TREC-2008 LEGAL TRACK
Interactive Task – Guidelines

Abstract

The TREC-2008 Legal Track features a completely revised Interactive Task that is quite different from the 2007 pilot Interactive Task. The modifications are designed to enable the task to model more completely and accurately the conditions and objectives of e-discovery in the real world. Among the most important modifications being introduced are: (1) the designation of a single individual (an attorney) to act as the authority for defining the intent and scope of a topic; (2) a provision that allows participants to engage with that authority for purposes of clarifying relevance to a topic; and (3) the specification of the task objective to be, for each topic, a binary assessment (relevant, not relevant) of all documents in the target collection.

The goal in introducing a more realistic task design is twofold: on the one hand, to provide the IR community with a better view into the conditions and objectives of document discovery in the legal domain; on the other hand, to foster greater awareness in the legal community (and among the e-discovery firms that support that community's document-retrieval needs) of the capabilities currently being explored by IR researchers. More generally, it is hoped that the task will encourage greater communication and collaboration among the IR, legal, and e-discovery communities in addressing today's increasingly steep document-discovery challenges.

Specific guidelines for registering for and executing the task are contained in the appendix to the task description.

Introduction

The ad hoc task of the TREC Legal Track is designed to facilitate comparison of retrieval systems by controlling as many conditions as possible, which necessarily leaves unmodeled a number of important aspects of the conditions in which text retrieval is actually conducted in the e-discovery domain. The 2007 Legal Track therefore introduced an "Interactive Challenge Task." The goal of that new task was to model more realistically the way in which retrieval queries might be generated, refined, and applied in the e-discovery domain; the task achieved this goal by making greater allowance for iterative human development and refinement of the queries intended to retrieve documents relevant to a target topic.

The results of the task, as discussed at TREC's November 2007 meeting, proved interesting on a number of points; chief among them were (a) that a team taking an iterative approach to the manual development and refinement of queries could achieve a relatively high rate of agreement with the relevance assessments made by an independent adjudicator and (b) that, within the constraints set by the task design, different teams, even teams from the same participating institution and following the same instructions, showed considerable variation in results.[1]

In 2008, the Legal Track will build on these results by introducing further modifications to the design of the Interactive Task, modifications that will enable the task to model still more closely and completely the conditions and goals of e-discovery in the real world. The purpose of this paper is to sketch the modified design.

The paper is organized as follows. We begin with an overview of the revised task protocol; here, we review the key objectives of the task and sketch the main features of a protocol we believe will meet those objectives. This section should suffice to give readers a view of the key modifications being proposed for the 2008 version of the task. We follow the overview with a

[1] Stephen Tomlinson, Douglas W. Oard, Jason R. Baron, Paul Thompson. Overview of the TREC 2007 Legal Track.

closer look at each of the main phases of the task, reviewing specific provisions for the key activities of (i) relevance definition, (ii) implementation, (iii) adjudication, and (iv) measurement. We conclude by summarizing what we see as the chief advantages and chief limitations of the task design. An appendix contains information on specific steps for registering and executing the task..

1. Overview of Protocol

1.1. Objectives of the Design

The goal of the Interactive Task is to model more accurately and more completely the real-world conditions in which companies and law firms, and the e-discovery firms they engage, must meet their document-retrieval objectives and obligations.

Pursuant to that goal, an objective for the 2008 design is to provide a more "end-to-end" representation of the task of e-discovery in the real world, incorporating aspects of the challenge of e-discovery not previously modeled. More specifically, the task will include, on the front end, the opportunity for participants to engage with an authority on the target topic in order to define more concretely the set of documents that are relevant to the topic (thereby modeling a crucial initial step in the e-discovery process). The task will also include, on the back end, the requirement that participants make a final binary assessment of the relevance of every document in the population (thereby modeling more closely the deliverable that an e-discovery firm, helping a client meet its production obligations, must make).

While meeting this modeling objective, the 2008 design must also satisfy other requirements. It must be feasible in terms of the resources required; it must be capable of completion within the timeframe set for other TREC tasks; and, more generally, it must be in accord with and support TREC's fundamental mission.

One of the guiding goals of the TREC series of workshops is "to increase communication among industry, academia, and government by creating an open forum for the exchange of research ideas." We hope that, by modeling the conditions of e-discovery more accurately and completely, the 2008 Interactive Task will encourage greater communication and collaboration among academic and industry researchers in the e-discovery domain.

1.2. Overview of the Design

When the lead attorney on a matter, the legal architect for a case, oversees a document production (or a document-retention or issue-coding effort), he or she will have formed, or be in the process of forming, an opinion as to what is responsive or relevant and what is not; when that attorney employs the products or services of an e-discovery firm, or, for that matter, the services of a traditional manual-review team, he or she does so with the goal of efficiently applying that conception of relevance across the full document population implicated by the matter. The e-discovery firm is not asked to consider, weigh, and resolve differences between all possible conceptions of relevance; the e-discovery firm is asked to replicate, across the document population, one conception of relevance, that of the legal architect who has hired the firm and who bears ultimate responsibility for the validity of the production.[2] The goal, therefore, of an e-discovery firm engaged to assess the relevance or responsiveness of documents in a population of interest is to replicate the relevance assessments the senior litigator in the matter would make, if

[2] Indeed, the Sedona Conference, in its commentary on best practices in e-discovery, has recognized that a key to the success of any retrieval effort is the legal architect's on-going and close involvement in the effort. Cf. The Sedona Conference® Best Practices Commentary on the Use of Search & Information Retrieval Methods in E-Discovery. *The Sedona Conference Journal*. 2007.

he or she had the time and leisure needed to review for relevance every document in the population.

The 2008 Interactive Task is designed to model this task and to measure participants' ability to perform it. In this section, we provide a brief overview of the proposed task design, summarizing key players, task workflow, and the results that can be expected from the implementation of the design.[3]

Key players. The following are the key roles required to implement the design.

