Overview of TREC 2006

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

1 Introduction

The fifteenth Text REtrieval Conference, TREC 2006, was held at the National Institute of Standards and Technology (NIST) 14 to 17 November 2006. The conference was co-sponsored by NIST and the Disruptive Technology Office (DTO). TREC 2006 had 107 participating groups from 17 different countries. Table 2 at the end of the paper lists the participating groups.

TREC 2006 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 2006 contained seven areas of focus called "tracks". Five of the tracks ran in previous TRECs and 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 two new tracks explored blog search and providing support for legal discovery of electronic documents.

There were two main themes in TREC 2006 that were supported by these different tracks. The first theme was exploring broader information contexts than in previous TRECs. This was accomplished by exploring both different document genres and different retrieval tasks. Traditional TREC document genres of newswire (in the QA track) and web pages (in the terabyte track) were still used, but these were joined by blogs (blog track), email (enterprise and spam tracks), corporate repositories (enterprise and legal tracks), and scientific documents (genomic and legal tracks). Retrieval tasks examined included ad hoc search (terabyte, enterprise-discussion, legal, genomics), known-item search (terabyte), classification (spam), specific response (QA, genomics, enterprise-expert), and opinion finding (blog). The second theme of the conference was a focus on creating new evaluation methodologies. These efforts included examining how to make fair comparisons when using massive data sets (terabyte and legal tracks), assessing the quality of a specific response (genomics, QA), balancing realism and privacy protection in experimental design (spam, enterprise), and constructing protocols for efficiency benchmarking in a distributed setting (terabyte).

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 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. The main task in the terabyte track and the legal track task are examples of ad hoc search tasks.

A *known-item search* is similar to an ad hoc search but the target of the search is a particular document (or a small set of documents) that the searcher knows to exist in the collection and wants to find again. Once again, the retrieval system's response is usually a ranked list of documents, and the system is evaluated by the rank at which the target document is retrieved. The named-page-finding task in the terabyte track is an example of a known-item search task.

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, while the opinion search task in the blog track and the discussion search task in the enterprise track are mixtures of ad hoc and categorization tasks.

Information retrieval has traditionally focused on returning entire documents that contain answers to questions rather than returning the answers themselves. This emphasis is both a reflection of retrieval systems' heritage as library reference systems and an acknowledgement of the difficulty of question answering. However, for certain types of questions, users would much prefer the system to answer the question than be forced to wade through a list of documents looking for the specific answer. To encourage research on systems that return answers instead of document lists, TREC has had a question answering track since 1999. In addition, the passage retrieval focus in the genomics track is a move toward question answering, and the expert-finding task in the enterprise track is a kind 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 [4, 8], and TREC continues this tradition. A test collection is an abstraction of an operational retrieval

```
<num> Number:  758
<title> Sugar tariff-rate quotas
<desc> Description:  Describe the nature and history of sugar
tariff-rate quotas in the United States.
<narr> Narrative:  Documents describing the system, its history and how
it works are relevant.  Proposed changes to the system or new agreements
explaining how it works are relevant.  Listings of current allocations
are not relevant.
```

Figure 1: A sample TREC 2006 topic from the terabyte track test set.

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. The document sets used in various tracks have been smaller and larger depending on the needs of the track and the availability of data. The terabyte track was introduced in TREC 2004 to investigate both retrieval and evaluation issues associated with collections significantly larger than 2 gigabytes of text.

While the initial TREC document sets consisted mostly of newspaper or newswire articles, later document sets have included recordings of speech, web pages, scientific documents, blog posts, email messages, and so forth. In each case, 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, text is kept as close to the original as possible. No attempt is 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—though some tracks don't use topics at all (e.g., spam) or use different formats to support the track (e.g., legal). 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 [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 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 [7] 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 [12]. He also found that the TREC collections were not biased against unjudged runs. In this test, he evaluated each run that contributed to the pools using both the official set of relevant documents published for that collection and the set of relevant documents produced by removing the relevant documents uniquely retrieved by the run being evaluated. For the TREC-5 ad hoc collection, he found that using the unique relevant documents increased a run's 11 point average precision score by an average of 0.5 %. The maximum increase for any run was 3.5 %. The average increase for the TREC-3 ad hoc collection was somewhat higher at 2.2 %.

