TREC-2006 Legal Track Overview

Jason R. Baron, National Archives and Records Administration, Office of General Counsel, Suite 3110, College Park, MD 20740, jason.baron@nara.gov

David D. Lewis, David D. Lewis Consulting, 858 W. Armitage Ave. #296, Chicago, IL 60614, trec06@DavidDLewis.com

Douglas W. Oard, College of Information Studies and Institute for Advanced Computer Studies, University of Maryland**,** College Park, MD 20742, oard@umd.edu

Abstract

This paper describes the first year of a new TREC track focused on "e-discovery" of business records and other materials. A large collection of scanned documents produced by multiple real world discovery requests was adopted as the basis for the test collection. Topic statements were developed using a process representative of current practice in e-discovery applications, with both Boolean and natural language queries being supported. Relevance judgments were performed by personnel who had received professional training, and often considerable experience, in review of similar materials for this task. Six research teams and one manual searcher submitted a total of 33 retrieved sets for each topic. These were pooled and a portion assessed to support evaluation of both the retrieved sets themselves and for future use of the collection.

1. Introduction

The use of information retrieval techniques in law has traditionally focused on providing access to legislation, regulations, and judicial decisions. Searching business records for information pertinent to a case (or "discovery") has also been important, but digitally searchable records were until recently the exception rather than the norm. That is rapidly changing, however. The motivating goal of this new legal track at the Text Retrieval Conference (TREC) is to assess the ability of information retrieval technology to meet the needs of the legal community for tools to help with retrieval of business records. This is an issue of increasing importance given the vast amount of information in electronic form to which access is required during litigation. Ideally, the results of our studies will also help to advance the discussion of the capabilities and limitations of automated support for e-discovery in the legal community.

The importance of doing well at e-discovery is hard to overstate. In the past few years, lawsuits involving giant corporations and single individuals alike have resulted in huge multi-million and even billion dollar adverse verdicts turning on the failure of a party to the litigation to properly preserve and provide access to various forms of electronic records, including most notably e-mail, and data on backup tapes (see, e.g., *Coleman, 2005*; *Zubulake, 2004*). Beyond the headlines, however, are a growing percentage of lawsuits that involve the production of responsive electronic data stored in vast corporate, governmental, and other repositories. Lawyers are struggling to keep up with the profusion of electronic data and metadata in all its forms, on desktops and networks. So too, troves of "legacy" documents, sometimes going back decades, continue to be maintained and need to be searched in response to discovery requests.

The results of the legal track are especially timely and important given recent changes in the U.S. Federal Rules of Civil Procedure that went into effect on December 1, 2006. The amended rules introduce a new category of evidence, namely, "electronically stored information" ("ESI") in "any medium," intended to stand on an equal footing with existing rules covering the production of "documents." Rule 26(f) specifically directs that at an initial conference of the parties, "any issues relating to disclosure or discovery of electronically stored information, including the form or forms in which it should be produced" are to be

discussed. Such issues will necessarily include the need to consider how appropriate ESI is made accessible to opposing parties. Providing access involves more than just search technology, of course—initial query formulation, iterative query refinement, and review of search results for relevance and privilege are important components of the entire process. The Advisory Committee notes to Rule 34 say that in talking about "ESI in any medium," the rules amendments were intended to "encompass future developments in computer technology," which speaks specifically to our goals for the TREC Legal Track.

Against the backdrop of the Federal Rules changes, the status quo in the legal profession, even in large and complex litigation, is continued reliance on free-text Boolean searching for satisfying document (and now ESI) production demands (Sedona Conference 2005). Thus, to the extent a trend exists in the case law, it is where courts have intervened at early stages to ensure that parties negotiate "search protocols." To date these have consisted solely of a static list of agreed upon query terms, rather than more complex forms of negotiations over, for example, complex (extended) Boolean queries (e.g., those specifying truncation and/or proximity operators). Moreover, as of the date of this paper, there is no reported case law in the United States where courts have been called upon to adjudicate the reasonableness of alternative forms of search methodologies (e.g., ranked retrieval). It is only a matter of time, however, before parties in litigation will more fully utilize alternative techniques, enter into negotiations regarding search system selection and/or query formulation, and, inevitably, conduct formal adjudication over the reasonableness and efficacy of such alternative approaches.

An important aspect of e-discovery and thus of the TREC legal track is an emphasis on recall over precision. In light of the fact that a large percentage of requests for production of documents (and now ESI) routinely state that "all" such evidence is to be produced, it becomes incumbent on responding parties to attempt to maximize the number of responsive documents found as the result of a search. All things being equal, lawyers would be expected to move towards alternative search methods that produce greater numbers of responsive documents for the same resources expended; conversely, alternatives that produce fewer responsive documents are likely to be judged as insufficient, even if greater precision (economy) is achieved overall. If recall comparable to the presently used techniques could be assured, then interest would likely exist in increasing precision (thereby diminishing the need to manually review false positive hits generated by automated means).

There have been to date few research efforts studying effectiveness of retrieval in civil discovery contexts. The seminal study (Blair & Maron, 1985), found that while attorneys believed they had found 75% of the relevant documents for litigation involving a train accident, in fact only an estimated 20% of relevant documents were discovered. The authors attributed this to the inherent ambiguity of language. At least one later study has looked at a comparison of Boolean and natural language searches in the context of a structured database of case precedents (Turtle 1994), and experiments with Boolean systems on outside the legal context have been reported at TREC (e.g., Lu et al. 1993; Jacobs 1995) and elsewhere.

The key goal of the TREC 2006 legal track was to apply objective benchmark criteria for comparing search technologies, using topics and documents approximating those of actual discovery settings. Given the reality of the use of Boolean search in present day litigation, of significant interest was comparing the efficacy of Boolean search using negotiated queries with alternative methods. The chosen collection, about seven million scanned documents from the tobacco Master Settlement Agreement, can also be used for technology-centered experiments comparing retrieval techniques based on metadata and/or optical character recognition.

The remainder of this paper is organized as follows. Section 2 describes the document collection. Section 3 then explains the topic development process. In Section 4, the process by which relevance judgments were created is presented. Section 5 identifies the participating research teams and presents some preliminary results. Section 6 concludes the paper.

2. Document Collection

The Legal Track required a collection reflecting the scope and diversity of documents searched in real discovery settings. Obtaining access to the internal documents of large enterprises for research purposes is difficult, but ironically discovery proceedings in real legal cases provide one source of such material. As the Legal Track test collection we chose the IIT CDIP Test Collection, version 1.0 (which we will refer to as "IIT CDIP 1.0") which is based on documents released under the tobacco "Master Settlement Agreement" (MSA).

The MSA settled a range of lawsuits by the Attorneys General of several US states against seven US tobacco organizations (five tobacco companies and two research institutes). One part of this agreement required those organizations to make public on the World Wide Web (through at least June 30, 2010) all documents produced in discovery proceedings in the lawsuits by the states, as well as all documents produced in a number of other smoking and health-related lawsuits. Notable among the provisions is that the tobacco organizations were required to provide to the National Association of Attorneys General (NAAG) a copy of metadata and the scanned documents from the websites, and are forbidden from objecting to any subsequent distribution of this material. The text of the MSA and accompanying appendices and other documents can be found at the websites of Attorneys General of several US states, including California (*http://ag.ca.gov/tobacco/msa.php*).

The University of California San Francisco (UCSF) Library, with support from the American Legacy Foundation, has created a permanent repository, the Legacy Tobacco Documents Library (LTDL), for tobacco documents (Schmidt, Butter & Rider 2002) in order to assure continued availability of these materials. The Illinois Institute of Technology (IIT) Complex Document Information Processing (CDIP) 1.0 collection is based on a snapshot, generated between November 2005 and January 2006, of the MSA subcollection of the LTDL. The snapshot consisted of 1.5 TB of scanned document images, as well as metadata records and optical character recognition (OCR) output produced from the images by UCSF. The IIT CDIP project subsequently reformatted the metadata and OCR, combined the metadata with a slightly different version obtained from UCSF in July 2005, and discarded some documents with formatting problems, to produce the IIT CDIP 1.0 collection (Lewis, et. al 2006).

The IIT CDIP 1.0 collection consists of 6,910,192 document records in the form of XML elements. The two subelements which provide the most conventional target for text retrieval are <ti> (the document title) and <ot> (the OCR text). The highly variable quality of the OCR, combined with the great variations in document length (from one page to thousands of pages) makes retrieval even on these fields a challenge. In addition to the text subelements, there are a wide range of other metadata subelements present in some or all of the records, including senders and recipients, important names mentioned in the document, controlled vocabulary categories, geographical and organizational context identifiers, and many others. The degree to which this information is present varies with the originating tobacco organization and other factors. Overall, the structure of the data is extremely rich and still not well understood.