- **Track Coordinators.** Lay the groundwork for the task; oversee its execution; analyze and prepare the track overview paper.
- **Topic Authorities.** Serve as an authoritative source of information for participating teams seeking to develop definitions of relevance; serve as final arbiters of the samples reviewed to measure participating teams' effectiveness. The Topic Authority represents the senior litigator who engages the services of an e-discovery firm.
- **Participating Teams.** Teams from academia and industry who endeavor to replicate the relevance assessments that would be made by a Topic Authority and report on their results in a paper for the TREC 2008 conference.
- **Document Reviewers.** Support the evaluation effort by reviewing samples of documents for relevance to the target topics (under the guidance of a Topic Authority).

Topic Authorities should be partner-level attorneys who have overseen large document-review efforts and who in court have vouched for the completeness and accuracy of a document production. Participating Teams are free to decide on their own composition, putting together whatever skills and expertise they believe will best enable them to perform the task. Document Reviewers should be attorneys (or law students) experienced in document review.[4]

Task workflow. The workflow of the task can be divided into three primary phases: (1) task set-up; (2) task implementation; and (3) results evaluation. A high-level summary of the steps to be taken at each phase follows.

Under task set-up fall all the steps that must be taken to lay the groundwork required for Participating Teams to execute the task. The following is a summary of the key steps taken in this phase.

- **Track Coordinators select target topics.** Given the effort required, by all parties, to implement the protocol, it is expected that the number of target topics will be small, probably two to three. In a larger-scale implementation of the protocol, there may be allowance for a greater number of topics; each Participating Team would be free to take on as many topics as it believed it had the resources for.
- **Track Coordinators select Topic Authorities and assign them topics.** Given the time that will be required of a Topic Authority to perform his or her duties, a Topic Authority should not be assigned responsibility for more than one topic.

---

[3] It should be noted that, as the 2008 run will be the first implementation of the revised design, and so is intended to function as an experimental run, it is expected that this run will be somewhat scaled back in terms of topics and participants. Should it be decided to go forward with an additional run in a future year, allowance may be made for a greater number of topics and participants.

[4] The question of recruitment of individuals to fill these roles is an important one, but one that is best addressed separately from this paper (which is focused on design considerations).

- **Track Coordinators assign Participating Teams to Topic Authorities.** A single Topic Authority may not be capable of meeting the knowledge-transfer needs of all Participating Teams; in this circumstance, we could adopt one of two possible solutions: we could either (1) allow there to be multiple Topic Authorities for a single topic or (2) limit to a number suitable for a single Topic Authority the number of teams that could work on a given topic. In the former case (multiple Topic Authorities for a single topic), more teams could work on the same topic; the results, however, of any one Participating Team would be directly comparable only to results of other teams assigned to the same Topic Authority.[5] In the latter case (no more than one Topic Authority per topic), the question of comparing the results of teams using different Topic Authorities is not raised; we would, however, need to assemble a larger pool of candidate topics.

Under task implementation fall all the steps Participating Teams follow in executing the task. Key steps for this phase are the following.

- **Participating Teams work with Topic Authorities to clarify and further specify topic definitions.** As in the 2007 Interactive Task, teams will be provided with materials (mock complaints and requests for production) that provide an initial characterization of the topics of interest. In the 2008 Interactive Task, modeling real-world opportunity and practice, provision will be made for teams to arrive at further clarification and specification of the target topics through communication with their designated Topic Authorities. Teams are free to choose the method they think best for eliciting these clarifications from the Topic Authority. The Topic Authority should not share clarifications obtained by one team with any other team; it is the responsibility of each team to elicit its own clarifications. Out of respect for Topic Authorities' time, and in imitation of real-world constraints, there will be limitations on the amount of time allowed for this communication; the limitation will take the form of a maximum number of hours (10 hours per topic) that a team can ask of a Topic Authority.
- **Participating Teams employ their chosen processes and technologies to retrieve documents relevant to the target topics.** Participating Teams are free to use any process or technology they believe will allow them accurately and completely to identify the target set of documents. Time permitted for the task will be unconstrained, subject only to the deadline for delivery of results.
- **Participating Teams deliver their results.** When an attorney vouches for the validity of a document production, he or she is vouching for the accuracy of a binary classification of the document population implicated by the litigation, a classification into the subset of the population that is responsive to the requests for production and the subset that is not. When an e-discovery firm supports an attorney in this effort, it must make a similar relevance determination. The 2008 task, modeling this requirement, specifies that each team's final deliverable will be a binary classification of the full population for relevance to each target topic. Teams are of course free to use relevance ranking as a means to arrive at their result sets, but the final deliverable must be a single binary classification (relevant/not relevant) of the full population of documents.

Under results evaluation fall all the steps taken to review the results submitted by Participating Teams and to obtain measures of the teams' effectiveness in performing the task. Key steps under this rubric are the following.

---

[5] Note that a side benefit of this approach is that it could provide insight into how similarly or dissimilarly different Topic Authorities defined the same topic.

- **Document Reviewers review samples of documents selected for the purpose of measuring the effectiveness of Participating Teams.** Document samples will be selected for the purpose of obtaining measures of Participating Teams' effectiveness in performing the task.[6] Under the instruction and guidance of Topic Authorities, Document Reviewers will review their assigned samples for relevance to their assigned topics.
- **Topic Authorities make a final adjudication of any disputed sample assessments.** Once sample review is complete, Participating Teams will be given access to sample results, allowing them to review any mismatches between their assessments and those of the Document Reviewers; teams will not, at this stage, be given access to the results submitted by any other team or be informed as to the stratum from which any given document was drawn (thereby preventing teams from knowing the impact of any particular document assessment on their final metrics). Teams will be permitted to appeal any Reviewer assessments they believe are directly and specifically contradicted by information given them by the Topic Authority during the relevance definition phase; they will not be permitted to appeal assessments that represent differences in interpretation.[7] The court of appeal and the final arbiter is the Topic Authority.
- **Track Coordinators use sample results to obtain measures of the effectiveness of Participating Teams.** Once sample assessments are finalized, Track Coordinators will use sample results to obtain estimates of the recall, the precision, and, as a summary index of effectiveness, the balanced F-measure achieved by each Participating Team on each topic. These estimates will be included in the track overview paper and provided to the participants for their use in preparing their site report papers.