A similar investigation of the TREC-8 ad hoc collection showed that every automatic run that had a mean average precision score of at least 0.1 had a percentage difference of less than 1 % between the scores with and without that group's uniquely retrieved relevant documents [10]. That investigation also showed that the quality of the pools is significantly enhanced by the presence of recall-oriented manual runs, an effect noted by the organizers of the NTCIR (NACSIS Test Collection for evaluation of Information Retrieval systems) workshop who performed their own manual runs to supplement their pools [5].

The leave-out-uniques (LOU) 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 recently. When pools are shallow *relative to the number of documents in the collection*, the sheer number of documents of a certain type fill up the pools to the exclusion of other types of documents. In particular, otherwise diverse retrieval methodologies will all 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. 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 [2]. But large document sets such as the one used in the terabyte track include so many documents containing topic title words that traditional pooling requires pools that are much far too large to be affordable to judge. One of the goals for the terabyte track is to investigate new pooling strategies to build reusable, fair collections at a reasonable cost despite collection size.

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, existing evaluation measures have been adapted and new evaluation measures have been devised. The details of the evaluation methodology used in a particular track are described in the track's overview paper.

3 TREC 2006 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

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
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
Web								17	23	30	23	27	18		
Video										12	19				
Novelty											13	14	14		
Genomics												29	33	41	30
HARD												14	16	16	
Robust												16	14	17	
Terabyte													17	19	21
Enterprise														23	25
Spam														13	9
Legal															6
Blog															16
Participants	22	31	33	36	38	51	56	66	69	87	93	93	103	117	107

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 2006 tracks. See the track reports later in these proceedings for a more complete description of each track.

3.1 The blog track

The blog track is a new track 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 track contained two tasks, an open task and an opinion retrieval task. Participants in the open task defined their own retrieval task and evaluation strategy using the blog corpus. These were pilot evaluations to inform the discussion of the track's future. The opinion retrieval task was a common task with topic development and relevance judgments performed at NIST.

The blog 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. A "document" in the collection (for the purposes of the opinion task) 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.

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 the Macbook Pro, Jon Stewart, and super bowl ads. Targets were drawn from a log of queries submitted to BlogPulse. The query from the log was used as the title field of the topic statement and the NIST assessor created the remaining parts of the topic statement.

While the systems' task was to retrieve posts expressing an opinion of the target without regard to the polarity of the opinion, the relevance assessments made for the track did differentiate among different types of posts to provide useful training data for future tasks. 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.

Fourteen groups participated in the blog opinion task, and an additional two groups participated in the open task. The primary measure used in the track was MAP when treating a document as relevant if it was both on-topic and opinionated. Runs were also evaluated using just on-topic as the definition of relevant. The correlation between the system rankings produced by the two definitions of relevant was high (a Spearman's ρ of 0.97 and a Kendall's τ of .88), suggesting that whether or not a document was on-topic dominated the retrieval results. A baseline run (created after relevance judging was complete) produced by the University of Glasgow's Terrier system with no opinion-specific processing was more effective than any of the submitted systems using either of the definitions of relevant. Thus more work is required to be able to separate opinionated posts from on-topic posts.

3.2 The enterprise track

The enterprise track started in TREC 2005. 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 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 both years of 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.).

The track contained two tasks, a discussion search task and a search-for-experts task. A total of twenty-five groups participated in the enterprise track.

