Overview of TREC 2007

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

1 Introduction

The sixteenth Text REtrieval Conference, TREC 2007, was held at the National Institute of Standards and Technology (NIST) November 6–9, 2007. The conference was co-sponsored by NIST and the Intelligence Advanced Research Projects Activity (IARPA). TREC 2007 had 95 participating groups from 18 countries. Table 2 at the end of the paper lists the participating groups.

TREC 2007 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 2007 contained seven areas of focus called “tracks”. Six of the tracks ran in previous TRECs and explored tasks in question answering, blog search, detecting spam in an email stream, enterprise search, search in support of legal discovery, and information access within the genomics domain. A new track called the million query track investigated techniques for building fair retrieval test collections for very large corpora.

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 blog post, an email message, or an invoice.

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 precedent in case law, and analysts searching archived news reports for particular events. A retrieval system's response to an ad hoc search is generally an ordered list of documents sorted such that documents the system believes are more likely to satisfy the information need are ranked before documents it believes are less likely to satisfy the need. The tasks within the million query and legal tracks are examples of ad hoc search tasks. The feed task in the blog track is also an ad hoc search task, though in this case the documents to be ranked are entire blogs rather than blog postings.

In a *categorization* task, the system is responsible for assigning a document to one or more categories from among a given set of categories. Deciding whether a given mail message is spam is one example of a categorization task. The polarity task in the blog track, in which opinions were determined to be pro, con or both, is a second example.

Information retrieval has traditionally focused on returning entire documents in response to a query. This emphasis is both a reflection of retrieval systems' heritage as library reference systems and an acknowledgement of the difficulty of returning more specific responses. Nonetheless, TREC contains several tasks that do focus on more specific responses. In the question answering track, systems are expected to return precisely the answer; the system response to a query in the expert-finding task in the enterprise track is a set of people; and the task in the genomics track explores the trade-offs between different granularities of responses (whole documents, passages, and aspects).

2.1 Test collections

Text retrieval has a long history of using retrieval experiments on test collections to advance the state of the art [4, 8], 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 initial TREC test collections contain 2 to 3 gigabytes of text and 500,000 to 1,000,000 documents. While the document sets used in various tracks throughout the years have been smaller and larger depending on the needs of the track and the availability of data, the general trend has been toward ever-larger document sets to enhance the realism of the evaluation tasks. Similarly, the initial TREC document sets consisted mostly of newspaper or newswire articles, but later document sets have included a much broader spectrum of

```
<num> Number:  951
<title> Mutual Funds
<desc> Description:  Blogs about mutual funds performance and trends.
<narr> Narrative:  Ratings from other known sources (Morningstar) or
relative to key performance indicators (KPI) such as inflation, currency
markets and domestic and international vertical market outlooks.  News
about mutual funds, mutual fund managers and investment companies.
Specific recommendations should have supporting evidence or facts linked
from known news or corporate sources.  (Not investment spam or pure,
uninformed conjecture.)
```

Figure 1: A sample TREC 2007 topic from the blog track feed task.

document types (such as recordings of speech, web pages, scientific documents, blog posts, email messages, and business documents). Each document is assigned an unique identifier called the DOCNO. For most document sets, high-level structures within a document are tagged using a mark-up language such as SGML or HTML. In keeping with the spirit of realism, the text is kept as close to the original as possible.

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. What is now considered the "standard" format of a TREC topic statement—a topic id, a title, a description, and a narrative—was established in TREC-5 (1996). But topic formats vary in support of the task. The spam track has no topic statement at all, for example, and the topic statements used in the legal track contain much more information as might be available from a negotiated request to produce. An example topic taken from this year's blog track feed task is shown in figure 1.

The different parts of the traditional topic statements allow researchers to investigate the effect of different query lengths on retrieval performance. The description ("desc") field is generally a one sentence description of the topic area, while the narrative ("narr") gives a concise description of what makes a document relevant. The "title" field has served different purposes in different years. In TRECs 1–3 the field is simply a name given to the topic. In later ad hoc collections (ad hoc topics 301 and following), the field consists of up to three words that best describe the topic. For some of the test collections where topics were suggested by queries taken from web search engine logs, the title field contains the original query (sometimes modified to correct spelling or similar errors).