IIT CDIP 1.0 had strengths and weaknesses as a collection for the Legal Track. The wide range of document genres (including letters, memos, budgets, reports, agendas, minutes, plans, transcripts, scientific articles, email, and many others) and the large number of documents are very typical of legal discovery settings. The fact that documents were scanned and OCR'd is representative of some discovery situations, but perhaps not those of most interest to those concerned with electronic discovery. The rich but variable quality metadata is also perhaps not typical. The fact that the MSA documents were themselves the *output* of legal discovery proceedings might suggest they are unrepresentative as *inputs* to TREC's simulation of a legal discovery situation. Our worries about that point are mitigated to some extent, however, by the fact that the MSA documents originated from seven different organizations in response to hundreds of distinct document requests in multiple legal cases. Thus their diversity is more representative of a diverse population of company records than perhaps might initially be imagined. We further addressed this concern by using a range of topics in the evaluation, some with content highly similar to MSA discovery requests, and others very different. The fact that documents originated from seven different organizations but were searched as a unit is decidedly anomalous from the perspective of federated search, and some future users of the collection may wish to treat the seven subcollections in a more separate manner.

Several minor glitches in the preparation of IIT CDIP 1.0 turned up during indexing of the data by Legal Track participants. In addition, a number of documents turned out to have XML records but no document images, which was both an immediate problem for relevance assessment, and also a problem for the types of document image retrieval and mining studies towards which the CDIP project is targeted (Agam et al. 2006). These problems are being investigated in ongoing work by the IIT CDIP project.

3. Topic Development

Topic development was modeled on U.S. civil discovery practice. In the litigation context, a "Complaint" is filed in court, outlining the theory of the case, including factual assertions and causes of action representing the legal theories of the case. In a regulatory context, often formal letters of inquiry serve a similar purpose by outlining the scope of the proposed investigation. In both situations, soon thereafter one or more parties create and transmit formal "requests for the production of documents" to adversary parties, based on the issues raised in the Complaint or Letter of Inquiry. (If in federal court, this type of demand is typically filed pursuant to Fed. R. Civ. P. 34, but may also be sent to third party non-defendants via subpoena under Fed. R. Civ. P. 45.) Requests to produce documents are typically very broadly worded, in an attempt to force the opposing party to provide a maximum number of responsive documents. In some cases, however, requests are purposely more narrowly tailored when the focus is on particular documents known to be in the possession of a party which are deemed useful at trial. A third category of requests are aimed at finding only particular types of documents (e.g. all "internal memoranda" on a designated topic.)

It is increasingly common for lawyers to consider requesting that specific search terms be used for the purpose of searching large databases for potentially responsive documents. Courts have begun referring to the development of "search protocols," which are to be developed either unilaterally or, to a greater or lesser extent, made subject to negotiations between parties prior to conducting searches. At present, it is typically assumed that an extended Boolean search (i.e., one with truncation and/or proximity operators) will be performed, although some legal technology firms now also support other types of search technology. Less well known is what percentage of cases have utilized a robust or sophisticated process of negotiations over how search terms, wildcards, Boolean logic, and proximity operators are to be combined to form queries. Nevertheless, for the purpose of the TREC 2006 legal track, it was deemed important to develop topics that stood in as proxies for real-life requests to produce documents in which a set of Boolean strings were developed by a negotiation process between two parties.

For the TREC 2006 legal track, five hypothetical complaints were created by members of the Sedona Conference®, a group of lawyers who are leading the development of professional practices for e-discovery. These complaints described: (1) an investigation into a fictional tobacco company's improper campaign contributions; (2) a consumer protection lawsuit challenging a fictional tobacco company's "product placement" decisions in television, film, and theatre shows watched by children; (3) an "insider training" securities lawsuit involving fictional tobacco executives; (4) an antitrust lawsuit involving the movement of commerce in California; and (5) a product liability lawsuit involving defective surgical devices as shown in animal testing. In using fictional names and jurisdictions, the track coordinators, on behalf of TREC, attempted to ensure that no third party would mistake the academic nature of the TREC legal track for an actual lawsuit against real-world companies, and any would-be link or association with either past or present real litigation involving such companies was entirely unintentional.

For each of the five complaints, a set of topics (formally, "requests to produce") were initially created by the creator of the complaint, and revised by the track co-coordinators. Revisions were considered necessary where the initial topic appeared to have too few or too many relevant documents for effective evaluation, or when it was feared assessors would find the topic too ambiguous. (In this respect, the TREC exercise models real-life objections that often are made to "overbroad," "vague," or "ambiguous" discovery requests, sometimes resulting in courts requiring parties to re-submit narrower and more focused requests.) In the end, 43 topics were selected by the track coordinators for use in the evaluation.

Two aspects of this screening process were less than ideal. First, the evaluation of breadth and ambiguity was done by the track organizers and a professional tobacco searcher, not by the eventual assessor for each

topic, as NIST has often been able to do in past TRECs. (Most assessors had not yet been recruited at the time topics were drawn up.) Second, the screeners did not have access to ranked retrieval search of the collection. Screening was done using the Boolean interface available from UCSF,[1] which at that time had only a beta version of OCR search.

For each of these 43 topics, the initial topic creator and a track coordinator took the roles of requester and respondent (respectively) in a discovery process, and engaged in an iterated negotiation over the form of a Boolean query for the topic. The final XML topic file contained 43 entries, each including the production request, the associated complaint (which for simplicity was repeated in full for each production request associated with that complaint), the extended Boolean query initially proposed by the (simulated) requesting party, the final extended Boolean query that was agreed upon, and any additional extended Boolean queries in the negotiation history. Human-readable versions of the complaints and the production requests were also prepared for use by relevance assessors and interactive searchers, and a cross-reference to each was recorded in the XML topic file. The topic file is available from the track Web page, *http://trec-legal.umiacs.umd.edu*.

4. Relevance Judgments

This section describes the process by which relevance judgments were created.

4.1. Creating Judgment Pools

The complexity of the CDIP documents and topics, and a report of pooling problems with other large collections (Buckley, et al 2006) generated some concern about the adequacy of conventional pooling approaches for the Legal Track. We adopted several strategies for addressing these problems, though none were a complete solution.

We invited track participants to submit up to eight runs (in an effort to maximize pool diversity), asked for runs to depth 5,000 (to facilitate computation of recall-oriented evaluation measures), and asked participating teams to designate their runs for inclusion in the assessment pools in priority order. We included in the assessment pools the top 100 documents from the highest priority run from each team and the top 10 documents from each of the other runs from that team. This yielded a maximum of 170 documents per team for any topic, although usually fewer documents than that were added to the pools because duplicates were removed (both within and across teams). A total of six participating teams submitted a total of 31 runs for official scoring. Two additional runs that were commissioned especially for the track were then used to further enrich the pools.

It is well known that expert searchers can and will often find documents that fully automated term-matching techniques would miss. The IIT CDIP project therefore contracted with an expert tobacco document searcher (Celia White, *http://professionalresearchservices.com*) to produce a set of approximately 100 documents for each topic to add to the pools. Working with a track coordinator, she attempted to find documents that were both relevant to a topic and unlikely to be highly ranked by ranked retrieval systems.

A particular interest in the Legal Track was to compare the effectiveness of the final negotiated Boolean query with the effectiveness of ranked retrieval systems. Hummingbird generously agreed to submit for our use as a baseline Boolean run the retrieved sets resulting from directly executing the negotiated Boolean query (with only a few format corrections, as described in the Open Text[2] team's paper). This run was not counted as an official submission of the Hummingbird team, but rather as a track baseline. We then drew a stratified sample (Cochran, 1977; Lewis, 1996) from the set of documents retrieved by the

[1] http://legacy.library.ucsf.edu/

[2] Hummingbird was acquired by Open Text Corporation in October 2006. Hence the Open Text Corporation paper describes the Hummingbird runs.

negotiated Boolean query for each topic in order to support unbiased estimation of certain evaluation measures for these sets.

Stratification was done by assigning each document from a baseline Boolean set to one of three strata based on whether and how that document occurred in the 31 official submitted runs. The three strata were:

- ***Stratum 1*** (documents occurring in the top 5,000 for at least one official run submitted by each of two or more of the six participating sites),
- ***Stratum 2*** (documents occurring in the top 5,000 for one or more official runs from exactly one of the six participating sites), and
- ***Stratum 3*** (documents not occurring in the top 5,000 for any official run submitted by any participating site).

For each topic, NIST drew a simple random sample of 100 documents from ***Stratum 1***, 50 from ***Stratum 2***, and 50 from ***Stratum 3*** to add to the pool for that topic. When a stratum was exhausted, leftover documents were drawn from the other strata in proportion to their original allocation. Using different stratification strategies for different topics could have improved our estimates, but would have complicated the sampling procedure. An unexpected downside of the above stratification was that ***Stratum 3*** often turned out to be empty. This may have resulted from use of terms from the negotiated Boolean query by ranked retrieval systems, which was allowed (and, indeed, encouraged) by the track guidelines.