In the discussion search task the systems were to retrieve the set of messages in the email lists that provided pro/con arguments for a particular choice such as "html vs. xhtml". The task was specifically focused on finding arguments for or against a decision rather than simply finding information about the topic. The motivation for the task is to assist users in understanding why a particular decision has been made.

The runs were evaluated both when relevance was defined simply as being on-topic as well as when relevance was defined as containing a pro/con argument. With a few exceptions including a manual run, the relative effectiveness of the runs was largely the same in both cases. Indeed, a more detailed look at the document rankings (see the track overview paper for details) showed that most runs did not consistently retrieve documents containing an argument earlier than documents that were simply on-topic. Thus, more

work is needed to develop argument detectors.

The motivation for the expert-finding task is being able to determine who to contact regarding a particular matter in a large organization. As operationalized in the track task, the expert search mines an organization's documents to create profiles of its people. Systems returned a ranked list of person-ids and a set of supporting documents per person in response to a topic such as "ontology engineering". Systems were given a mapping between names and person-ids of W3C members. The supporting documents were a set of up to 20 documents that the system believed demonstrated why the person was an expert on the topic. Topic creation and relevance assessments were performed by the track participants.

The better expert-finding runs had a mean reciprocal rank (MRR) score greater than 0.9 showing that those systems were generally able to return a true expert at rank one. Corresponding P(10) scores were approximately 0.7 showing that the majority of candidate experts suggested by those runs were in fact experts.

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 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 access. The TREC 2006 track consisted of a single passage retrieval task, though that task was evaluated in a number of different ways to explore a variety of facets. The task was 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 set used in the track was a set of full-text articles from several biomedical journals which 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 162,259 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 topics were derived from the topics used in the TREC 2005 track. The form of the topic was a natural language question, though these were created using a set of "generic topic templates" such as *Find articles describing the role of a gene involved in a given disease*. The test set contained 28 questions, seven questions each from four templates.

Relevance judgments were made by 10 people with expertise in the domain. 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 assigned one or more MeSH terms to that passage as the definition of the *aspect* that the passage pertained to. A maximum span could contain multiple answer passages; the same aspect could be covered by multiple answer passages and a single answer passage could pertain to multiple aspects.

Using these relevance judgments, runs were then evaluated at the document, passage, and aspect 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 30 participants. The passage-level task is apparently a difficult task as evaluation scores for this task were generally low. Effectiveness for both the aspect and document levels was much better, suggesting that the difficulty for the passage level is in finding the appropriate extent of the required information.

3.4 The legal track

The legal track was a new track in 2006. It focused on a specific aspect of retrieval in the legal domain, that of meeting the needs of lawyers to engage in effective discovery of digital documents. Currently, it is common for the two sides involved in litigation to negotiate a Boolean expression that defines the set of documents that are then examined by humans to determine which are responsive to a discovery request. The goal of the track is to evaluate the effectiveness of other search technologies in facilitating this process.

From the retrieval perspective, the task in the track was an ad hoc search task using a set of hypothetical complaints and requests for the production of documents as topics. The document set used in the track was the IIT Complex Document Information Processing collection, which consists of approximately seven million documents drawn from the Legacy Tobacco Document Library hosted by the University of California at 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 a metadata record.

The production requests used as topics 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. Systems could use the negotiated Boolean query in any way they saw fit (including ignoring it completely) for the TREC runs. Stephen Tomlinson of Open Text (Hummingbird) ran the track's reference run, which consisted of running just the negotiated Boolean query for each topic.

The relevance assessments were made by legal professionals who followed their typical work practices. Pools were created using traditional pooling for the TREC submissions received from the six participating groups plus a stratified sample of the baseline Boolean run. In addition, the track organizers arranged for a professional searcher familiar with the document collection to (manually) produce a set of approximately 100 documents for each topic that the searcher expected to be relevant to the topic and unlikely to be retrieved by the other methods. These documents were also added to the pools.

