving retrieval effectiveness, many questions regarding how best to use it remain. Note, for example, that the best automatic run from the robust track (that used no interaction) was more effective than any of the automatic runs from the HARD track.

3.4 The question answering (QA) track

The goal of the question answering track is to develop systems that return actual answers, as opposed to ranked lists of documents, in response to a question. The main task in the TREC 2005 track was very similar to the TREC 2004 task, though there were additional tasks as well in TREC 2005.

The questions in the main task were organized into a set of series. A series consisted of a number of "factoid" (questions with fact-based, short answers) and list questions that each related to a common, given target. The final question in a series was an explicit "Other" question, which systems were to answer by retrieving information pertaining to the target that had not been covered by earlier questions in the series. The score for a series was computed as a weighted average of the scores for the individual questions that

comprised it, and the final score for a run was the mean of the series scores.

The document set used in the track was again the AQUAINT corpus. The test set consisted of 75 series of questions where the target was either a person, an organization, an entity to be defined (e.g., "kudzu"), or an event. Events were new to the TREC 2005 task.

One of the concerns expressed at both the SIGIR 2004 IR4QA workshop and the QA track workshop at the TREC 2004 meeting was a desire to build infrastructure that would allow a closer examination of the role document retrieval techniques play in supporting QA technology. To this end, participants in the main task were required to submit a document ranking of the documents their system used in answering the question for each of 50 individual questions (not series). While not all QA systems produce a ranked list of documents as an initial step, some ranking (even if it consisted of only a single document) was still required. The submitted document rankings were pooled as in a traditional ad hoc task, and NIST assessors judged the pools using "contains an answer to the question" as the definition of relevant. The judged pools thus give the number of instances of correct answers in the collection, a statistic not computed for other QA test sets. The ranked lists will also support research on whether some document retrieval techniques are better than others in support of QA.

The relationship task was an optional second task in the track. The task was based on a pilot evaluation that was run in the context of the ARDA AQUAINT program (see `http://trec.nist.gov/data/qa/add_qaresources.html`). AQUAINT defined a relationship as the ability of one entity to influence another, including both the means to influence and the motivation for doing so. Eight spheres of influence were noted, including financial, movement of goods, family ties, communication pathways, organizational ties, co-location, common interests, and temporal. Systems were given a topic statement that set the context for a final question asking about one of the types of influence. The system response was a set of "information nuggets" that provided the evidence (or lack thereof) for the relationship hypothesized in the question. The relationship task test set contained 25 topics. Submissions to the relationship task were allowed to be either automatic (no manual processing at all) or manual.

Thirty-three groups participated in the main task, including three groups that performed only the document ranking task. Six groups participated in the relationship task as well. The document ranking task results demonstrated only a weak correlation between the effectiveness of the initial document ranking as measured by R-precision and the ability of the system to answer factoid questions.

3.5 The robust track

The robust track looks to improve the consistency of retrieval technology by focusing on poorly performing topics. Previous editions of the track have demonstrated that average effectiveness masks individual topic effectiveness, and that optimizing standard average effectiveness usually harms the already ineffective topics.

The task in the track is an ad hoc retrieval task where effectiveness is measured as a function of worst-case behavior. Measures of poor performance used in earlier tracks were problematic because they are relatively unstable when used with as few as 50 to 100 topics. A new measure developed during the final analysis of the TREC 2004 robust track results appears to give appropriate emphasis to poorly performing topics in addition to being stable with as few as 50 topics. This "gmap" measure is based on a geometric, rather than arithmetic, mean of average precision over a set of topics, and was the main effectiveness measure used in this year's track.

As discussed in the HARD track section, the HARD and robust tracks used the same test collection in 2005. The collection consists of the AQUAINT document set and 50 topics that had been used in previous

years' robust tracks. The 50 topics were topics that had low median effectiveness (across TREC submissions) when run against TREC disks 4&5 and are therefore considered difficult topics. The topics were selected from a larger set by choosing only those topics that had at least three relevant documents in the AQUAINT collection as judged by NIST assessors. Different assessors judged the topics this year against the AQUAINT document set from those that initially judged the topics against the disks 4&5 collection.