One participating team, Hummingbird, leveraged the track's approach to constructing assessment pools (which was known by the participants) to do their own stratified sampling experiment. Their main run (humL06tvz) actually drew documents from various depths of a standard ranked run, enabling them to compute unbiased estimates of precision (and the number of unjudged relevant documents) to depth 9,000. Details can be found in their paper (Tomlinson, 2006). This strategy almost certainly increased the diversity of the assessed pools (at the cost of some richness in relevant documents) by increasing the number of lower-ranked documents assessed. It also invalidated our computation of standard evaluation measures for that run (which are shown in Figure 3 only for completeness).

4.2. Relevance Judgment Process

A total of 35 volunteers from government, law firms, legal technology firms, and law schools (plus two unaffiliated individual volunteers) assessed a total of 32,738 documents in the judgment pools for 40 of the topics. Due to lack of assessment capacity, no assessments were performed for the three remaining topics, and they were thus removed from the evaluation. The volunteers included eight lawyers, ten law students (with 1st, 2nd and 3rd year students all represented), three paralegals with substantial legal experience, one professional archivist, one historian, and several individuals with degrees with science or finance. The affiliations of volunteers for primary assessments were the National Archives and Records Administration (8 topics), George Washington University Law School (D.C., 8 topics), H5 Technologies Inc. (San Francisco, 7 topics), Lewis & Roca LLP (Phoenix, 4 topics), Preston Gates LLP (Seattle, 3 topics), Bank of America (Charlotte, 2 topics), FTI Consulting, (New York City, 2 topics), one topic each by George Mason University School of Law (Virginia), Reasonable Discovery LLC (Virginia), New Mexico State Attorney General's Office, and one topic each from three private individuals (in Florida, California, and the U.K.).

The assessors used a beta version of a Web-based platform to view the scanned MSA documents and record their relevance judgments. (The platform was designed by David D. Lewis Consulting, and implemented by Smokescreen Consulting, as part of the IIT CDIP project.) We provided the assessors with a "How To Guide" (Baron, Lewis & Oard, 2006) that explained that the project was modeled on the ways in which lawyers make and respond to real requests for documents, including in electronic form. Assessors were told to assume that they had been requested by a senior partner, or hired by a law firm or another company, to review a set of documents for "relevance." No special, comprehensive knowledge of the matters discussed in each complaint was expected (e.g., no need to be an expert in federal election law, product liability, etc.). The heart of the exercise was to look for relevant and nonrelevant documents within a topic. Relevance, consistent with all known legal definitions from Wigmore to Wikipedia, was to be

defined broadly. Specifically, assessors were instructed that a document should be considered relevant when the reference to the topic was found in the document. Assessors were reminded that a document may be relevant even if it fails to contain any of the words in the topic request, and conversely, that a document may end up being considered not relevant despite containing one or more words from the topic request. Assessors were also informed that for some topics, the *document type* would circumscribe the scope of the topic (e.g., all *internal memoranda* of a company on topic x), and that (for a very few topics) the scope might be limited by a specified date span (e.g., all documents created in 1992). Relevance judgments were to be recorded as a binary value (yes or no), although a third "unsure" category was also available in the assessment platform.

The first phase of assessment (the only phase initially planned) began on August 7, 2006, and was completed on September 15, 2006. This was the first time that distributed assessment of document images had been used in TREC, and a few complications unsurprisingly arose. It became apparent during assessment that the collection contained some extremely long documents (e.g. a 3,500 page card catalog) and that the participating systems had retrieved a disproportionate number of these long documents. The assessment guidelines were changed in mid-August to allow assessors to mark documents longer than 300 pages as "unsure" if their relevance could not be determined by examining the available metadata and a few pages of the document. Documents marked as unsure were treated as unjudged. When surveyed after completion of their work, some assessors suggested that graded relevance judgments be supported in future years, so as to distinguish between mere "passing references" to a topic (which were recorded as relevant for this year's track) and documents that materially or substantively discuss a topic (which were also recorded as relevant this year).

Some of the assessors went beyond the text of the topic (the complaint, the production request, and the Boolean queries) to perform additional legal research which they viewed as helpful to the exercise. For example, the assessor for Topic 30 researched at greater length what the numbered statutory code provisions were corresponding to the California Cartwright Act, to ensure that all documents containing such references, with or without reference to the Cartwright Act itself, would be marked as responsive. The assessor on Topic 10 performed independent research into the ban on tobacco advertising, as an aid to understanding what documents might be expected to be found in response to a topic involving tobacco product placement in television or film. One assessor asked for assistance on the definition of one of the keywords in the topic, leading to additional research conducted on the Internet.

Some differences were observed in how liberally or narrowly assessors viewed the scope of their discretion to find responsiveness. In some exceptional cases, assessors were willing to find responsiveness even where a key term might be missing, if the document was otherwise sufficiently generic and might yet be viewed as responsive with the aid of further research. For example, the assessor for Topic 9 ("All documents discussing, referencing or relating to payment of compensation to 20th Century Fox Corporation for placement of products and/or brands in a film production"), marked certain documents as relevant even if the film company was not expressly mentioned, where the context indicated that the company might be involved. In most cases, however, assessors appeared to adopt relatively restrictive interpretations on what met the mark for relevance.

Assessors reported some confusion as to whether they should exclude documents that might be within the literal scope of a production request when read in isolation, but which weren't relevant to the main thrust of the associated complaint (i.e., the document had nothing to do with the causes of action in the lawsuit or investigation). The question of scope arose in particular for production requests associated with the one complaint that on its face did not involve allegations against the tobacco industry (but which was instead about medical devices). Topic 49, which coupled that complaint with a production request for "[a]ll documents created between 1962 and 1999 referencing or including warnings or draft warnings used in the United States," proved to be particularly problematic because it was read by the assessor as being aimed at warnings for faulty medical devices. Not surprisingly, no relevant documents were found for topic 49. It was therefore removed from the evaluation because topics with no known relevant documents can not be used to compare the effectiveness of alternative system designs. Results are therefore reported for the remaining 39 topics.

As is often the case, assessors found some unintended ambiguity in the topics, either due to grammatical construction of the topic (e.g., what did the word "their" refer to), or due to inherent ambiguity embedded within words or concepts (e.g., what constitutes "lobbying efforts," "advertising," "marketing," and "promotion"). For one assessor, the word "event" (in a topic asking for all documents relating to the placement of product logos at events held in California), prompted them to consult the Random House Dictionary, where the word is defined as "something that occurs in a certain place during a particular interval of time." Therefore, in this assessor's view, documents that mentioned such activities as the America's Cup Race, speed skiing, auto racing, Hispanic Cultural events, Swing jam weekend, an antiviolence campaign, a country music festival, and an anti-smoking campaign called "Tobacco is Whacko," were all properly within the scope of the topic.

Another miscellaneous concern of one or more assessors involved how to deal with documents containing foreign language text. The track coordinators instructed assessors to make judgments based on English portions of documents, or otherwise mark the document as unsure.

In general, assessors took their jobs very seriously. A number of assessors made a second pass through their document set to resolve anomalies or to revisit judgments based on knowledge gained on the first pass. Many requests were directed to the track coordinators for help in resolving technical concerns.

It turned out that a nontrivial portion of the documents in the judgment pools could not be assessed at all using the assessment platform. While the same set of UCSF XML records provided the starting point for both the IIT CDIP version 1.0 collection and for the assessment platform's database, a few records with formatting problems were inadvertently treated differently by the two groups. In addition, a substantial number of XML records with variant formatting could not be loaded until assessment was already underway. More importantly, an even larger number of documents could not have their page images displayed during much of the assessment period. The total number of documents affected was less than 5% of the total collection, although somewhat more than 5% of the assessment pools were affected because longer documents were more likely to be affected. We addressed these problems by asking assessors to view documents at the LTDL Web site *(http://legacy.library.ucsf.edu/)* if their images could not be viewed on the CDIP platform, and record their assessments using the CDIP platform. In a very few cases, no record at all was loaded on the CDIP platform and assessments were sent by email. Also in a very few cases document images were found to be partial or missing on the LTDL Web site as well. In those few cases, assessors were asked to make a judgment based on the metadata record if possible, or to mark the document as "unsure" (which was treated as unassessed).

The track coordinators asked assessors to record how much time they spent in performing assessment review. Based on post-assessment survey responses and related emails, assessment time data is available for 16 participants representing 39% of the overall assessment effort (12,743 of the 32,738 assessments). The reported review rate of documents reviewed per hour ranged from a low of 12.33 (Topic 31) to a high of 67.5 (Topic 25). The average review rate constituted 24.7 documents per hour. Note that each of the assigned topics included within it a highly varied set of documents, in terms of both differences in subject matter complexity as well as in total length.