To understand how ranked retrieval approaches can assist discovery, it is necessary to compare ranked retrieval results to the results obtained by the negotiated Boolean queries. Thus, one of the goals of the track was the development of an evaluation methodology that provides for the fair comparison of such runs on a large document set where only a sample of documents is judged. This is a very complicated issue that this first running of the track has just begun to address. In the interim, one measure used in the track was R-precision, a measure that probably favors ranked retrieval runs since the "first" R documents is not well-defined in a pure Boolean run. However, each of the Boolean runs submitted to the track including

the reference run were ranked in some fashion after the Boolean constraint was applied, so R-precision is defined for the track runs. Using R-precision as the measure, the reference Boolean run and several of the best ranked runs were equally effective.

While the average R-precision for the better runs was approximately the same, different runs were relatively better for different topics and each run found relevant documents that the other systems did not retrieve. In particular, the collection contains many relevant documents that do not match the negotiated Boolean queries. This is an important finding for current practice since legal discovery is a recall-oriented task.

3.5 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 2006 track contained two tasks, the main task that was a series task similar to the task used in TRECs 2004 and 2005, and a complex interactive QA (ciQA) task.

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 *AQUAINT Corpus of English News Text* (LDC catalog number LDC2002T31, see `www.ldc.upenn.edu`).

In a change from previous years, time-dependent factoid questions were required to be answered with regard to a particular timeframe (as opposed to the timeframe of an arbitrary document containing an answer). For factoid questions phrased in the present tense, the implicit timeframe was the date of the latest AQUAINT document, i.e., the system was required to answer with the most up-to-date information possible. For factoid questions phrased in the past tense, either the question specified the timeframe (*What cruise line attempted to take over NCL in December 1999?*) or the timeframe of the series that included the question was the implied timeframe (for a target of "France wins soccer's World Cup", the question *Who was the coach of the French team?* is to be interpreted as the coach at the time of the World Cup).

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. In a second change from previous years, the weights given to factoid, list, and other questions in the average were equal. This change lessened the importance of factoid questions in the final score.

In absolute terms, the series scores for participating systems have decreased since 2004. This reflects the increasing difficulty—and realism—of the evaluation conditions. In particular, the new requirement for answers to be correct with respect to the date of the latest document in the collection is a significant departure from previous requirements.

The ciQA task was a blend of the TREC 2005 relationship QA task and the TREC 2005 HARD track. The goal of the task was 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.

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, where each nugget provides evidence for the relationship in question.

The limited interaction with the user (using the assessor as the surrogate user) was accomplished through

forms as in previous HARD tracks. Participants were allowed to create one HTML-based form per question per run. The form contained a task for the assessor to perform, and assessors were limited to no more than 3 minutes per form. The result of the interaction with a form were returned to the participant, who (presumably) incorporated the results into a new question answering run.

Six groups participated in the ciQA task. In addition, the University of Maryland provided an initial baseline run constructed by retrieving sentences using the Lucene search engine, and a corresponding final baseline run that eliminated those sentences that the assessor marked not relevant during the clarification form interaction. This baseline set was among the best of the runs, excluding a manual run set that was clearly more effective than all other submissions. This is yet another example in TREC 2006 where it has proved difficult to improve on the effectiveness of standard retrieval technology for more specialized tasks.

Thirty-one groups participated in the QA track.

3.6 The spam track

The spam track was first run in TREC 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. Nine groups participated in the TREC 2006 spam track.

The 2006 track included an on-line filtering task as in the 2005 track, plus an enhancement to that task and a new active learning task. For each task the track used 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. In the original 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 eventually receives the correct designation for each 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. In the new active learning task, the classifier must determine the designations for the final 10% of an email stream based on learning the correct designations for exactly N messages of its own choosing from the first 90% of the stream (where N was much smaller than 90% of the collection size).

The track used two private email streams and two public email streams. The private streams and one of the public streams were predominately English streams (some spam messages could be in other languages) while the second public stream was predominately Chinese. 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 corpora and returned the evaluation results to NIST, who then forwarded the evaluation results to the appropriate participant.