As in the robust 2004 track, a second requirement in the track was for systems to submit a ranked list of the topics ordered by perceived difficulty. A system assigned each topic a number from 1 to 50 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. This task is motivated by the hope that systems will eventually be able to use such predictions to do topic-specific processing. The quality of a prediction is measured using the area between two curves each of which plots the MAP score computed over all topics except the run's worst X topic. X ranges from 0 (so, all topics are included) to 25 (so, the average is computed over the best half of the topics). In one curve, the worst topics are defined from the run's predictions, while in the second curve the worst topics are defined using the actual average precision scores.

Seventeen groups participated in the robust track. As in previous robust tracks, the most effective strategy was to expand queries using terms derived from resources external to the target corpus. The relative difficulty of different topics, as measured by the average score across runs, differed between the disks 4&5 collection and the AQUAINT collection.

3.6 The spam track

The spam track is a second new track in 2005. The immediate goal of the track is to evaluate how well systems are able to separate spam and ham (non-spam) when given an email sequence. Since the primary difficulty in performing such an evaluation is getting appropriate corpora, longer term goals of the track are to establish an architecture and common methodology for a network of evaluation corpora that would provide the foundation for additional email filtering and retrieval tasks.

There are a number of reasons why obtaining appropriate evaluation corpora is difficult. Obviously making real email streams public is not an option because of privacy concerns. Yet creating artificial corpora is also difficult. Most of the modifications to real email streams that would protect the privacy of the recipients and senders also compromises the information used by classifiers to distinguish between ham and spam. The track addressed this problem by having several corpora, some public and some private. The track also made use of a test jig developed for the track that takes an email stream, a set of ham/spam judgments, and a classifier, and runs the classifier on the stream reporting the evaluation results of that run based on the judgments.

Track participants submitted their classifiers to NIST. Track coordinator Gord Cormack and his colleagues at the University of Waterloo used the jig to evaluate the submitted classifiers on the private corpora. In addition, the participants used the jig themselves to evaluate the same classifiers on the public corpora and submitted the raw results from the jig on that data back to NIST.

Several measures of the quality of a classification are reported for each combination of corpus and classifier. These measures include

ham misclassification rate: the fraction of ham messages that are misclassified as spam;

spam misclassification rate: the fraction of spam messages that are misclassified as ham;

ham/spam learning curve : error rates as a function of the number of messages processed;

ROC curve: ROC (Receiver Operating Characteristic) curve that shows the tradeoff between ham/spam misclassification rates;

ROC ham/spam tradeoff score: the area above an ROC curve. This is equivalent to the probability that the spamminess score of a random ham message equals or exceeds the spamminess score of a random spam message.

Thirteen groups participated in the spam track. In addition, the organizers ran several existing spam classifiers on the various corpora and report those results as well in the spam track section of Appendix A. On the whole, the filters were effective, though each had a misclassification rate that was observable on even the smallest corpus (8000 messages). Steady-state misclassification rates were reached quickly and were not dominated by early errors, suggesting that the filters would continue to be effective in actual use.

3.7 The terabyte track

The goal of the terabyte track 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 was the same collection as was used in the TREC 2004 track: 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 contains approximately 25 million documents and is 426 GB. While smaller than a full terabyte, this collection is at least an order of magnitude greater than the next-largest TREC collection. The collection is distributed by the University of Glasgow, see `http://ir.dcs.gla.ac.uk/test_collections/`.

The track contained three tasks, a classic ad hoc retrieval task, an efficiency task, and a named-page-finding task. Manual runs were encouraged for the ad hoc task since manual runs frequently contribute unique relevant documents to the pools. The efficiency and named page tasks required completely automatic processing only.

The ad hoc retrieval task used 50 information-seeking topics created for the task by NIST assessors. While systems returned the top 10 000 documents per topic so various evaluation strategies can be investigated, pools were created from the top 100 documents per topic.

The efficiency task was an extension of the ad hoc task and was designed as a way of comparing the efficiency and scalability of systems given participants all used their own (different) hardware. The "topic" set was a sample of 50 000 queries mined from web search engine logs plus the title fields of the 50 topics used in the ad hoc task. Systems returned a ranked list of the top 20 documents for each query plus reported timing statistics for processing the entire query set. To measure the effectiveness of the efficiency runs, the results for the 50 queries that corresponded to the ad hoc topic set were added to the ad hoc pools and judged by the NIST assessors during the ad hoc judging.