4.3. Inter-Assessor Agreement

In order to assess the effects of differing assessor interpretations, we performed a limited amount of dual assessment after completion of the first phase of assessments. A sample of 50 documents (25 that had been judged as relevant, and 25 that had been judged as not relevant) was drawn from the pool for each of the 40 assessed topics. (Topic 49 was included for dual assessments, even though it could not be used for evaluating systems.) When fewer than 25 relevant documents had been identified, the number of non-relevant documents was increased to keep the total at 50. These sets were then assessed by a different assessor, without knowledge of the previous judgments. A total of 12 volunteers assessed documents in this second round, seven first-round veterans who received new topics to review, plus five new recruits.

Figure 1 shows the values of Cohen's kappa (Shoukri, 2004, Sec. 3.3), a chance-corrected measure of agreement, for each topic, as computed from the sample of 50 documents. Let:

n_{00} = number documents judged nonrelevant by main and secondary assessor,
n_{01} = number documents judged nonrelevant by main, but relevant by secondary,
n_{10} = number documents judged relevant by main, but nonrelevant by secondary, and
n_{11}= number of documents judged relevant by both main and secondary assessor.

where $n = n_{00} + n_{01} + n_{10} + n_{11}$ is for us equal to *50*. To compute kappa, one first computes the observed proportion of agreement between the assessors:

$$p_o = (n_{00} + n_{11}) / n$$

and the proportion agreement expected by chance under the assumption the assessors make their judgments independently with their particular observed frequencies of relevant and nonrelevant:

$$p_e = (n_{00} + n_{01})(n_{00} + n_{10}) /n^2 + (n_{10} + n_{11})(n_{01} + n_{11}) /n^2.$$

Cohen's kappa is then:

$$K = (p_o - p_e)/(1 - p_e) .$$

The mean value of kappa over the 40 topics was +0.49, indicating moderate overall agreement between assessors (kappa ranges between -1 for complete disagreement to +1 for complete agreement), although considerable variation was evident across topics. The kappa values shown in Figure 1 are based on a sample of documents with (usually) 25 documents that the main assessor judged positive, and 25 they judged negative. The kappa value would have been different if a random sample from the pool had been judged by both assessors. We can compute an approximation of what kappa on the pool would have been by treating the 50 documents as a stratified sample and computing the expected values of the four contingency table cells that go into kappa. This is not quite an unbiased estimate of what kappa would have been on the pool, since kappa is a nonlinear function of the contingency table cells, but it is a reasonable approximation. Table 1 (which can be found at the end of this paper) shows the raw values of the contingency table entries along with kappa and other associated statistics. Table 2 (also at the end of the paper) shows the stratified estimates of what the contingency table cells would be for the full pool, along with approximations to the agreement measures computed by plugging the expected values of the contingency table cells into the formula for each measure.

As Voorhees has shown, moderate inter-annotator agreement can yield comparisons that are stable when one set of assessments are substituted for the other (Voorhees 2000). Evaluation measures should, therefore, be interpreted on a comparative rather than an absolute basis.

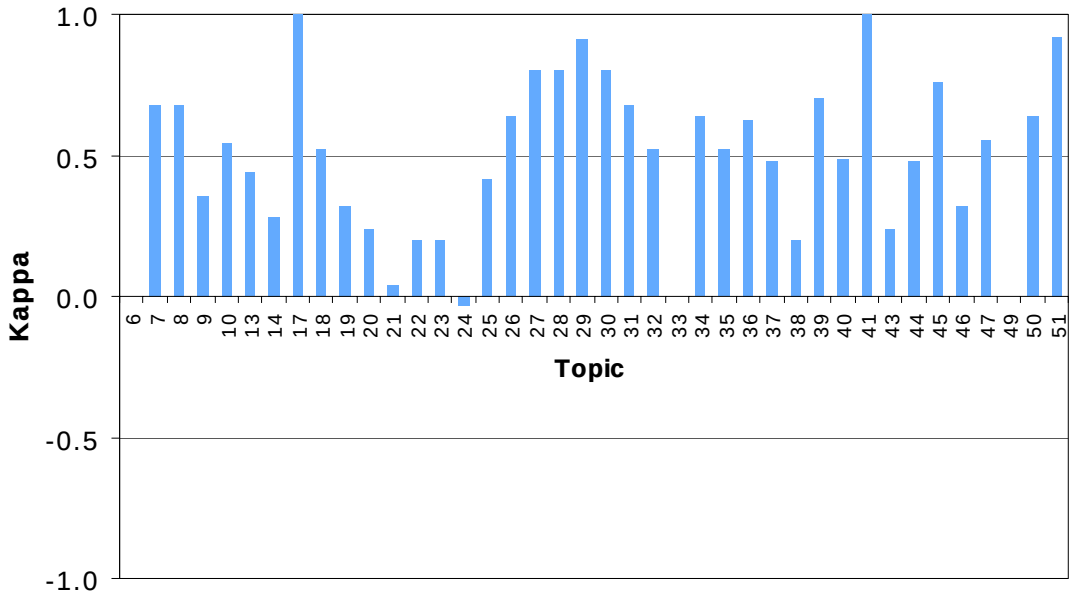

Figure 1. Chance-corrected inter-annotator agreement, by topic.

5. Results

Six participating sites submitted 31 ranked runs with no more than 5,000 documents per topic. Three of those runs applied a Boolean restriction when producing the document sets—those three runs consisted of substantially fewer than 5,000 documents for some topics. The baseline Boolean run, on the other hand, was not required to be ranked (although in practice it was first subjected to the Boolean constraint and then resulting Boolean set was ranked), so no upper bound on the size of the retrieved set was imposed in that case. The actual sizes of the submitted sets for the baseline Boolean run varied from 1 to 128,195 documents across topics. In addition to these 32 runs, the sets of approximately 100 documents found by the human expert for each topic (described in Section 4.1) were scored as if they were a 33rd run, (although as described below this comparison is not a fair one). Runs were given names beginning with an abbreviation that identified the submitting site. In this section, we briefly review the techniques used by each site; additional details can be found in the papers posted on the TREC Web site (http://trec.nist.gov).

- Hummingbird (hum). Hummingbird (now Open Text Corporation) submitted eight runs that explored the effects of alternative ways of formulating queries, different choices of index terms, and blind relevance feedback, plus the reference Boolean run (humL06B). The documents were indexed using the Livelink ECM-eDocs SearchServer system. The OCR field was indexed in every case, and all metadata was indexed together with OCR for seven runs, including the reference Boolean run (the exceptions being humL06dvo and humL06tvo). Queries were constructed automatically in six cases (the exceptions being humLo6B—the reference Boolean run, humL06t—the same run with a cutoff at 5,000, and humL06t0—a contrastive Boolean run using the first query in the negotiation history rather than the last query). For five of those six runs, the queries were automatically constructed from words in the Boolean queries (but without the use of Boolean or proximity operators); for the sixth run (humL06dvo) the queries were automatically constructed from the production request field.

- National University of Singapore (NUS). The National University of Singapore submitted two runs to explore the effects of evidence combination from multiple topic fields. The contents of the OCR field were indexed using the Lucene text retrieval system, and queries were formed from words found in the production request and the Boolean queries (but without the use of Boolean or proximity operators).

- Sabir Research (Sab). Sabir Research submitted seven runs to explore the effects of vocabulary filtering on OCR indexing and blind relevance feedback. The contents of the OCR and all metadata fields were indexed together using a vector space text retrieval system with pivoted document length normalization. Queries were formed from words in the production request and words in the Boolean Query for five of those runs; one run used only words from the production request (SabLeg06ar1) and one run used words from the production request, words from the Boolean query (without Boolean or proximity operators) and words from the Complaint (SabLeg06aa1).

- University of Maryland (Umd). The University of Maryland submitted four runs that explored the effects of different sources of query terms. The contents of the OCR and all metadata fields were indexed together using the Indri text retrieval system. Queries were formulated automatically for three runs: UmdBase (from words in the production request field), UmdBoolAuto (from words found in the final Boolean query, but without Boolean or proximity operators), and UmdComb (from both). For the fourth run (UmdBool), Indri queries were manually constructed to approximate the Boolean operators as closely as possible using Indri's query language (which does not directly support some required operators).

- University of Missouri-Kansas City (UMKC). The University of Missouri-Kansas City submitted eight runs that explored the effects of blind relevance feedback. The contents of the OCR field were indexed using the Lucene text retrieval system. Queries were formed automatically from words in the Boolean query (with Boolean operators, and sometimes also with proximity operators).

- York University (york). York University submitted two runs that explored the effects of blind relevance feedback. The contents of the OCR and all metadata fields were indexed together using Okapi BM 25 term weights. Queries were formulated automatically from words found in the Boolean query negotiation history (but without Boolean or proximity operators).