Expected results. Implementation of the protocol outlined above should enable a number of positive results.

- We should be able to obtain informative measures of the effectiveness of participating teams in performing a task that closely models conditions and goals of e-discovery in the real-world.
- Information scientists who participate should gain greater insight into the information-retrieval needs and conditions of the legal community and of the e-discovery firms that support that community.
- The more complete and accurate representation of the task of e-discovery should encourage greater interest and participation of e-discovery firms in the TREC Legal Track.
- By requiring greater participation by the legal community (e.g., as Topic Authorities), the protocol may foster greater awareness, among members of the community, of the work of TREC and of the relevance of that work to the document management and discovery challenges they face on a daily basis.

In the remainder of this document, we elaborate further on specific aspects of the protocol.

---

[6] Below we propose a sampling design we believe will permit the accurate and efficient estimation of performance. There are, however, a number of sampling designs that could suit the purposes of the task.

[7] It is up to the team to constrain, to the extent possible, the room for such differences in interpretation during the relevance definition phase.

2. Protocol Specifics

As already noted, the revised task design incorporates a number of elements not included in the 2007 running of the task. In the following, we further define the protocol for these new elements, starting with the provision for relevance definition (2.1), then turning to implementation and delivery (2.2), adjudication (2.3), and the sampling and metrics protocol (2.4).

2.1. Relevance Definition

The objective of an e-discovery process is to replicate, over the entire collection of documents implicated by the litigation in question, the relevance assessments that the lead attorney litigating the matter would make. In order to meet that objective, those implementing the process will need some guidance as to what the lead attorney would count as relevant and what he or she would not. Court-filed documents (Complaints, Requests for Production, Responses to Requests for Production, etc.) can provide some direction, but typically there is both the need and the opportunity for those providing e-discovery services to gain further direction from the lead attorney. One modification of the 2008 task design is make provision for such relevance clarification.

As in the 2007 running, participants will be supplied with materials (mock complaints and requests for production) that will provide direction as to the scope and intent of the target topics; new to the 2008 running, Participating Teams will also be given the opportunity to engage with a designated Topic Authority in order gain further clarification as to the proper scope and intent of the topic. The rules governing this engagement are as follows.

- A Participating Team is free, subject to the constraints specified below, to choose the form of interaction (telephone interview, assessment of example documents, etc.) with the Topic Authority that will suit their purposes best.
- A Participating Team may not communicate with a Topic Authority other than the one(s) to which they have been assigned; a Participating Team may consult a Topic Authority only on the topic for which the given Authority is responsible.
- A Participating Team may consult, in addition to the Topic Authority, any additional resources it chooses as a way to learn more about a topic, and the team is free to discuss those resources with the Topic Authority. Teams should understand that, for purposes of appealing an assessment in the adjudication phase, the only points of reference will be the clarifications gathered from the designated Topic Authority.
- Communication with a Topic Authority may be conducted by email, by telephone, or by internet meeting. In order to avoid giving any possible unfair advantage to participants with the resources for travel, participants may not schedule in-person meetings with a Topic Authority.
- A Topic Authority may not share information provided to one participant with any other participant (unless, of course, that other participant has independently asked for the same information). It is the responsibility of each Participating Team to obtain its own clarifications.
- The amount of time that a Participating Team may ask of a Topic Authority is limited as follows.
    - There is a maximum number of hours that a Participating Team may ask of a Topic Authority for clarification purposes; this maximum is set at 10 hours. That time allotment would include any time a Topic Authority spends in response to a Team's

request (time on the telephone, time responding to emails, time reviewing documents sent by a Participating Team, and so on).

We believe that, if implemented in accordance with these rules, this component of the task will achieve the goal of modeling real-world conditions while remaining within practical constraints.

2.2. Implementation & Delivery

With regard to the execution of the task, the 2008 running, like that of 2007, imposes few restrictions. Participating Teams may be of any size and configuration. Teams may avail themselves of any process or technology they believe will enable them to perform the task; as in 2007, teams are free to make use of publicly accessible web applications for searching the target document population.[8] The only substantive constraint on execution is one of time: teams must submit their results by the deadline set for doing so (date to be specified prior to the start of the task).

With regard to objective of the task, the deliverable asked of participants, the 2008 running does, as noted above, introduce some modifications intended to model more accurately the real-world deliverable of an e-discovery process.

When an attorney certifies that the client he or she represents has met its document-discovery obligations, he or she is certifying, in almost all instances, not merely that the client has produced *some* documents responsive to the opposing party's requests (however "good" those documents may be), but that the client has produced *all* (non-privileged) documents responsive to the requests (or at least made a reasonable, good-faith, effort at doing so). What the attorney is certifying, therefore, is, in large part, that, in the not-produced part of the population, few to no responsive documents remain; and it is on the not-produced part of the population that any challenge to the adequacy of the production is likely to focus, for that is the part of the population that is the most likely site of sanctionable document-discovery deficiencies.[9] Any e-discovery firm that would help an attorney meet his or her document-discovery objectives must enable the attorney to make this certification and to answer any challenges to its validity.

The attorney's deliverable, then, and the deliverable of the e-discovery firm that would support the attorney, can be thought of as a binary classification of the entire document population implicated by the litigation into two subsets: (i) the responsive subset (to be produced) and (ii) the nonresponsive subset (not to be produced). What the attorney, and the e-discovery firm, must be able to claim about this deliverable is that the subset classified as responsive is largely free of documents that are actually nonresponsive and that the subset classified as nonresponsive is, to an extent commensurate with a reasonable good-faith effort, free of documents that are actually responsive. This is the deliverable we seek to model in the 2008 running of the Interactive Task.