The overall results were consistent across the four email streams. Detecting spam is more difficult when given delayed feedback than when immediate feedback is available; the active learning task is even more difficult. Nonetheless, filters are able to detect the vast majority of spam with high accuracy, and there is no indication that this year's (more recent) spam is any harder to detect than earlier spam.

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 previous TREC collections.

The document collection used in the track was the same collection created for the initial running of the track in TREC 2004: 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. 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 strongly encouraged for the ad hoc task since manual runs frequently contribute unique relevant documents to the pools. As part of the inducement for manual runs, an (unspecified) prize was offered to the group that returned the greatest number of unique relevant documents. The efficiency and named page tasks required completely automatic processing only.

Fifty new information-seeking topics were created by NIST assessors for the track. Manual runs used only these 50 topics; automatic runs were required to use the set of 149 topics created for the track from TRECs 2004–2006. Systems returned the top 10,000 documents per topic. In an attempt to overcome the bias toward topic title word documents described in section 2.1.3, pools were created in multiple stages with only the initial stage using traditional pooling. See the terabyte track overview paper for more details.

The more effective automatic ad hoc runs used a variety of retrieval models. Most of these runs used features such as phrases or term proximity factors, and pseudo-relevance feedback was generally put to good use. None of the top eight runs made special use of anchor text, and only one used link analysis in producing the retrieved set.

The efficiency task 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 100,000 queries mined from web search engine logs. To be selected for the query set, the query was required to have a minimum number of hits in the GOV2 collection. The title fields from the ad hoc and named-page tasks' topics were added to this set but were not distinguished in any way. The queries were distributed in four different sets to represent four query streams. Queries in a given stream had to be processed in the order in which they appeared in the stream, but queries from different streams could be interleaved in any manner. Participants ran their systems using the entire query set and returned the top 20 documents per query plus reported the average processing time per query and the total time for all queries. Finally, participants were asked to submit one run using one of three open-source information retrieval systems whose efficiency characteristics are known as a way of normalizing for hardware differences. The queries corresponding to the ad hoc and named-page topics were used to measure the effectiveness of the efficiency runs.

Both effectiveness and efficiency varied greatly across participants. As to be expected, systems could realize effectiveness gains by being less efficient (i.e., a system's most effective run differed from its most efficient run).

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 and a set of target topics created

by the participants.

In contrast to the ad hoc task, the more effective named-page finding runs exploited some combination of link structure, anchor text and document structure (for example, giving greater weight to document title words). The most effective named-page run, `indri06Nsdp` from the University of Massachusetts that had a mean reciprocal rank score of 0.512, used all three factors.

Twenty-one groups participated in the terabyte track.

4 The Future

Initial plans for TREC 2007 were formulated during the TREC 2006 conference. All of the 2006 tracks except the terabyte track will continue into 2007; the terabyte track will pause while the feasibility of collecting and using an even larger document set than GOV2 is explored.

TREC 2007 will contain a new track optimistically called the "Million Query" track. While it is unlikely that a test collection with literally 1,000,000 queries will be constructed, the goal of the track is to test the hypothesis that a test collection built from very many, very incompletely judged queries (topics) is a better research tool than a traditional TREC pooled test collection. Both NIST assessors and TREC participants will judge on the order of 50 documents for a query. Queries will be mined from web search engine logs with existing TREC topics (title fields) included as part of the query set. The documents to be judged will be selected from participant submissions according to a particular sampling strategy such as those suggested by Yilmaz and Aslam [11] or Carterette et al. [3]. (Particular strategies will be randomly assigned to queries.) The expectation is that this will allow different sampling strategies to be compared on both the validity of the resulting test collection and the expense of producing the collection.

Acknowledgements

Special thanks to the track coordinators who make the variety of different tasks addressed in TREC possible. The track summaries in section 3 are based on the track overview papers authored by the coordinators.