- Expert manual searcher "run" (EXPMANUAL). As described in section 4.1, the expert manual searcher used an interactive search system to identify up to 100 documents per topic that she felt would be unlikely to be retrieved by fully automated systems.

5.1. Uniques Analysis

One way of characterizing the results of different approaches to searching is to examine the contribution of each approach to the total set of known relevant documents. Figure 2 shows one way of looking at those statistics. As the grey bars show, on average across the 39 topics, 57% of the known relevant documents[3] were found by the reference Boolean query (i.e., either uniquely by the reference Boolean system, or by the reference Boolean system and also one or more other systems). As the analysis in Section 5.3 shows, our pooling strategy results in an underestimate of the actual number of relevant documents found by the reference Boolean system for topics with large numbers of relevant documents. Nevertheless, we this serves as a useful reference point from which to start an analysis of documents uniquely retrieved by other techniques.

The black bars stacked above the grey bars show the additional relevant documents found by the expert manual searcher but not by the reference Boolean system. On average across the 39 topics, the expert searcher found an additional 11% of the known relevant documents. In this case, the counts are accurate, since every document added to the pools by the expert searcher was judged. From this, we can conclude that by reformulating their query the expert searcher was able to find a substantial number of relevant documents that were not found by the reference Boolean system.

[3] In this section, and through the paper, the "known relevant documents" that we refer to are those judged as relevant by the primary assessor. Documents identified as relevant only by the second assessor in the inter-annotator agreement studies were not treated as relevant in the uniques analysis or when computing effectiveness metrics.

The white bars stacked above the black and grey bars show the additional relevant documents that were found by some system other then the reference Boolean system or the expert manual searcher. On average across the 39 topics, these other systems found an additional 32% of the known relevant documents. Our pooling strategy, which focuses on documents near the top of at least one ranked list and which includes no more than 100 documents from any one system, likely underestimates the number of relevant documents that ranked retrieval systems can find. Indeed, results for the "depth probe" run reported in the Hummingbird (Open Text) paper suggest that this underestimate may be substantial for at least some topics. Nonetheless, we can state with confidence that there were a large number of known relevant documents (1,417 across 39 topics) that were not found by the reference Boolean system or by the expert searcher. There was, therefore, scope for ranked retrieval systems to substantially outperform both the reference Boolean system and the expert manual searcher because there were a substantial number of known relevant documents that neither of those systems found. As we will see below, that did not happen.

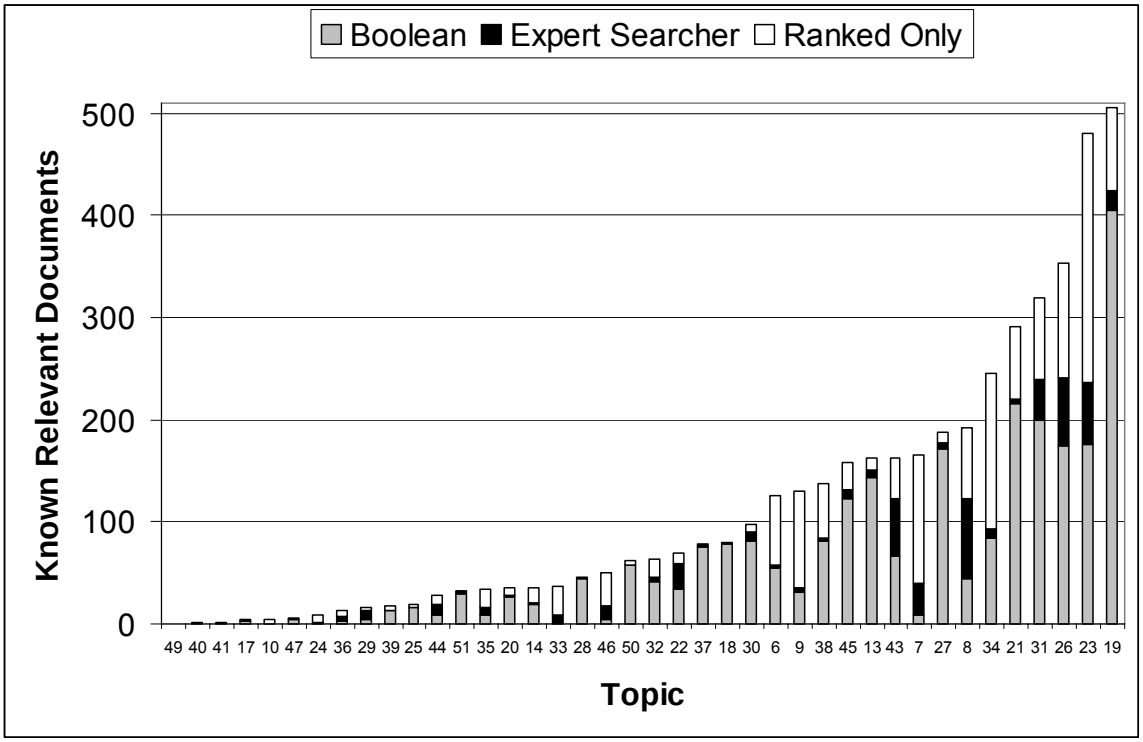

Figure 2. Known relevant documents found by the Reference Boolean system (grey), found by the expert searcher but not the reference Boolean system (black), and found uniquely by at least one other system (white).

5.2. R-Precision

Although our principal focus is on recall rather than precision, it is convenient to begin with a precision-oriented measures because precision-oriented measures are well understood, widely reported, and easily computed. Figure 3 compares the ranked retrieval runs using mean R-precision, a precision-oriented measure computed as the average across topics of the density of relevant documents at rank R (where R is the number of known relevant documents for that topic). The seven dark bars show the best scoring run from each participating team (and from the manual searcher). For comparison, all other runs (in order: the expert manual search, the reference Boolean run, and three Boolean runs from participating teams) are shown to the left of the ranked runs. Because R-precision is focused early in the ranked list, this measure would be expected to favor ranked retrieval systems. All four Boolean runs were, however, ranked in some way after being subjected to the Boolean constraint. The result is, therefore, in some sense fair in those cases. The expert human searcher "run" is disadvantaged in this comparison, however. It consisted of only

about 100 documents, those documents were not intentionally ranked by probability of relevance, and the searcher focused on finding diverse relevant documents to enrich the pool rather than the easiest relevant documents to boost measured effectiveness.

Three results are clearly evident in this data. First, the best runs from three of the participating sites were nearly indistinguishable by the R-precision measure, and one of those three runs (humL06t) was subjected to a Boolean constraint. Indeed, the reference Boolean run did about as well on this precision-oriented measure as the best unconstrained ranked retrieval runs. This is notable because Boolean runs can retrieve only documents that satisfy the Boolean query, while the ranked runs had no such constraint. From this we can conclude that (when averaged over 39 topics), little adverse effect resulted from respecting the Boolean constraint. Of course, with only six participating systems we are nowhere near exhausting the design space for search techniques, so ways may yet be found to achieve improvements that are not available to a Boolean system. All we can say at this point is that such improvements have not yet been demonstrated in the TREC legal track.

The second obvious result is that Boolean systems are not all created equal—two of the four Boolean runs did about twice as well (by this precision-oriented measure) as the other two! In one case (Hummingbird) this appears to result from using the initial rather than the final Boolean queries. In the other case (Maryland) the differences appear to result from incomplete support for extended Boolean operators. When we first proposed this track, one of our shorthand goals was to see if someone could "beat Boolean." This year's results indicate that might be easily achieved in the wrong way (by inadvertently creating an underperforming "Boolean" baseline), and that careful attention to the process by which the Boolean queries are created and used will be important if we are to produce meaningful comparisons.

Third, the expert manual searcher's submitted sets had, despite the factors discussed above that would tend to decrease R-precision scores, noticeably higher R-precision than any of actual submitted runs (all of which were essentially fully automatic, although in a few cases some query reformatting was done manually). This suggests that focusing attention on interactive search might yield interesting results.