Modeling this deliverable is for the most part straightforward: we simply require that each Participating Team make, for each target topic, a binary assessment (responsive, not responsive) of every document in the population. Teams may make use of relevance ranking, or any other approach they choose, as a means of arriving at their submitted assessments; what will matter, however, in terms of measuring effectiveness, is just the final binary classification. Each team's effectiveness in making that binary classification will then be measured by obtaining estimates of the recall and precision achieved.

---

[8] One such system is the Legacy Tobacco Document Library (http://legacy.library.ucsf.edu/); this is the system used by all three 2007 Interactive Task participants. LTDL is not, however, designed to produce full-collection result sets, so some post-processing will likely be required.

[9] The produced part of the population coming into play in cases of egregious overproduction.

Now, while the requirement is straightforward enough, and certainly justifiable as a representation of real-world objectives, we recognize that an exhaustive classification of the target population may not be feasible, in the time allotted for the task, for all would-be participants. The time required by some approaches to document-retrieval is more volume-dependent than is the time required by others, and we do not want to exclude from participation teams that would like to test an approach that, though of practical or scientific interest, would not permit the assessment of the full population in the time set for the execution of the task. Cases of incomplete assessment of the full population will be treated as follows.

- Any document that does not receive a positive assessment (i.e., any document that a team has not assessed as relevant), whether because the document has been assessed and found to be non-relevant or because the document has not been assessed at all, will be treated, for purposes of adjudication and task-wide summarization of results, as having been given a negative assessment. As a consequence, a process that does not assess the full population may receive a lower (full-population) recall score than it would have received had it been able to assess the population completely.
- Any team that does not complete assessment of the full population is free, in its individual summary of findings, to report its performance on just that part of the population that was given a definitive positive or negative assessment.[10] This will enable a scalability-independent view of the effectiveness of a tested process.

We believe that this approach will allow a meaningful modeling of real-world requirements while still permitting participation from teams employing volume-dependent approaches.[11]

2.3. Adjudication

As we have seen, the objective for participants in this year's Interactive Task is to replicate, across the full document population (or across a sample of it), the relevance assessments that a designated Topic Authority would make, if he or she had the time required, across that same document population. As we have also seen, each Participating Team is given the opportunity to engage with their designated Topic Authority for the purpose of clarifying the understanding of relevance that they should replicate. The measurement of a team's effectiveness in carrying out the task will be based on samples of documents that have been assessed by Document Reviewers under the direction of the appropriate Topic Authority. Now, it is to be expected that the direction provided by the Topic Authority will go some way to ensuring that the assessments entered by the Document Reviewers are consistent with the Topic Authority's understanding of relevance; it is not to be expected, however, that it will go all the way to ensuring that outcome. We therefore think it both necessary and fair to include a provision whereby Participating Teams, under certain conditions, can appeal the relevance assessments rendered by Document Reviewers.

The procedures for appealing sample assessments and for reaching final adjudication of them is as follows.

- Samples are selected in accordance with the design sketched below.

---

[10] The Track Coordinators may provide software for computation of an alternative score if a consensus is reached on a specific alternative that participants would prefer.

[11] Note that the nature of the objective also has implications for the way a Participating Team can submit its results. Because the assessment is binary (responsive, not responsive), and because any unassessed document will be counted as a negative assessment (not responsive), a complete list of a Participating Team's positive assessments will suffice to determine the team's performance on the task. What Participating Teams will be asked to submit, therefore, will be simple lists of the documents (identified by the DOCNO) that have been assessed positively for each target topic.

- Samples are reviewed by Document Reviewers under the direction of the Topic Authority. The Topic Authority's guidance takes the form of (a) an initial training session for the Document Reviewers led by the Topic Authority, (b) ongoing clarification Q&A between Document Reviewers and their Topic Authority, and (c) occasional spot checking of the Document Reviewers' assessments by the Topic Authority.
- Results of the sample assessment are given to Participating Teams to review.
  - In reviewing these results, a Participating Team has access to the assessments entered by the Document Reviewers and to its own submitted assessments.
  - In reviewing these results, a Participating Team does not have access to the assessments returned by any other Participating Team or to information regarding the stratum from which any given document was drawn; a Participating Team will therefore not be able to determine the impact overturning a given assessment will have on its final results.
- Participating Teams will be permitted to appeal any Reviewer assessments they believe are directly and specifically contradicted by information given them by the Topic Authority during the relevance definition phase.
  - A Participating Team may appeal only assessments that are directly and specifically contradicted by information given them by the Topic Authority.
  - A Participating Team may not appeal an assessment that represents only a difference in interpretation; it is the responsibility of the team, during the relevance clarification stage, to minimize the scope for such differences in interpretation.
  - A Participating Team may not appeal an assessment on the basis of information gathered independently of the Topic Authority. If, of course, during the relevance clarification phase, the Topic Authority has had the opportunity to opine on the significance of information a team has independently gathered, then that opinion can serve as the basis for an appeal.
  - In submitting any appeal, a team must clearly specify the grounds of its appeal, referring both to the pertinent parts of the document in question and to the pertinent information provided in the relevance clarification stage.
  - A 9-day window will be specified during which teams may make their appeals; begin and end dates for the appeal window will be specified prior to the start of the task.
- The Topic Authorities consider all appeals of assessments for which they have responsibility and render a final relevance assessment.
  - The Topic Authority's (post-appeal) assessment is final and not open to further appeal.

2.4. Measurement

Once participants have submitted their results, we will want to gauge how well they have performed their assigned task. We will want, in order to do so, first, to identify the metrics best suited to gauging effectiveness in performing the task and, second, to design a sampling protocol that will allow us efficiently to obtain accurate estimates of performance achieved on the target measures.

2.4.1. Metrics

We will use three metrics to gauge effectiveness in performing the Interactive Task: recall, precision, and, as a summary measure of effectiveness, the balanced F-measure (F_1).