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 topics are generally constructed specifically for the task they are to be used in. When outside resources such as web search engine logs are used as a source of topics the sample selected for inclusion in the test set is vetted to insure there is a reasonable match with the document set (i.e., neither too many nor too few relevant documents). Topics developed at NIST are created by the NIST *assessors*, the set of people hired to both create topics and make relevance judgments. Most of the NIST assessors are retired intelligence analysts. The assessors receive track-specific training by NIST staff for both topic development and relevance assessment.

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 [6]. 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 [9].

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 infeasible. For example, with one million documents and assuming one judgment every 15 seconds (which is *very* fast), it would take approximately 4100 hours to judge a single topic. Thus by necessity TREC collections are created by judging only a subset of the document collection for each topic and then estimating the effectiveness of retrieval results from the judged sample.

The technique most often used in TREC for selecting the sample of documents for the human assessor to judge is pooling [7]. In pooling, the top results from a set of runs are combined to form the pool and only those documents in the pool are judged. Runs are subsequently evaluated assuming that all unpooled (and hence unjudged) documents are not relevant. In more detail, the TREC pooling process proceeds 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 (frequently $X = 100$) documents per topic are added to the topics' pools. 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 critical factor in pooling is that unjudged documents are assumed to be not relevant when computing traditional evaluation scores. This treatment is a direct result of the original premise of pooling: that by taking top-ranked documents from sufficiently many, diverse retrieval runs, the pool will contain the vast majority of the relevant documents in the document set. If this is true, then the resulting relevance judgment sets will be "essentially complete", and the evaluation scores computed using the judgments will be very close to the scores that would have been computed had complete judgments been available.

Various studies have examined the validity of pooling's premise in practice. Harman [5] and Zobel [10] independently showed that early TREC collections in fact had unjudged documents that would have been judged relevant had they been in the pools. But, importantly, the distribution of those "missing" relevant documents was highly skewed by topic (a topic that had lots of known relevant documents had more missing relevant), and uniform across runs. Zobel demonstrated that these "approximately complete" judgments produced by pooling were sufficient to fairly compare retrieval runs. Using the leave-out-uniques (LOU) 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 %.

As document sets continue to grow, the proportion of documents contained in standard-sized pools shrinks. At some point, pooling's premise must become invalid. The test collection created in the Robust and HARD tracks in TREC 2005 showed that this point is not at some absolute pool size, but rather when pools are shallow relative to the number of documents in the collection [2]. With shallow pools, the sheer number of documents of a certain type fill up the pools to the exclusion of other types of documents. This produces judgments sets that are biased against runs that retrieve the less popular document type, resulting in an invalid evaluation.

Several recent TREC tracks have investigated new ways of sampling from very large documents sets to obtain judgment sets that support fair evaluations. The primary goal of the terabyte track that was part of TRECs 2004–2006 was to investigate new pooling strategies to build reusable, fair collections at a reasonable cost despite collection size. The new million query track is a successor to the terabyte track in that it has the same goal, but a different approach. The goal in the million query track is to test the hypothesis that a test collection containing very many topics, each of which has a modest number of well-chosen documents judged for it, will be an adequate tool for comparing retrieval techniques. The legal track has used a different sampling strategy still to address the challenging problem of comparing recall-oriented (see below) searches of large document sets for both ranked and unranked result sets.

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. (An 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. For a single topic, recall and precision at a common cut-off level 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.

The measures described above are traditional retrieval evaluation measures that assume (relatively) complete judgments. As concerns about traditional pooling arose, new measures and new techniques for estimating existing measures given a particular judgment sampling strategy have been investigated. Bpref is a measure that explicitly ignores unjudged documents in the retrieved sets, and thus it can be used when judgments are known to be far from complete [3]. It is defined as the inverse of the fraction of judged irrelevant documents that are retrieved before relevant ones. The sampling strategies used in the million query and legal tracks have corresponding methods for estimating the value of evaluation measures based on the sampled documents. The track overview paper gives the details of the evaluation methodology used in that track.