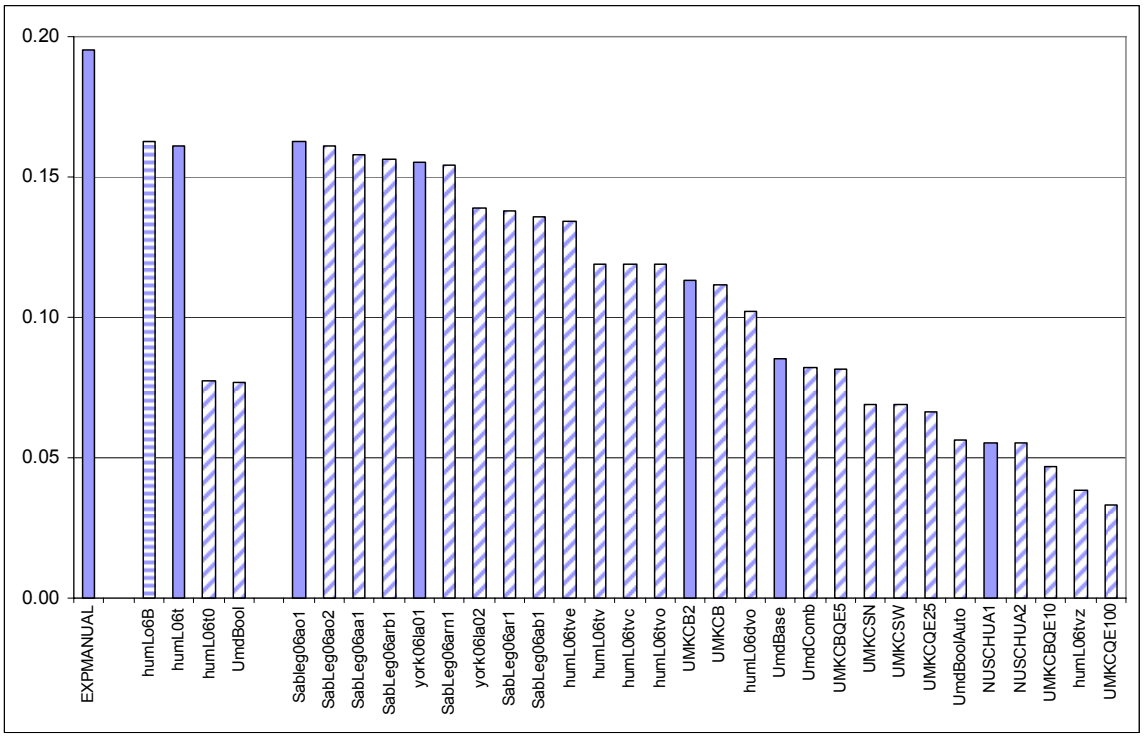

Figure 3. Mean precision at R (the actual number of known relevant documents for each topic). Ranked runs on left side, Reference runs on right side. Best run for each team shown as solid bar.

Runs EXPMANUAL and humL06tvz were not conventionally ranked and thus are disadvantaged by this measure.

5.3. P@B

A set-oriented comparison of ranked retrieval with the reference Boolean run was possible for 22 topics for which 5,000 or fewer documents were included in the Boolean set.[4] Let B be the size of the submitted set for the baseline Boolean run for a particular topic. The idea is to treat the top B documents of a ranked run for that topic as if it were a submitted set of size B and then compute P@B, the density of relevant documents in that set (treating unassessed documents as not relevant). Although the true number of relevant documents is not known, the precision at any fixed cutoff is proportional to the recall at that same cutoff, so we can interpret P@B for any individual topic as a measure of recall. Averaging across topics yields somewhat different results than a direct computation of recall would, however, since the constant of proportionality varies by topic.

In Figure 4 we compare P@B values for SabLeg06ao2 (one of the top-scoring runs by P@R) with those of the baseline Boolean run. For 12 of 22 topics, P@B favors the reference Boolean run, while for 7 of 22 the ranked run is favored. Three topics had tied values of P@B that were near 0.

The above analysis understates the true value of P@B since the assessed pools are incomplete and biased in favor of documents ranked highly by submitted runs. This problem is worse for a set-based measure like P@B than for measures like R-precision that focus on the documents closest to the top of a ranked list. We had no alternative to pooling for evaluating the ranked run, but for the baseline Boolean run an unbiased estimate of P@B could be computed using stratified sampling.

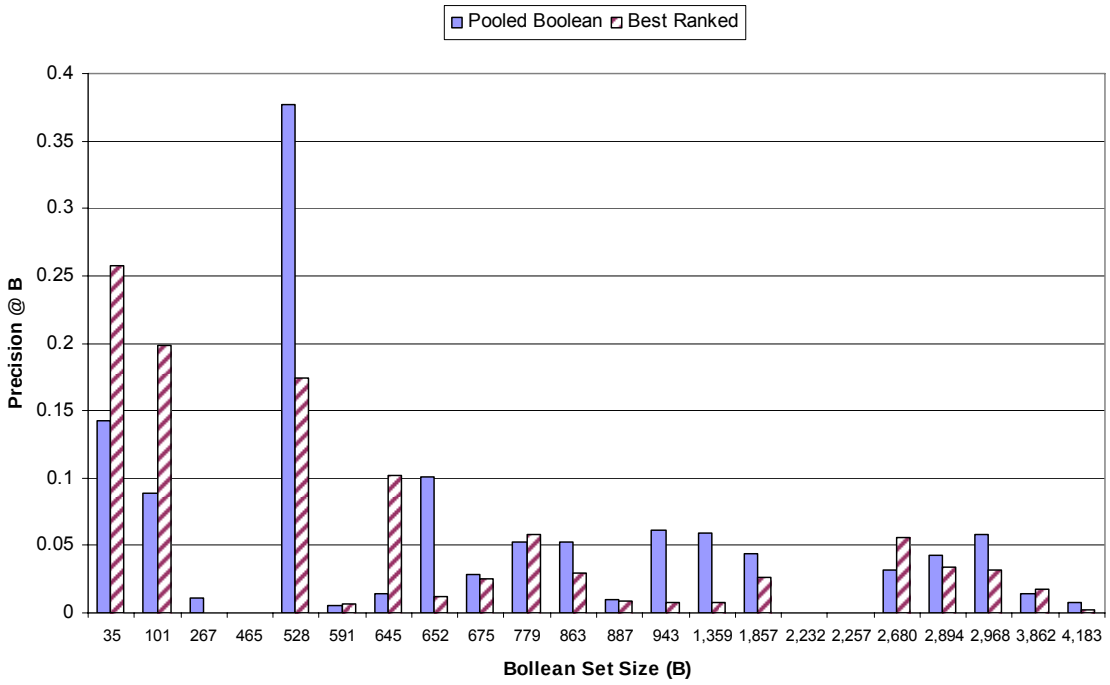

Figure 4. Recall-oriented effectiveness measure, by topic, in increasing order of Boolean set size. Topic 33 (for which B=1) not shown.

[4] There were actually 23 topics with B≤5,000, but using topic 33, for which B=1, would not be informative because when B=1 precision can only be 0 or 1. Precision at B for topic 33 was 0 for the reference Boolean run, and 1 for the best ranked run.

It turned out that the identity of the original stratified samples (Section 4.1) from the baseline Boolean run could not be recovered at the time that evaluation measures were computed because of a hardware failure. Further, the original stratification could not be reconstructed from the pools themselves because documents meeting the strata definitions could have come from ranked runs, the expert manual run, or the stratified sampling process. However, we were able to define new strata in a way that still allowed the computation of unbiased estimates of set-based effectiveness measures for the baseline Boolean run.

We separated the documents in the baseline Boolean set for each topic into four strata based on which other runs they occurred in:

- ***Stratum 0'***: Documents occurring in the top 100 of any site's main run, top 10 of any run from any site, or in the expert manual set.

- ***Stratum 1'***: Documents in former ***Stratum 1***, but not in ***Stratum 0'***.

- ***Stratum 2'***: Documents in former ***Stratum 2***, but not in ***Stratum 0'***.

- ***Stratum 3'***: Documents in former ***Stratum 3***, but not in ***Stratum 0'***.

By putting all documents added to the pool by a run other than the baseline Boolean run into ***Stratum 0'***, we can treat any remaining documents as if they had been drawn randomly from the newly defined strata. ***Stratum 0'*** is treated as having all its documents sampled, while the number of documents treated as sampled from ***Strata 1', 2',*** and ***3'*** varies by topic. We used the resulting stratified samples to produce unbiased estimates of P@B for the baseline Boolean run, as well as computing a 95% confidence interval for these estimates using the Gaussian approximation to the binomial (Lewis, 1996). Because these new strata generally contain fewer documents than under the original stratification, our estimates of P@B usually have a higher sampling variance than they would have with the original stratification.

As Figure 5 shows, pooling and stratified sampling produce the same estimate of P@B when B is at or below 267. The situation is quite different as B grows, however. In 10 of the 16 cases for which B is 528 or higher, and for which the pooled estimate of P@B is nonzero, the pooled estimate falls below the lower limit of the confidence interval on the stratified estimate. This result reinforces our earlier that our pool-based effectiveness measures do not provide a measure of the absolute effectiveness of any of the participating systems. Further, the large gap between the pool-based P@B and the true value (or at least an unbiased estimate of it) means more danger that biases in pool construction will affect even comparisons of relative effectiveness.

Analysis reported in the Hummingbird (Open Text) paper indicates that similar effects are present in at least the one ranked “depth probe” run for which a kind of stratified sampling was done (humL06tvz). Our future work on comparison of ranked and Boolean runs will require a more nuanced strategy than we have yet applied.