In looking at the results of a relevance-assessment effort, the attorney who initiated the effort will want to know the answer to two primary questions. (1) Does the result set capture a high proportion of the material I am looking for (and so put me in a position to meet my obligations and to answer any challenges brought by opposing counsel)? (2) Does the result set minimize the amount of non-relevant material I have to review or preserve (thereby saving me from a wasteful and ineffective review and allowing me to make a more complete purge of unwanted material)?

Of the metrics available, recall and precision provide the most direct and meaningful answers to these questions. Because recall and precision condition on the marginal sums directly pertinent to the questions (total actually relevant documents, total documents assessed as relevant), they serve as the most sensitive gauges of the aspects of performance that matter most to an attorney looking to employ the products or services of an e-discovery firm.

A goal of the Interactive Task is to compare the effectiveness of the different approaches that have been taken to performing the task. While recall and precision are sensitive and informative performance metrics, it is also desirable, for purposes of making comparisons, to have a single summary measure of the effectiveness of each of the approaches tested. For this purpose, we use the balanced F-measure.

The F-measure is a function of the aspects of performance that are measured separately by recall and precision and is intended to provide a single measure of the overall performance of a system. When recall and precision are given equal weight, the F-measure resolves to the harmonic mean of the two metrics ($= 2/(R^{-1} + P^{-1}) = 2RP/(R + P)$). It is also possible, however, to make use of variants of the F-Measure that give unequal weights to recall and precision, should conditions dictate giving priority to one aspect of performance over the other.

In e-discovery, both recall and precision are important, and the relative importance of each may vary from project to project and even, within a project, from topic to topic. In these circumstances, it is not possible to arrive at an unbalanced formulation of the F-measure that will be appropriate for all circumstances; the formulation of the F-measure that will be best suited to the widest variety of cases is the balanced one (also denoted F_1). If F_1 is accompanied by recall and precision numbers, one will have both a single index of overall effectiveness (in F_1) and a view into the aspects of performance that contribute to that index (in recall and precision). As our single index of effectiveness in performing the Interactive Task, we will use F_1.

We may find it helpful to supplement our primary metrics with additional gauges of effectiveness; our focus, however, in measuring participants' performance, will be on recall, precision and the balanced F-measure.

2.4.2. Sampling

Given the size of the document population that is the domain for the Interactive Task (nearly 7 million documents), we will not be able to obtain exact values for the recall, precision, and F-measure that each participant has achieved on each of the target topics; we will, however, be able to obtain sample-based estimates of those values. A number of sampling designs might be used to obtain these estimates; in the following, we sketch a sampling design that we believe will efficiently meet the goal of obtaining estimates of the metrics in which we are interested.

The sampling design we propose is fairly straightforward, its salient features being results-based stratification and disproportionate representation of strata. The key characteristics are as follows.

- A distinct sample is drawn for each set of Participating Teams who, for purposes clarifying a given topic, shared the same Topic Authority.
- For each sample, the document population is partitioned into strata, with strata being defined by the cross-classification of results submitted by each of the teams whose

performance is to be measured via the sample. For example, in the case of three teams sharing one Topic Authority, there will be eight strata, one for each of the possible combinations of binary assessments (R/R/R, R/R/NR, R/NR/R, R/NR/NR, and so on).[12]

- In constructing the sample, strata are represented disproportionately to their full-population proportions, with preference being given to strata defined for positive assessments. This is done in order to ensure that, even for a low-frequency topic, each possible result combination is sufficiently covered and that even an outlier participant will have a sufficient number of its mismatches with the others assessed.

Such an approach will allow us to obtain, using reasonably-sized samples, reasonably precise estimates of the recall, precision, and balanced F-measure achieved by each Participating Team on each topic.

To illustrate the approach we propose, it will be helpful to walk through a hypothetical example.

Suppose, for example, we have three Participating Teams (Entries A, B, and C) who engage with the same Topic Authority for clarifying the scope and intent of a given topic; suppose, moreover, that the full-population yield of documents relevant to the target topic is relatively low, in the neighborhood of 2%. When we partition the population on the basis of the results submitted by the participants, the result might look something like the following.

Scenario 1 Three Entries (LH, HL, HH) Yield = 2%				
Stratum	**Entry A**	**Entry B**	**Entry C**	**N**
1	R	R	R	44,801
2	R	R	NR	11,490
3	R	NR	R	11,272
4	R	NR	NR	20,273
5	NR	R	R	45,257
6	NR	R	NR	122,270
7	NR	NR	R	38,796
8	NR	NR	NR	6,705,841
TOTAL				**7,000,000**

What the table shows is that, of the 7 million documents in the population, 44,801 have been assessed as relevant to the target topic by all three participants, 11,490 documents have been assessed as relevant by Entries A and B but not relevant by Entry C, 11,272 documents have been assessed as relevant by Entries A and C but not relevant by Entry B, and so on.

Suppose, then, that we draw a simple random sample (without replacement) of 500 documents from each of the strata defined for a positive assessment by at least one team (strata 1-7) and a simple random sample (without replacement) of 2,000 documents from the stratum containing cases in which all three participants have made a negative assessment (stratum 8); our sample then totals 5,500 documents and has been composed in such a way as to ensure that, for each participant, 2,000 positive assessments and 3,500 negative assessments will be reviewed. Suppose we review the 5,500-document sample for relevance to the target topic and, following the adjudication procedures outlined above, reach a final assessment of the relevance of each of the sample documents. The result might look something like the following.

[12] Note that this feature of the sampling design means that sample review for performance measurement cannot begin until after teams have submitted their results, introducing an additional constraint on the overall timeline for the task; see Appendix A below.