3 TREC 2007 Tracks

TREC's track structure began 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 2007 tracks. See the track reports later in these proceedings for a more complete description of each track.

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

	TREC															
Track	’92	’93	’94	’95	’96	’97	’98	’99	’00	’01	’02	’03	’04	’05	’06	’07
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	31	28
Web								17	23	30	23	27	18			
Video										12	19					
Novelty											13	14	14			
Genomics												29	33	41	30	25
HARD												14	16	16		
Robust												16	14	17		
Terabyte													17	19	21	
Enterprise														23	25	20
Spam														13	9	12
Legal															6	14
Blog															16	24
Million Q																11
Participants	22	31	33	36	38	51	56	66	69	87	93	93	103	117	107	95

3.1 The blog track

The blog track first started in TREC 2006. Its purpose is to explore information seeking behavior in the blogosphere, in particular to discover the similarities and differences between blog search and other types of search. The TREC 2007 track contained three tasks, an opinion retrieval task that was the main task in 2006; a subtask of the opinion task in which systems were to classify the kind of the opinion detected (the polarity task); and a blog distillation (also called a feed search) task.

The document set for all tasks was the blog corpus created for the 2006 track and distributed by the University of Glasgow (see `http://ir.dcs.gla.ac.uk/test_collections`). This corpus was collected over a period of 11 weeks from December 2005 through February 2006. It consists of a set of uniquely-identified XML feeds and the corresponding blog posts in HTML. For the opinion and polarity tasks, a “document” in the collection is a single blog post plus all of its associated comments as identified by a Permalink. The collection is a large sample of the blogosphere as it existed in early 2006 that retains all of the gathered material including spam, potentially offensive content, and some non-blogs such as RSS feeds. Specifically, the collection is 148GB of which 88.8GB is permalink documents, 38.6GB is feeds, and 28.8GB is homepages. There are approximately 3.2 million permalink documents.

In the opinion task, systems were to locate blog posts that expressed an opinion about a given target. Targets included people, organizations, locations, product brands, technology types, events, literary works, etc. For example, three of the test set topics asked for opinions regarding Coretta Scott King, JSTOR, and Barilla brand pasta. Targets were drawn from a log of queries submitted to a commercial blog search engine. The query from the log was used as the title field of the topic statement; the NIST assessor who selected the query created the description and narrative parts of the topic statement to explain how he or she interpreted that query.

The systems' job in the opinion task was to retrieve posts expressing an opinion of the target without regard to the kind (polarity) of the opinion. Nonetheless, the relevance assessors did differentiate among different types of posts during the assessment phase as they had done in 2006. A post could remain unjudged if it was clear from the URL or header that the post contains offensive content. If the content was judged, it was marked with exactly one of: irrelevant (not on-topic), relevant but not opinionated (on-topic but no opinion expressed), relevant with negative opinion, relevant with mixed opinion, or relevant with positive opinion. These judgments supported the polarity subtask. For the polarity subtask, participants' systems labeled each document in the ranking submitted to the opinion task with the predicted judgment (positive, negative, mixed) of that document.

The goal in the blog distillation task was for systems to find blogs (not individual posts) with a principal, recurring interest in the subject matter of the topic. Such technology is needed, for example, when a user wishes to find blogs in an area of interest to follow regularly. The system response for the feed task was a ranked list of up to 100 feed ids (as opposed to permalink ids.) Topic creation and relevance judging for the feed task were performed collaboratively by the participants.

Twenty-four groups total participated in the blog track including 20 in the opinion task, 11 in the polarity subtask, and 9 in the feed task.

To address the question of specific opinion-finding features that are useful for good performance in the opinion task, participants were asked to submit both a topic-relevance-only baseline and an opinion-finding run. Results from this comparison were mixed, with some systems showing a marked increase in effectiveness over good baselines by using opinion-specific features, but others showing serious degradation. Nonetheless, as in the 2006 track the correlation between topic-relevance effectiveness and opinion-finding effectiveness remains very high, indicating that topic-relevance effectiveness is still a dominant factor in good opinion finding.

3.2 The enterprise track

TREC 2007 was the third year of the enterprise track, a track whose goal 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 reports, intranet web sites, and email, and a goal is to have search systems deal seamlessly with the different data types.