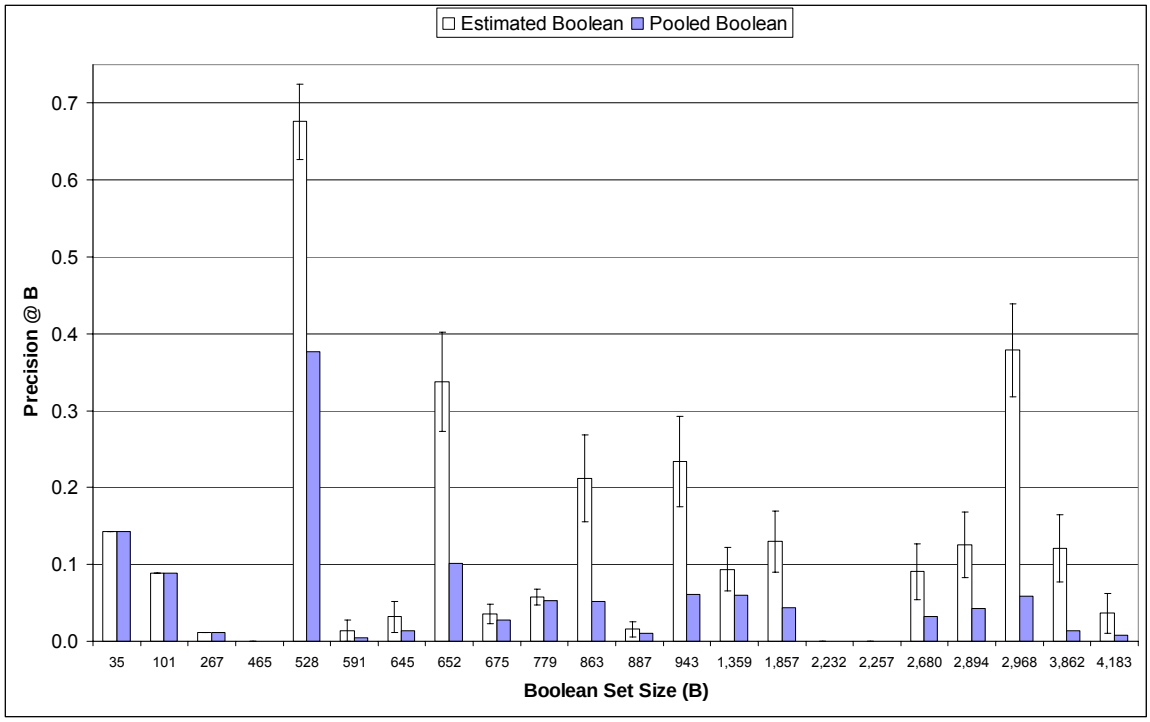

Figure 5. Comparison of stratified estimate of P@B with pool-based estimate of P@B.

6. Conclusion

This first year of the TREC Legal Track has produced a new test collection that models present practice in e-discovery, and that will also be of interest to researchers working on retrieval from scanned document images and to researchers working on the integrated use of structured metadata with document text as a basis for retrieval. Six research teams participated in the evaluation, contributing to the creation of relevance assessment pools that were judged in a manner representative of the human review process that precedes release in an e-discovery process. These judgments provide a basis for both this year's evaluation and for development of new approaches that are tuned to the unique characteristics of this task.

Analysis of the results yielded a number of useful insights. Perhaps the most striking result is the strong performance of the Boolean queries. The reference Boolean run did about as well (by R-precision) as the best ranked runs, and the top seven ranked runs (again, by R-precision) all used terms from the Boolean queries as part of their automatic query formulation process. This suggests that the negotiated Boolean queries are information-rich, which has implications both for practice (propounding Boolean queries is a productive activity) and for system design (leveraging manually constructed Boolean queries when they are available can yield improved retrieval effectiveness). A second important result is objectively quantifying the fact that there are many relevant documents to be found beyond those identified by strict application of negotiated Boolean queries. This should not be surprising, of course, since it is well known that formulating queries that are both sufficiently inclusive and sufficiently precise is difficult. Perhaps the most important implication of this observation is that exploring system designs based on relaxation of the Boolean query and based on augmenting queries using terms from other sources (e.g., the production request) may ultimately yield better retrieval effectiveness than strict application of Boolean logic. While that potential was not realized in the TREC 2006 legal track (at least not by the P@R measure), this year's relevance judgments are exactly what is needed to explore the space of possible system designs to determine whether such gains can indeed be achieved.

From the perspective of evaluation design, the clearest conclusion is that additional work on statistical estimation for recall-oriented measures is needed. The analyses in this paper and in the Open Text paper

indicate that statistical estimates of retrieval effectiveness for both the reference Boolean run and for one ranked run yield markedly different results from the more commonly used metrics in which unassessed documents are treated as not relevant. Additional analysis will be needed before we can directly compare those two runs, and the potential for statistical estimation for other ranked retrieval runs from 2006 is limited by the sampling strategies that were employed when forming the assessment pools. It will therefore be important to revisit both our choice of measures and our sampling strategies for the 2007 Legal Track.

Our focus in this first year of the Legal Track was on the design of automated systems, but of course automated systems are ultimately used by people. Our expert searcher run yielded some interesting insights, however, finding an average of 13 documents per topic that the reference Boolean query had missed and achieving better retrieval effectiveness (by the P@R measure) than any other run. This suggests that a focused effort to explore interactive search techniques in the TREC 2007 legal track might yield additional insights.

Perhaps the greatest accomplishment of the TREC 2006 Legal Track is that it happened at all. More than 50 volunteers contributed to assembling and distributing the collection, creating topics, developing systems, managing submissions, creating pools, judging relevance, developing metrics, creating scoring software, analyzing results, and coordinating those activities. This has yielded the results that we would hope for from any TREC track in its first year: (1) a reusable test collection to support future research, (2) a set of baseline results to which future research can be compared, and (3) a community of researchers who bring a variety of perspectives to these important challenges. The coordinators trust that a second year of research will continue to yield important results.

Acknowledgements

This track would not have been possible without the generous support of the IIT CDIP project (including Gady Agam, Shlomo Argamon, Ophir Frieder, and Dave Grossman, plus special thanks to David Roberts), the University of California at San Francisco Library (particularly Karen Butter, Albert Jew, Kirsten Neilsen, and Heidi Schmidt), members of the Sedona Conference®, and the volunteer relevance assessors and their participating firms and institutions. In particular, the coordinators wish to thank Ryan Bilbrey, Conor Crowley, Joe Looby, and Stephanie Mendelsohn, for their greatly appreciated assistance in writing draft complaints, in topic development, and for participating in "Boolean negotiations," and Anna Marshall at George Washington University School of Law for her extraordinary assessor recruitment efforts. Special acknowledgement is due Richard Braman, Executive Director of The Sedona Conference ®, for all of his assistance in facilitating a successful outcome of year 1 of the Legal Track. Thanks also to Michael Tacelosky and Keith Ivey of Tobacco Documents Online and *smokefree.net* for access to and help with their collection of tobacco documents and for their work on our assessment platform. Finally, special thanks also go to our colleagues at NIST for handling much of the logistics.

References

Agam, G., Argamon, S., Frieder, O., Grossman, D. and Lewis, D., "Complex Document Information Processing: Prototype, Test Collection, and Evaluation," *Document Recognition and Retrieval XIII*, SPIE Proceedings vol. 6067, pp. 60670N-1 to 60670N-11, 2006.

Baron, J., "Toward a Federal Benchmarking Standard for Evaluating Information Retrieval Products Used in E-Discovery," *Sedona Conference Journal*, vol. 6, pp. 237-246, 2005 (available from Westlaw and LEXIS)

Baron, J., Lewis, D., and Oard, D. " "How To" Guide for Assessors – TREC Legal Track 2006." Version 4, Aug. 20, 2006. *http://trec-legal.umiacs.umd.edu/TRECLegal_HowToGuide_Version4Final.doc*

Blair, D., and Maron, M. "An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System," *Communications of the ACM*, 28(3)289-299, 1985.

Buckley, C., Dimmick, D., Soboroff, I and Voorhees, E., "Bias and the Limits of Pooling," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 619-620, Seattle, 2006

Buckley, C. and Voorhees, E. "Retrieval System Evaluation," in *TREC: Experiment and Evaluation in Information Retrieval*, E. M. Voorhees and D. K. Harman, eds., MIT Press, pp. 53-75, 2002.

Cochran, W. *Sampling Techniques*, 3rd edition. John Wiley & Sons, New York, 1977.