Scenario 1 Three Entries (LH, HL, HH) Yield = 2%							
Stratum	**Entry A**	**Entry B**	**Entry C**	**N**	**n**	**Rel**	**p**
1	R	R	R	44,801	500	499	99.8%
2	R	R	NR	11,490	500	487	97.4%
3	R	NR	R	11,272	500	497	99.4%
4	R	NR	NR	20,273	500	69	13.8%
5	NR	R	R	45,257	500	495	99.0%
6	NR	R	NR	122,270	500	46	9.2%
7	NR	NR	R	38,796	500	144	28.8%
8	NR	NR	NR	6,705,841	2000	2	0.1%
TOTAL				**7,000,000**	**5,500**	**2,239**	

What the table shows is that, of the 500 documents sampled from Stratum 1 (the set of documents all three entries assessed as relevant), 499 (or 99.8%) were, via the review and adjudication process, found to be actually relevant; of the 500 documents sampled from the second stratum, 487 (97.4%) were found to be actually relevant, and so on down to Stratum 8 in which just 2 out of 2,000 documents were found to be actually relevant.

From these results, estimation of the full-population yield of documents relevant to the target topic is, using a conventional stratified estimator, straightforward; the results are as follows.

Full-Population Estimates			
Summary	**Est**	**95% Low**	**95% High**
t	**143,837**	133,897	153,777
p	**2.1%**	1.9%	2.2%

An estimate of the recall achieved by a given participant can then be obtained by taking the estimate of the number of actually relevant documents in just the strata defined for that participant's positive assessments (e.g., for Entry A, Strata 1-4) out of the full-population estimate of actually relevant documents. An estimate of the precision achieved by a given participant is obtained by taking the estimate of the number of actually relevant documents in the strata defined for the participant's positive assessments out of all the documents in those same strata. An estimate of F_1 is obtained by taking the harmonic mean of the recall and precision estimates. The results for our hypothetical scenario are as follows.

Entry-Specific Performance									
	Recall			**Precision**			**F1**		
Entry	**Est**	**95% Low**	**95% High**	**Est**	**95% Low**	**95% High**	**Est**	**95% Low**	**95% High**
Entry A	**48.6%**	45.2%	52.0%	**79.6%**	78.8%	80.3%	**60.3%**	57.7%	63.0%
Entry B	**77.8%**	72.0%	83.6%	**50.0%**	48.6%	51.4%	**60.9%**	58.8%	63.0%
Entry C	**77.8%**	72.3%	83.3%	**79.9%**	78.7%	81.0%	**78.8%**	75.9%	81.7%

As can be seen, the approach sketched here allows us to obtain informative measures of effectiveness at a reasonable cost in terms of sample review and adjudication.

In order to supplement this illustration, it may be helpful also to consider results that might be obtained under a four-participant scenario; such a scenario, utilizing an 8,000-document sample, follows.

Scenario 2 Four Entries (LL, LH, HL, HH) Yield = 2%								
Stratum	**Entry A**	**Entry B**	**Entry C**	**Entry D**	**N**	**n**	**Rel**	**p**
1	R	R	R	R	22,400	400	399	99.8%
2	R	R	R	NR	5,603	400	398	99.5%
3	R	R	NR	R	5,601	400	398	99.5%
4	R	R	NR	NR	1,578	400	355	88.8%
5	R	NR	R	R	22,405	400	398	99.5%
6	R	NR	R	NR	6,733	400	333	83.3%
7	R	NR	NR	R	5,881	400	381	95.3%
8	R	NR	NR	NR	69,771	400	8	2.0%
9	NR	R	R	R	22,401	400	398	99.5%
10	NR	R	R	NR	5,887	400	380	95.0%
11	NR	R	NR	R	5,671	400	395	98.8%
12	NR	R	NR	NR	18,695	400	30	7.5%
13	NR	NR	R	R	22,853	400	392	98.0%
14	NR	NR	R	NR	115,537	400	19	4.8%
15	NR	NR	NR	R	32,914	400	68	17.0%
16	NR	NR	NR	NR	6,636,070	2000	2	0.1%
TOTAL					**7,000,000**	**8,000**	**4,354**	

Full-Population Estimates			
Summary	**Est**	**95% Low**	**95% High**
t	**144,787**	135,135	154,439
p	**2.1%**	1.9%	2.2%

Entry-Specific Performance									
	Recall			**Precision**			**F1**		
Entry	**Est**	**95% Low**	**95% High**	**Est**	**95% Low**	**95% High**	**Est**	**95% Low**	**95% High**
Entry A	**48.2%**	44.9%	51.5%	**49.9%**	49.1%	50.6%	**49.0%**	47.3%	50.8%
Entry B	**48.2%**	45.0%	51.4%	**79.4%**	78.8%	80.0%	**60.0%**	57.5%	62.5%
Entry C	**77.1%**	71.7%	82.5%	**49.9%**	48.8%	51.0%	**60.5%**	58.7%	62.4%
Entry D	**77.1%**	71.9%	82.4%	**79.7%**	78.8%	80.6%	**78.4%**	75.7%	81.1%

Conclusion: benefits and limitations

In this paper, we have outlined a set of modifications to the design of the Interactive Task, modifications that, in large part, are motivated by an interest in making the 2008 running of the task a more realistic representation of real-world e-discovery conditions and objectives. It may be helpful, in conclusion, to review what we see as the chief benefits, as well as possible limitations, of the proposed modifications to this year's task design.

First, the benefits.

- **More realistic.** As redesigned, the task includes elements that will allow it to serve as a more accurate end-to-end representation of the goals and conditions of e-discovery in the real world. Key elements in this regard are:
  - Participants are assigned the task of replicating the relevance assessments of a single authority, modeling the objective, in a real-world matter, of replicating the judgment of the legal architect overseeing document discovery.
  - Participants are given the opportunity to engage with experts in order to clarify the scope and intent of a topic, modeling opportunities in an actual e-discovery project.
  - Participants are required to make binary relevance assessments across the entire document population (or, in some cases, across a sample of it), mirroring real-world e-discovery requirements.
- **Informative metrics.** The task includes provision for efficiently obtaining informative measures of the performance of participating teams.
- **Greater participation by the legal and e-discovery communities.** It is expected that the more realistic design of the task, together with some specific task requirements (e.g., for

the role of the Topic Authority) will encourage greater participation in the task by the legal and e-discovery communities.