Because of the track's focus on supporting a user of an organization's data, the data set and task abstraction are particularly important. The document set in the first two years of the track was a crawl of the World-Wide Web Consortium web site. This year the document set was instead a crawl of `www.cisro.au`, the web site of the Commonwealth Scientific and Industrial Research Organisation (CSIRO), which is Australia's national science agency. CSIRO employs people known as science communicators who enhance CSIRO's public image and promote the capabilities of CSIRO by managing information and interacting with various constituencies. In the course of their work, science communicators can come upon an area of focus for which no good overview page exists. In such a case a communicator would like to find a set of key pages and people in that area as a first step in creating an overview page (or to stand as a substitute for such a page). This "missing page" problem was the motivation for the two tasks in the track.

In the document search task systems were to retrieve a set of key pages related to the target topic. As in previous years, a key page was defined as an authoritative page that is principally about the target topic. In the search-for-experts task systems returned a ranked list of email addresses representing individuals who are experts in the target topic. Unlike previous years, there was no a priori list of people made available to the systems. Instead, systems were required to mine the document set to find people and decide whether they are experts in a given field. Systems were required to return a list of up to 20 documents in support of the nomination of an expert.

The topics for the track were developed by current CSIRO science communicators, with the same set of topics used for both tasks. Communicators were given a CSIRO query log and asked to develop topics using queries taken from the log or something similar to those. In addition to the query, the communicators were also asked to supply examples of key pages for the area of the query, one or two CSIRO staff members who are experts in that area, and a short description of the information they would consider relevant to include in the overview page.

Systems were provided with the query and description as the official topic statement. Systems could also access the communicator-provided key page examples for relevance feedback experiments. The experts supplied by the science communicators were used as the relevance judgments for the expert search task. Document pools were judged by participants based on the full topic statements to produce the relevance judgments for the document task.

Twenty groups total participated in the enterprise track, with 16 groups participating in the document task and 16 in the expert search task. Comparison between feedback and non-feedback runs in the document task shows that successfully exploiting the example key pages was challenging: only a few teams submitted feedback runs that were more effective than their own non-feedback runs. The results from the expert-finding task suggest that systems are finding only people associated with a given topic rather than actual expertise. For example, systems suggested the science communicators as experts for some topics.

3.3 The genomics track

The goal of genomics track is to provide a forum for evaluation of information access systems in the genomics domain. It was 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 access. The task in the TREC 2007 track was similar to the passage retrieval task introduced in 2006. In this task systems retrieve excerpts from the documents that are then evaluated at several levels of granularity to explore a variety of facets. The task is motivated by the observation that the best response for a biomedical literature search is frequently a direct answer to the question, but with the answer placed in context and linking to original sources.

The document collection used for 2007 was the same as that used for 2006. This document collection is a set of full-text articles from several biomedical journals that were made available to the track by Highwire Press. The documents retain the full formatting information (in HTML) and include tables, figure captions, and the like. The test set contains about 160,000 documents from 49 journals and is about 12.3 GB of HTML. A passage is defined to be any contiguous span of text that does not include an HTML paragraph token (<p> or <\p>). Systems returned a ranked list of passages in response to a topic where passages were specified by byte offsets from the beginning of the document.

The format of the topic statements differed from that of 2006. The 2007 topics were questions asking for lists of specific entities such as drugs or mutations or symptoms. The questions were solicited from practicing biologists and represent actual information needs. The test set contained 36 questions.

Relevance judgments were made by domain experts. The judgment process involved several steps to enable system responses to be evaluated at different levels of granularity. Passages from different runs were pooled, using the maximum extent of a passage as the unit for pooling. (The maximum extent of a passage is the contiguous span between paragraph tags that contains that passage, assuming a virtual paragraph tag at the beginning and end of each document.) Judges decided whether a maximum span was relevant (contained an answer to the question), and, if so, marked the actual extent of the answer in the maximum span. In addition, the assessor listed the entities of the target type contained within the maximum span. A maximum span could contain multiple answer passages; the same entity could be covered by multiple answer passages and a single answer passage could contain multiple entities.