Coleman v. Morgan Stanley, 2005 WL 679071 (Fla. Cir. Ct. Mar. 1, 2005)

Collaborative Expedition Workshop #45, *Advancing Information Sharing, Access, Discovery and Assimilation of Diverse Digital Collections Governed by Heterogeneous Sensitivities*, held Nov. 8, 2005, see http://colab.cim3.net/cgi-bin/wiki.pl?AdvancingInformationSharing_DiverseDigitalCollections_HeterogeneousSensitivities_11_08_05

Jacobs, P., "GE in TREC-2: Results of a Boolean Approximation Method for Routing and Retrieval," in *The Second Text Retrieval Conference (TREC-2)*, Gaithersburg, MD, August, 1993, pp. 191-200.

Lewis, D. "The TREC-5 Filtering Track," in *The Fifth Text Retrieval Conference (TREC-5)*, Gaithersburg, MD, November, 1996, pp. 75-96.

Lewis, D., Agam, G., Argamon, S., Frieder, O., Grossman, D., and Heard, J. "Building a Test Collection for Complex Document Information Processing," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 665-666, Seattle, 2006.

Lu, X., Allan, J., Miller, D. "Boolean Systems Revisited: Its Performance and Behavior," *The Fourth Text Retrieval Conference (TREC-4)*, pp. 459-474, Gaithersburg, MD, November, 1995.

Schmidt, H.; Butter, K.; and Rider, C. "Building Digital Tobacco Document Libraries at the University of California, San Francisco Library/Center for Knowledge Management," *D-Lib Magazine*, 8(2), 2002.

Sedona Conference, *The Sedona Principles: Best Practices Recommendations & Principles for Addressing Electronic Document Production* (2005 version), Principle 11, see http://www.thesedonaconference.org/content/miscFiles/publications_html

Turtle, H., "Natural Language vs. Boolean Query Evaluation: A Comparison of Retrieval Performance," *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 212-220, Dublin, 1994.

U.S. Federal Rules of Civil Procedure, Rules 26 & 34, as amended (Dec. 1, 2006)

Voorhees, E., "Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness," *Information Processing and Management*, 36(5)697-716, 2000.

Zubulake v. UBS Warburg, 217 F.R.D. 309 (S.D.N.Y. 2004)

Topic	n	n11	n01	n10	n00	Agree	Agree R	Agree N	Kappa
6	50	0	0	25	25	0.5	0	0.667	0
7	50	21	4	4	21	0.84	0.84	0.84	0.68
8	50	19	2	6	23	0.84	0.826	0.852	0.68
9	49	9	0	16	24	0.673	0.529	0.75	0.355
10	50	2	0	3	45	0.94	0.571	0.968	0.545
13	50	11	0	14	25	0.72	0.611	0.781	0.44
14	50	11	4	14	21	0.64	0.55	0.7	0.28
17	50	4	0	0	46	1	1	1	1
18	50	20	7	5	18	0.76	0.769	0.75	0.52
19	50	8	0	17	25	0.66	0.485	0.746	0.32
20	50	7	1	18	24	0.62	0.424	0.716	0.24
21	50	5	4	20	21	0.52	0.294	0.636	0.04
22	50	5	0	20	25	0.6	0.333	0.714	0.2
23	50	9	4	16	21	0.6	0.474	0.677	0.2
24	50	0	1	9	40	0.8	0	0.889	-0.037
25	50	7	6	5	32	0.78	0.56	0.853	0.414
26	50	21	5	4	20	0.82	0.824	0.816	0.64
27	50	22	2	3	23	0.9	0.898	0.902	0.8
28	50	21	1	4	24	0.9	0.894	0.906	0.8
29	50	16	1	1	32	0.96	0.941	0.97	0.911
30	50	22	2	3	23	0.9	0.898	0.902	0.8
31	50	23	6	2	19	0.84	0.852	0.826	0.68
32	50	20	7	5	18	0.76	0.769	0.75	0.52
33	50	0	0	25	25	0.5	0	0.667	0
34	50	20	4	5	21	0.82	0.816	0.824	0.64
35	50	14	1	11	24	0.76	0.7	0.8	0.52
36	50	9	3	4	34	0.86	0.72	0.907	0.627
37	50	14	2	11	23	0.74	0.683	0.78	0.48
38	50	17	12	8	13	0.6	0.63	0.565	0.2
39	50	15	4	3	28	0.86	0.811	0.889	0.7
40	50	1	2	0	47	0.96	0.5	0.979	0.485
41	50	1	0	0	49	1	1	1	1
43	50	10	4	15	21	0.62	0.513	0.689	0.24
44	50	12	0	13	25	0.74	0.649	0.794	0.48
45	50	19	0	6	25	0.88	0.864	0.893	0.76
46	50	8	0	17	25	0.66	0.485	0.746	0.32
47	50	4	3	2	41	0.9	0.615	0.943	0.558
49	50	0	32	0	18	0.36	0	0.529	0
50	50	19	3	6	22	0.82	0.809	0.83	0.64
51	50	24	1	1	24	0.96	0.96	0.96	0.92
MEAN						0.765	0.627	0.810	0.490

Table 1: Raw contingency table entries from interassessor comparison study. We show agreement, i.e. *(n00 + n11)/n*, agreement on relevant, i.e. *2*n11 / (2*n11 + n01 + n10)*, agreement on nonrelevant, i.e. *2*n00 / (2*n00 + n01 + n10)*, and Cohen's kappa.

Top	pool	E[n11]	E[n01]	E[n10]	E[n00]	~E[Agree]	~E[AgreeR]	~E[AgreeN]	~E[Kappa]
6	840	0	0	125	715	0.851	0	0.92	0
7	854	138.6	110.2	26.4	578.8	0.84	0.67	0.894	0.57
8	857	145.9	53.2	46.1	611.8	0.884	0.746	0.925	0.671
9	849	46.8	0	83.2	719	0.902	0.529	0.945	0.488
10	858	2	0	3	853	0.997	0.571	0.998	0.57
13	837	71.3	0	90.7	675	0.892	0.611	0.937	0.559
14	716	15.8	108.8	20.2	571.2	0.82	0.197	0.899	0.129
17	767	4	0	0	763	1	1	1	1
18	769	64	192.9	16	496.1	0.728	0.38	0.826	0.263
19	919	161.6	0	343.4	414	0.626	0.485	0.707	0.298
20	938	9.8	36.1	25.2	866.9	0.935	0.242	0.966	0.209
21	893	58.2	96.3	232.8	505.7	0.631	0.261	0.754	0.046
22	853	13.8	0	55.2	784	0.935	0.333	0.966	0.315
23	832	173.2	56.3	307.8	295.7	0.563	0.487	0.619	0.183
24	924	0	22.3	9	892.7	0.966	0	0.983	-0.014
25	961	11.1	148.7	7.9	793.3	0.837	0.124	0.91	0.092
26	935	297.4	116.2	56.6	464.8	0.815	0.775	0.843	0.62
27	916	165.4	58.2	22.6	669.8	0.912	0.804	0.943	0.747
28	910	38.6	34.6	7.4	829.4	0.954	0.648	0.975	0.625
29	875	16	26	1	833	0.969	0.542	0.984	0.529
30	781	85.4	54.7	11.6	629.3	0.915	0.72	0.95	0.672
31	707	294.4	92.9	25.6	294.1	0.832	0.832	0.832	0.668
32	770	51.2	197.7	12.8	508.3	0.727	0.327	0.828	0.225
33	570	0	0	37	533	0.935	0	0.966	0
34	810	196	90.4	49	474.6	0.828	0.738	0.872	0.611
35	542	19	20.3	15	487.7	0.935	0.519	0.965	0.484
36	872	9	69.7	4	790.3	0.916	0.196	0.955	0.175
37	863	43.7	62.8	34.3	722.2	0.887	0.474	0.937	0.412
38	741	93.2	289.9	43.8	314.1	0.55	0.358	0.653	0.118
39	887	15	108.6	3	760.4	0.874	0.212	0.932	0.183
40	832	1	33.9	0	797.1	0.959	0.056	0.979	0.053
41	876	1	0	0	875	1	1	1	1
43	820	64.8	105.3	97.2	552.7	0.753	0.39	0.845	0.236
44	821	13.4	0	14.6	793	0.982	0.649	0.991	0.641
45	755	120.1	0	37.9	597	0.95	0.864	0.969	0.834
46	627	16	0	34	577	0.946	0.485	0.971	0.464
47	733	4	49.6	2	677.4	0.93	0.134	0.963	0.121
49	983	0	629.1	0	353.9	0.36	0	0.529	0
50	756	47.1	83.3	14.9	610.7	0.87	0.49	0.926	0.426
51	936	31.7	36.2	1.3	867.8	0.96	0.628	0.979	0.61
mean	825					0.854	0.462	0.901	0.396

Table 2: Stratified estimates of what the interassessor agreement contingency table values would be on the full pools, along with approximate expected values of agreement, agreement on relevant, agreement on nonrelevant, and kappa.