Since the document set used in the track is a crawl of a cohesive part of the web, it can support investigations into tasks other than information-seeking search. One of the tasks that had been performed in the web track in earlier years was a named-page finding task, in which the topic statement is a short description of a single page (or very small set of pages), and the goal is for the system to return that page at rank one. The terabyte named page task repeated this task using the GOV2 collection.

Nineteen groups participated in the track, including 18 groups participating in the ad hoc task, 13 groups in the efficiency task, and 13 groups in the named page task. While there was a wide spread in both efficiency

and effectiveness across groups, runs submitted by the same group do demonstrate that devoting more query-processing time can increase retrieval effectiveness.

4 The Future

A significant fraction of the time of one TREC workshop is spent in planning the next TREC. Two of the TREC 2005 tracks, the HARD and robust tracks, will be discontinued as tracks in TREC 2006. A variant of the HARD track's clarification form task will continue as a subtask of the question answering track; the evaluation methodology developed in the robust track will be incorporated in other tracks with ad hoc tasks. The discontinued tracks make room for two new tracks to begin in TREC 2006. The blog track will explore information seeking behavior in the blogosphere. The goal in the legal track is to develop search technology that meets the needs of lawyers to engage in effective discovery in digital document collections.

Acknowledgements

Special thanks to the track coordinators who make the variety of different tasks addressed in TREC possible. The analysis of the pools from the HARD/robust tracks and the terabyte track was done in collaboration with Chris Buckley and Ian Soboroff.

References

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

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

[3] 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.

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

[5] 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.

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

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

[8] Ellen M. Voorhees. Overview of the TREC 2005 robust retrieval track. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*, 2005.

[9] 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`.

[10] 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 2005

Academia Sinica	Arizona State University (2 groups)
University of Alaska Fairbanks	Beijing University of Posts and Telecommunications
Breyer, Laird	Chinese Academy of Sciences (3 groups)
CL Research	Carnegie Mellon University (2 groups)
Coveo	CSIRO ICT Centre
California State University San Marcos	The Chinese University of Hong Kong
CRM114	Dalhousie University
DaLian University of Technology	OOO Datapark
DFKI GmbH (Saarland University)	Drexel University
Dublin City University	Ecole des Mines de Saint-Etienne
Erasmus MC	Fudan University (2 groups)
Harbin Institute of Technology	The Hong Kong Polytechnic University
Hummingbird	IBM Research Lab Haifa
IBM India Research Laboratory	IBM Almaden Research Center
IBM T.J. Watson (3 groups)	Institute for Infocomm Research
Illinois Institute of Technology	Indiana University
Institut de Recherche en Informatique de Toulouse	The Johns Hopkins University
Jozef Stefan Institute	LangPower Computing, Inc.
Language Computer Corporation	LexiClone
LowLands Team	Macquarie University
Massey University	Max-Planck Institute for Computer Science
Meiji University	Microsoft Research
Microsoft Research Asia	Microsoft Research Ltd
Massachusetts Institute of Technology	The MITRE Corporation
Monash University	National Library of Medicine - University of Maryland
National Library of Medicine (Wilbur)	National Security Agency
National Taiwan University	National University of Singapore
Oregon Health & Science University	Peking University
Pontifi cia Universidade Catolica Do Rio Grande Do Sul	Queen Mary University of London
Queens College, CUNY	Queensland University of Technology
Queen's University	RMIT University:
Rutgers University (2 groups)	Sabir Research, Inc.
SAIC OIS	Simon Fraser University
SUNY Buffalo	SUNY Stony Brook
TNO and Erasmus MC	Tokyo Institute of Technology
Tsinghua University	University of Albany
University of Amsterdam (2 groups)	University of Central Florida
University College Dublin	University of Colorado School of Medicine
University of Duisburg-Essen	U. of Edinburgh and U. of Sydney
University of Geneva	University of Glasgow
University of Illinois at Chicago	University of Illinois at Urbana-Champaign
University of Iowa	University of Limerick
University of Magdeburg	University of Maryland
University of Massachusetts	The University of Melbourne
The University of Michigan-Dearborn	Universit degli Studi di Milano
University of North Carolina	Universite de Neuchatel
University of North Texas	University of Padova
Universite Paris-Sud (2 groups)	Universitat Politcnica de Catalunya
University of Pisa	University of Pittsburgh
University of Sheffi eld	University of Strathclyde
University of Tampere	The University of Tokyo
University of Twente	University of Waterloo (2 groups)
University of Wisconsin	York University