Using these relevance judgments, runs were then evaluated at the document, passage, and aspect (entity) levels. A document is considered relevant if it contains a relevant passage, and it is considered retrieved if any of its passages are retrieved. The document level evaluation was a traditional ad hoc retrieval task (when all subsequent retrievals of a document after the first were ignored). Passage- and aspect-level evaluation was based on the corresponding judgments. Aspect-level evaluation is a measure of the diversity of the retrieved set in that it rewards systems that are able to find more different aspects. Passage-level evaluation is a measure of how well systems are able to find the particular information within a document that answers the question.

The genomics track had 25 participants. Results from the track showed that effectiveness as measured at the three different granularities was highly correlated. As in the blog track, this suggests that basic recognition of topic relevance remains a dominating factor for effective performance in each of these tasks.

3.4 The legal track

The legal track was started in 2006 to focus specifically on the problem of e-discovery, the effective production of digital or digitized documents as evidence in litigation. Since the legal community is familiar with the idea of searching using Boolean expressions of keywords, Boolean search is used as a baseline in the track. The goal of the track is thus to evaluate the effectiveness of Boolean and other search technologies for the e-discovery problem.

The TREC 2007 track contained three tasks, the main task, an interactive task, and a relevance feedback task. The document set used for all tasks was the IIT Complex Document Information Processing collection, which was also the corpus used in the 2006 track. This collection consists of approximately seven million documents drawn from the Legacy Tobacco Document Library hosted by the University of California, San Francisco. These documents were made public during various legal cases involving US tobacco companies and contain a wide variety of document genres typical of large enterprise environments. A document in the collection consists of the optical character recognition (OCR) output of a scanned original plus metadata.

The main task was an ad hoc search task using as topics a set of hypothetical requests for production of documents. The production requests were developed for the track by lawyers and were designed to simulate the kinds of requests used in current practice. Each production request includes a broad complaint that lays out the background for several requests and one specific request for production of documents. The topic statement also includes a negotiated Boolean query for each specific request. Stephen Tomlinson of Open Text, a track coordinator, ran the negotiated Boolean queries to produce the task's reference run. Participants could use the negotiated Boolean query, the set of documents that matched the Boolean query, and the size of the retrieved set of the Boolean query (B) in any way (including ignoring them completely) for their submitted runs. For each topic systems returned a ranked list of up 25 000 documents (or up to B documents if B was larger than 25 000).

Because of the size of the document collection and the legal community's interest in being able to evaluate the effectiveness of the (unranked) Boolean run, special pools were built from the submitted runs to support Estimated-Recall-at-B as the evaluation measure. The pooling method sampled a total of approximately 500 documents from the set of submitted runs respecting the property that documents at ranks closer to one had a higher probability of being selected for inclusion in the pools. (See the track overview paper for more details.) Note that it is not currently known how reusable the resulting collection is (that is, whether the judgments can be usefully exploited to evaluate runs that did not contribute to the pools). The relevance assessments were made by legal professionals (mostly law students) who followed the legal community's typical work practices.

Iterative search methods generally offer increased effectiveness as compared to the single running of a static query, even if that query is the result of prior negotiation. The feedback and interactive search tasks were introduced into the legal track to explore the level of performance obtainable by iterative search methods in e-discovery and to investigate how best to evaluate those techniques. Both tasks used a subset of the topics from the TREC 2006 legal track.

The goal in the interactive task was for a user to find as many relevant documents as possible for a topic while actively engaging with the retrieval system. Twelve topics were available for this task, ranked in priority order. Participants in the interactive task could do as many of the twelve topics as desired, but were required to perform them in priority order. Submissions consisted of up to 100 documents per topic, which were scored using a utility measure (gaining one point for each relevant document retrieved and losing a half point for each nonrelevant retrieved).

For the relevance feedback task, systems re-ran the TREC 2006 topics exploiting the relevance judgments produced as a result of the TREC 2006 track. Documents that had been judged in 2006 were removed from the submissions ("residual collection" evaluation) and new pools were formed for 10 topics (a subset of the 12 topics used in the interactive task)[1]. The main evaluation measure used in the task was again Estimated-Recall-at-(residual)-B.