- **More productive collaboration among the IR, legal, and e-discovery communities.** It is expected that, as members of the legal and e-discovery communities increase their participation in the track, a track in which, until now, the primary participant has been the IR community, all three communities will learn more about each others' needs and capabilities, thereby laying the groundwork for finding document-discovery tools and processes better suited to real-world needs.

Now, the limitations.

- **Direct comparison of all participants.** As designed, the task permits direct comparison of only participants who shared the same Topic Authority for purposes of topic clarification; it might be desirable to be able to make a direct comparison of all participants in the task (or who targeted the same topic). We might address this limitation in two ways.
  - While direct comparison of the performance of teams who do not share the same Topic Authority is not possible, it remains true that the objective of those teams is the same: to replicate the relevance judgments of their respective Topic Authorities. The metrics we obtain will tell us how close each team came to achieving its objective, even if we have to allow that the conditions in which each team operated (i.e., the Topic Authority) were not the same.
  - If in the future we chose more than one Topic Authority for a given topic, we might at that time also consider adding an additional layer of guidance (a "Super" Topic Authority) as a means of ensuring consistency across all Topic Authorities for the topic. We believe that this may be an option worth exploring, but for the 2008 running of the task we think the simpler design sketched above will be sufficiently informative.
- **Reusability of results.** A goal of TREC is to produce annotated collections that can serve as the basis for future research; one might ask whether the collections that result from this task, because the assessments are specific to a given Topic Authority's interpretation, will serve that purpose. To this question we have three answers.
  - The challenge posed by assessment inconsistency is not new to this task. One might argue, in fact, that explicitly tying assessments to the interpretation of a single Topic Authority brings greater transparency to the sources of inconsistency and, as a consequence, will put future research that would rely on those assessments on a sounder footing.
  - The sampling and measurement protocol will result in data that can be used for the future testing and measurement of retrieval processes targeting the task topics. The sampling and measurement protocol will result, for each topic, in an estimate of the yield of relevant documents in the full collection. This estimate can serve as the basis by which future researchers estimate the recall achieved by any newly-tested process. If one knows the number of positive assessments that a process has resulted in, and if one is able to obtain an estimate of the precision of those positive assessments, and, finally, if one has a valid estimate of the yield of actually relevant documents in the target collection, then one has all the elements one needs to arrive at an estimate of the recall achieved by the tested process.
  - We plan to include a provision for archiving, for each topic, the information provided to participants by each Topic Authority. This will allow future researchers to make a

better informed use of the assessments that have been recorded for a given topic (as defined by a given Topic Authority). Some specifics regarding the form and content of the topic archives remain to be decided.

- **Number of topics.** As designed, the resource requirements of the task preclude running a large number of topics. While a larger number of topics would certainly be desirable, we believe that a more realistic representation of the task of e-discovery, even if it requires the acceptance of a smaller number of topics, will serve the interested communities better.[13] We might add that, over time, once any flaws in the design have been identified and resolved, we may be able to increase the number of topics.
- **Cost.** Making a more realistic task does, as should be clear from our discussion, require additional resources; we believe, however, that, with the appropriate outreach to the interested communities, we should be able to recruit the resources necessary. It is, after all, in the interest of all parties (the IR community, the legal community, the e-discovery community) to participate in a project that will advance our understanding of the processes and technologies capable of meeting the increasingly steep challenges of large-scale document discovery and that can also serve as a forum for on-going communication and collaboration.

We believe that the benefits to be expected from the revised design for the Interactive Task considerably outweigh the limitations. We look forward to a productive discussion with the Legal Track community on the new design and, after making any improvements called for by that discussion, to an exciting and informative initial running of the revised design in 2008.

[13] Note that the 2007 Interactive Task also targeted a limited set of topics.

Appendix A: Specific Guidelines

In the preceding, we have endeavored to provide a reasonably full description of the task and of the reasoning behind its various features. In this appendix, we review specific requirements for registering and executing the task.

A.1. Prerequisites and Registration

We welcome new and returning participants; prior participation in TREC or in the Legal Track is not a prerequisite. By way of prerequisites, all that we ask is that those considering participation review the preceding task description in order to understand the task requirements. If there are any questions regarding the task design, please do not hesitate to contact Bruce Hedin.

Registration for the Interactive Task is a matter of three steps.

1. Register your intent to participate in TREC 2008; do so by following the instructions at http://lists.si.umich.edu/pipermail/clair/2007-December/000139.html.
2. Join the Legal Track mailing list, if you have not already done so. Contact oard (at) umd.edu to be added to the list.
3. Send an email to the Track Coordinators indicating that you will be participating in the Interactive Task and indicating the number of topics for which you will be submitting results and your preferred topics (see below for more on topics). Email addresses for the Track Coordinators are as follows:

Jason R. Baron	jason.baron (at) nara.gov;
Bruce Hedin	bhedin (at) h5.com;
Douglas W. Oard	oard (at) umd.edu;
Stephen Tomlinson	stephent (at) magma.ca.

In addition to the Interactive Task, the 2008 Legal Track also includes Ad Hoc and Relevance Feedback tasks; for additional information on those tasks, see the guidelines posted on the Legal Track website (http://trec-legal.umiacs.umd.edu/).

A.2. Document Collection

The document collection used for the Interactive Task will be the same as that used in the first two years of the Legal Track.

> The set of documents for the track will be the IIT Complex Document Information Processing test collection. This collection consists of roughly 7 million documents (approximately 57 GB of metadata and OCR text uncompressed, 23 GB compressed) 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.
>
> The metadata and OCR can be obtained by FTP at no charge. For teams unable to transfer this quantity of data by FTP, the collection will also be available by mail as a set of DVD's from NIST.