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. In *Proceedings of the Twenty-Ninth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, pages 619–620, 2006.

[3] Ben Carterette, James Allan, and Ramesh Sitaraman. Minimal test collections for retrieval evaluation. In *Proceedings of the Twenty-Ninth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, pages 268–275, 2006.

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

[5] Noriko Kando, Kazuko Kuriyama, Toshihiko Nozue, Koji Eguchi, Hiroyuki Kato, and Souichiro Hidaka. Overview of IR tasks at the first NTCIR workshop. In *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pages 11–44, 1999.

[6] 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] 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`.

[11] Emine Yilmaz and Javed A. Aslam. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pages 102–111, Arlington, Virginia, November 2006.

[12] Justin Zobel. How reliable are the results of large-scale information retrieval experiments? In W. Bruce Croft, Alistair Moffat, C.J. van Rijsbergen, Ross Wilkinson, and Justin Zobel, editors, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 307–314, Melbourne, Australia, August 1998. ACM Press, New York.

Table 2: Organizations participating in TREC 2006

Arizona State University	Australian National University & CSIRO
Beijing University of Posts and Telecommunications	Carnegie Mellon University
Case Western Reserve University	Chinese Academy of Sciences (2 groups)
The Chinese University of Hong Kong	City University London
CL Research	Concordia University (2 groups)
Coveo Solutions Inc.	CRM114
Dalhousie University	DaLian University of Technology
Dublin City University	Ecole des Mines de Saint-Etienne
ErasmusMC, TNO, & University of Twente	Fidelis Assis
Fudan University (2 groups)	Harbin Institute of Technology
Humboldt University, Berlin & Strato AG	Hummingbird
IBM Research Haifa	IBM T.J. Watson Research Center
Illinois Institute of Technology	Indiana University
Institute for Infocomm Research	ITC-irst
Jozef Stefan Institute	Kyoto University
Language Computer Corporation (2 groups)	LexiClone Inc.
LowLands Team	Macquarie University
Massey University	Max-Planck Institute for Informatics
Massachusetts Institute of Technology	The MITRE Corp.
National Institute of Informatics	National Library of Medicine
National Security Agency	National Taiwan University
National University of Singapore	NEC Laboratories America, Inc.
Northeastern University	The Open University
Oregon Health & Science University	Peking University
Polytechnic University	Purdue U. & Carnegie Mellon U.
Queen Mary University of London	Queensland University of Technology
Ricoh Software Research Center Beijing	RMIT University
Robert Gordon University	Saarland University
Sabir Research, Inc	Shanghai Jiao Tong University
Stan Tomlinson	State University of New York at Buffalo
Technion - Israel Institute of Technology	Tokyo Institute of Technology
TrulyIntelligent Technologies	Tsinghua University
Tufts University	UCHSC at Fitzsimons
University of Alaska Fairbanks	University of Albany
University of Amsterdam (2 teams)	U. of Arkansas at Little Rock
U. of California, Berkeley	U. of California, Santa Cruz
University of Edinburgh	University of Glasgow
University of Guelph	University of Hannover
University and Hospitals of Geneva	U. of Illinois at Chicago (2 groups)
U. of Illinois at Urbana-Champaign	University of Iowa
U. of Karlsruhe & Carnegie Mellon U.	University of Limerick
U. Maryland Baltimore County & APL, Johns Hopkins U.	University of Maryland
University of Massachusetts	The University of Melbourne
Universit degli Studi di Milano	University of Missouri-Kansas City
Universite de Neuchatel	University of Pisa
University of Pittsburgh	University of Rome "La Sapienza"
University of Sheffi eld	University of Strathclyde
University of Tokyo	U. of Ulster & Saint Petersburg State U.
University of Washington	University of Waterloo
University of Wisconsin	Weill Medical College of Cornell U.
York University	