A total of 14 groups participated in the legal track: 12 in the main task, 3 in the interactive task, and 3 in the relevance feedback task. Results from the TREC 2007 tasks confirm results from the TREC 2006 track with respect to the Boolean run. Collectively the runs produced by track participants retrieve many relevant documents not retrieved by the negotiated Boolean queries of the reference run, but the average effectiveness of the reference Boolean run is at least as great as the average effectiveness of the other individual runs (with respect to Estimated-Recall-at-B). In other words, all of the runs, including the reference Boolean run, have significant room for improvement with respect to consistently obtaining high recall.

3.5 The million query track

The million query track was a new track in TREC 2007. One of the main goals of the track was to investigate a specific retrieval evaluation hypothesis: that a test collections built using many topics with few, shallow judgments may be a better evaluation tool than a test collection built from fewer topics with relatively thorough judgments. The track also provided an opportunity for participants to explore ad hoc retrieval on a large document set.

The retrieval task of the track was an ad hoc search task over the GOV2 document set. GOV2 is a collection of web pages from within the `.gov` domain spidered in early 2004. The collection contains about 25 million documents and is available from the University of Glasgow (see `http://ir.dcs.`

[1] The new judgments made for the 2007 tasks were created by a different assessor from the one who judged the topic in TREC 2006.

gla.ac.uk/test_collections). The topics for the track were taken from a web search engine log and consisted only of the equivalent of the standard TREC topic statement's title field (some of these topics later had standard topic statements developed for them during the assessing phase). The test set consisted of 10,000 queries, including the title field from some of the topics that had been used in previous years' terabyte tracks.

Relevance judging was performed by both NIST assessors and track participants. The judging procedure was as follows:

1. The assessment system presented the judge with 5 queries randomly selected from the test set.
2. The judge selected one of the queries; the others were returned to the query pool.
3. The judge wrote a description and narrative for this query, thus creating a standard TREC topic statement.
4. The system presented a GOV2 document to the judge and obtained a 3-way judgment (highly relevant, relevant, not relevant) for it.
5. The process continued until at least 40 documents were judged. The judge could continue past 40 documents if he or she wanted to.

The documents to be judged were selected by one of two different sampling methods, the minimal test collection method and the statistical evaluation method, each of which supports a particular evaluation strategy. The details of the sampling and corresponding evaluation methods are given in the track overview paper in these proceedings. The target was to have half the queries that were judged have 20 documents selected by both methods, a quarter of the queries have 40 documents selected by the minimal test collection sampling method, and the remaining queries have 40 documents selected by the statistical evaluation method. Approximately 1800 queries were judged, with a small set receiving judgments from multiple people.

The judgments gathered in this way allow evaluation using the appropriate measure(s) associated with the selection method. The use of the terabyte topics allows runs to be evaluated over those topics using trec_eval and the standard NIST-produced relevance judgments created in the terabyte track as a third evaluation strategy. The 24 runs submitted by 11 groups were each evaluated using the three evaluation strategies in turn. The three different strategies agreed with one another with respect to "big picture" results: all three strategies found the same three clusters of systems with similar effectiveness. More fine-grained comparisons differed across strategies, though, in that rankings of systems within clusters varied depending on the evaluation strategy used. The rankings produced by the two sampling-based evaluation methods were more similar to each other than either was to the ranking produced by evaluation over the terabyte topics.

3.6 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 2007 track contained two tasks, the main task that was a series task similar to the task used since 2004, and a complex interactive QA (ciQA) task introduced in 2006.

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. Answers were required to be supported by a document from the corpus used in the track.

The 2007 main task differed from the task in earlier years in that the corpus consisted of both newswire documents (the AQUAINT-2 collection) and blog documents (the same corpus as was used in the blog track). Introducing blogs into the track created two significant new challenges for QA systems. First, since language use in blogs can be much more informal than in newswire, systems were required to handle language that is not well-formed. Second, blog data also contains discourse structures that are less formal and reliable than newswire, so systems had to do more vetting of candidate responses to determine if those responses were indeed answers.