In order to download the collection, please go to the IIT CDIP Test Collection web page (http://www.ir.iit.edu/projects/CDIP.html); fill out the form at the bottom of the page and you will be contacted with the ftp information.

A.3. Topics

Three topics have been selected as the retrieval targets for the Interactive Task (the resource-intensive nature of the task prevents, at least for 2008, a greater number of topics). A Participating Team is free to take on one, two, or all three topics, as it chooses.

All three topics are associated with the same mock complaint (an updated version of a complaint used in the 2006 Legal Track). Two of the topics (102, 103) are entirely new for 2008; one (104) was used in a prior years (in the 2006 Ad Hoc task and in the 2007 Relevance Feedback Task), but users should be aware that modifications to the complaint and the addition of the Topic Authority's guidance could well reorient the topic, making it essentially a "new" topic. To each topic at least one Topic Authority will be assigned; as noted above, the Topic Authority will serve as a resource for teams seeking clarification of the scope of a topic and as the final arbiter of relevance in the review and adjudication phase of the task. Both the complaint and the topics can be found at http://trec-legal.umiacs.umd.edu/.

In order to be able to link teams to their topics and to their Topic Authorities, we ask that each team, when registering, indicate the number of topics for which it wishes to submit results (due to resource constraints, we can accept for evaluation only one submission per topic per team) and its preferred topics. The coordinators will do their best to give all teams their preferred topics; because, however, of the need to balance the load among Topic Authorities, we may find it necessary to assign a team a topic other than its first choice. Those who indicate their preferences early will be more likely to get their first choices.

A.4. Submission of Results

For the Interactive Task, because, as noted above, the assessment is binary (relevant, not relevant), and because any unassessed document will be counted as a negative assessment (not relevant), a complete list of a Participating Team's positive assessments will suffice to evaluate a team's submission for a topic.

What we ask, then, by way of results submission, is that each team, for each topic for which it is submitting results for evaluation, submit a separate file. The filename should indicate both the team submitting the results and the topic for which the results are being submitted (e.g., "ParticipatingTeam_Topic102"). The content of the file should be a simple list of the TID values (unique document identifiers; e.g., "aaa00a00", "aaa01aa0", etc.) of all the documents a team has found to be relevant to the target topic. Any document on the list will be considered to have been deemed relevant by the team; any document not on the list will be considered to have been deemed not relevant by the team.

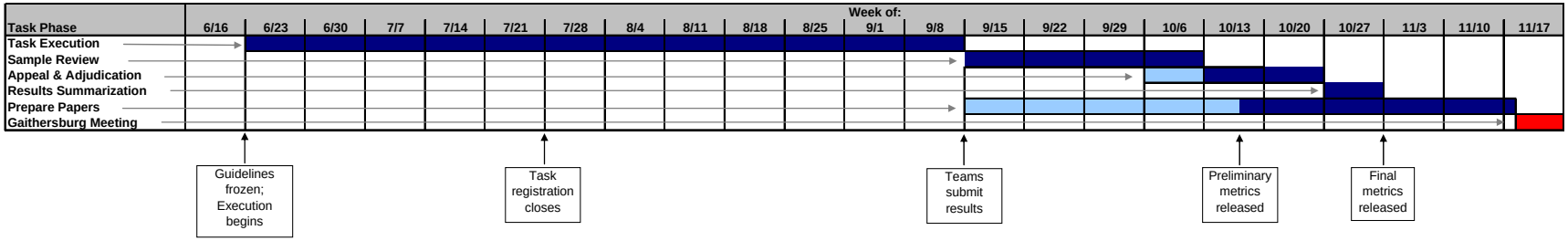
A.5. Schedule

Key dates for the Interactive Task are as follows.

Guidelines Frozen; Topics Released	06/23/08
Window for Task Execution	06/23/08 – 09/12/08
Window for Task Registration Closes	07/25/08
Deadline for Teams to Submit Results	09/12/08
Evaluation Sample Review	09/15/08 – 10/10/08
Appeal & Adjudication	10/06/08 – 10/24/08
Preliminary Metrics Released	10/15/08
Final Metrics Released	10/31/08
TREC 2008, Gaithersburg, MD	11/18/08 – 11/21/08

The diagram on the following page summarizes the task timeline.

TREC-2008 Interactive Task Timeline



Key Dates

Teams Execute Task	06/23/08 - 09/12/08
Teams Deliver Results	09/12/08
Sample Review	09/15/08 - 10/10/08
Appeal & Adjudication	10/06/08 - 10/24/08
Preliminary Metrics Released	10/15/08
Final Metrics Released	10/31/08
Gaithersburg Meetings	11/18/08 - 11/21/08

Legend

- Partial Engagement
- Full Engagement
- Gaithersburg Meetings

A.6. Reporting of Results

As in past years, Track Coordinators will prepare a report summarizing findings from the Legal Track. A preliminary draft of the report will be prepared for the November 2008 meetings in Gaithersburg; a final version of the report will be made available in spring 2009.

Individual teams are also expected to prepare reports of their findings. Preliminary drafts are submitted in late October for use only by participants in the November TREC confeerence, with revised versions being posted on the TREC Web site in January 2009. Teams should note that the deadline for the submission of reports to be included on the conference CD is typically three weeks prior to the start of the conference. Given the task timeline, reports prepared to meet that deadline will most likely be based primarily on preliminary metrics; final (fully adjudicated) metrics will be made available by the November TREC conference and can be the basis for final reports on findings.

A.7. Additional Information

The Legal Track website at http://trec-legal.umiacs.umd.edu/ contains links to resources and background information. The track mailing list archives can be reached through a link from that page. For additional questions, please contact one of the track coordinators:

Jason R. Baron	jason.baron (at) nara.gov;
Bruce Hedin	bhedin (at) h5.com;
Douglas W. Oard	oard (at) umd.edu;
Stephen Tomlinson	stephent (at) magma.ca.