Despite the introduction of the blog data, which was expected to increase the difficulty of the QA task, individual component scores for the best systems were greater in 2007, after having generally declined each year since TREC 2004. While it is possible that the questions in the 2007 test set are intrinsically easier than previous years, no procedural changes in the way questions were formed were instituted, so large changes in difficulty are not likely.

The ciQA task was introduced in TREC 2006 and is a blend of the TREC 2005 relationship QA task and the TREC 2005 HARD track. The goal of the task is to extend systems' abilities to answer more complex information needs than those covered in the main task and to provide a limited form of interaction with the user in a QA setting.

As in 2006, the questions used in the task contained two parts, a specific question derived from templates of relationship question types, and a narrative that provided more explanation for the specific question. The system response to a question was a ranked list of information "nuggets" supported by AQUAINT documents (the blog corpus was not used in the ciQA task), where each nugget provides evidence for the relationship in question.

The interaction was accomplished using the NIST assessor as the surrogate user and web forms to implement the interface. Unlike 2006, the forms were hosted at the individual participants' home site, so any type of web-based QA system could be used in the task. For each topic, the assessor was given a list of URLs, one URL per participating run. The lists of URLs for different topics were sorted differently, and assessors processed each list in the order given, to control for presentation order effects. Assessors clicked on a URL to begin an interaction and had a maximum of five minutes to finish the task for that pair of run/topic. Participants were responsible for instrumenting the application to capture the results of the interaction.

The protocol for the ciQA task had participants submit initial runs prior to the interaction, perform the interaction, and then submit final runs that (presumably) made use of the information gathered in the interaction. Retrieval results were scored using Pyramid nuggets F-score. In addition, an exit questionnaire gathered data on the assessors' perceptions of the interactions.

Results from the ciQA task showed that, unlike in TREC 2006, most runs were more effective than a sentence-retrieval baseline run. However, many interactions degraded effectiveness; that is, the final run score was less than the corresponding initial run's score. Analysis of the data collected from the exit questionnaire suggested a possible contributing factor for the decrease in effectiveness through interaction: NIST assessors are unusual users in that they already know a lot about the topic, yet the typical users assumed by many participating systems were naive users searching for basic information. Future instantiations of interactive tasks will need to take this mismatch into consideration.

A total of 28 groups participated in the QA track. The main task had 21 participants and the ciQA task had 7 participants.

3.7 The spam track

The spam track was first run in TREC 2005. The goal of the track is to evaluate how well systems are able to separate spam and ham (non-spam) when given an email sequence. The TREC 2007 track repeated the three 2006 tasks using new data. The tasks all involved classifying email messages as ham or spam, differing in the amount and frequency of the feedback the system received.

For each task the track used a test jig that implements a simple interface between the evaluation infrastructure on the one hand and a participant's classifier on the other. The jig 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. In the main on-line filtering task, the classifier receives the correct designation for a message as soon as it classifies the message (this represents ideal user feedback). In the delayed feedback extension to the task, the classifier might eventually receive the correct designation for a message, but the designation for a given message m may come after some number of intervening messages that must be classified before the feedback for m is received, or the feedback may never come at all. In the partial feedback extension to the task, feedback is provided only for messages sent to a subset of the users of a mail server, though the filter is expected to filter messages to all users. In the active learning task, the classifier must explicitly request the correct designation for a document, and may do so for only a given number N of messages.

The track used both a private email stream and a public email stream. Participants ran their own filters on the public corpora using the jig and submitted the evaluation output to NIST. For the private corpora, participants submitted their filters to NIST. NIST passed the filters onto the University of Waterloo after stripping all identification of which filters came from which participant. The University of Waterloo used the jig to run the filters on the private stream and returned the evaluation results to NIST, who then forwarded the evaluation results to the appropriate participant.

Twelve groups participated in the spam track. As in previous years of the track, the general effectiveness of the track's filters has improved relative to the then-current state-of-the-art. Comparison among the different types of training show that both delayed and partial feedback degrade filter effectiveness with respect to ideal feedback, but longer delay periods do not appear to cause more deterioration than shorter delay periods.

4 The Future

TREC 2007 contained a brainstorming session designed to get feedback as to what research areas individuals in the TREC community were personally interested in. In the spirit of true brainstorming, we asked for any ideas without initial filtering by feasibility concerns such as data availability or privacy issues. The session was lively with approximately 40 ideas suggested before discussion was stopped due to time constraints. Enough people expressed interest in three broad areas for those ideas to be further explored informally over a group lunch at the conference and discussion lists after the conference. The goal of the discussions was to formulate a proposal for a TREC track in the area to begin in TREC 2009. The three areas included:

informal text: a track to focus on data access tasks within social media contexts such as instant messaging systems or social tagging;

scientific literature: a track to focus on providing access to the scientific literature more broadly than within a single topic domain as in the genomics track; and

user interaction: a reprise of the TREC interactive track where the focus is on understanding how best to support humans in the search process.

There are five confirmed tracks for TREC 2008. The blog, enterprise, legal, and million query tracks will continue. A new track to examine the effectiveness of relevance feedback across different retrieval models and under different conditions (such as amount of relevance data) will begin. The question answering track will move to a new NIST evaluation conference called the Text Analysis Conference (TAC), see `http://www.nist.gov/tac`. The genomics and spam tracks are ending as TREC tracks, though tasks similar to those investigated in these tracks are expected to appear in other venues.

Acknowledgements

The track summaries in section 3 are based on the track overview papers authored by the track coordinators. My thanks to the 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, Darrin Dimmick, Ian Soboroff, and Ellen Voorhees. Bias and the limits of pooling for large collections. *Information Retrieval*, 10:491–508, 2007.

[3] Chris Buckley and Ellen M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 25–32, 2004.

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

[5] Donna Harman. Overview of the fourth Text REtrieval Conference (TREC-4). In D. K. Harman, editor, *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*, pages 1–23, October 1996. NIST Special Publication 500-236.

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

[7] 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.

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

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

[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 2007

Arizona State University
Beijing U. of Posts & Telecommunications
Carnegie Mellon University
Carnegie Mellon University & U. Karlsruhe
Chinese Academy of Sciences (2 groups)
Concordia University (2 groups)
CRM114 Team
CSIRO ICT Centre
Dalhousie University
Dalian University of Technology
Dartmouth College
Drexel University
EffectiveSoft Ltd
European Bioinformatics Institute
Exegy Inc.
Fitchburg State College
Fondazione Ugo Bordoni, Italian National Research Council, & U. Roma ‘Tor Vergata’
Fudan University
Heilongjiang Institute of Technology
IBM Cairo
IBM Research Lab, Haifa
Indian Institute of Technology, Delhi
Illinois Institute of Technology
Indiana University
International Institute of Information Technology
Jozef Stefan Institute
Kobe University (2 groups)
Kyoto University
Language Computer Corporation
Long Island University
Lymba Corporation
Massachusetts Institute of Technology
Michigan State University
The MITRE Corporation
National Library of Medicine
National Taiwan University
National University of Defense Technology
Northeastern University
Oregon Health & Science University
Open Text Corporation
The Open University
Peking University
Queens College, CUNY
RMIT University
The Robert Gordon University
Saarland University
Sabir Research, Inc
Shanghai Jiao Tong University (2 groups)
South China University of Technology
St. Petersburg State U. & INRIA
SUNY Albany
SUNY Buffalo
Technical University Berlin
TNO, Twente University & EMC
Tokyo Institute of Technology
Tsinghua University
Tufts University
Twente University
University of Alaska Fairbanks
University of Alicante
University of Amsterdam (2 groups)
University of Arkansas at Little Rock
University of Colorado School of Medicine
University of Glasgow
University and Hospitals of Geneva
University of Illinois at Chicago (2 groups)
University of Illinois at Urbana-Champaign
University of Iowa (2 groups)
University of Lethbridge
University of Maryland, College Park
University of Massachusetts
The University of Melbourne (2 groups)
University of Missouri at Kansas City
University of Neuchatel
University of North Carolina
Università di Roma ‘La Sapienza’
University of Strathclyde
University of Texas, Austin
University of Washington
University of Waterloo (2 groups)
Ursinus College
Weill Cornell Medical College
Wuhan University
York University
Zhejiang University