TREC–2010 Legal Track
Interactive Task Guidelines

Abstract

In 2010, the TREC Legal Track will again be featuring an interactive task (along with a learning task, for more on which see the Learning Task Guidelines [4]). This document contains guidelines for this year's interactive task, with particular focus on aspects of task design that are new in 2010; the document also covers specific steps that those interested in participating should take to begin the task.

1 Introduction

In 2008, the TREC Legal Track, seeking to develop an exercise that modeled more completely and accurately the task of reviewing documents for responsiveness to a request for production in civil litigation, introduced a redesigned interactive task (see the 2008 Interactive Task Guidelines [1]). The task saw participation from four teams (two academic and two commercial) and produced interesting results, both with regard to the effectiveness of the approaches evaluated and with regard to the evaluation design itself (see Overview of the TREC 2008 Legal Track [6]).

In 2009, the Legal Track again featured the interactive task, this time with a new test collection and with a few minor modifications to the task design (see the 2009 Interactive Task Guidelines [2]). The 2009 exercise saw participation from eleven teams (three academic and eight commercial) and again produced interesting results, both from the perspective of retrieval methodology and from the perspective of evaluation design (see Overview of the TREC 2009 Legal Track [5]).

In 2010, the Legal Track will again be featuring the interactive task. This document describes key features of the design of the task. We begin with a brief review of the main elements of the interactive task (Section 2), then turn to a discussion of modifications (e.g., featuring a privilege topic) we are introducing for the 2010 exercise (Section 3), and conclude with a summary of the specific parameters that will define the 2010 interactive task (Section 4).

2 Overview of Task

The Legal Track's interactive task models the conditions and objectives of a review for responsiveness; that is to say, the task models the conditions and objectives of a search for documents that are responsive to a request for production that has been served during the discovery phase of a civil lawsuit. A full discussion of the circumstance modeled and of the general design of the exercise can be found in the 2008 task guidelines [1]. For purposes of the current overview, we briefly summarize the key features of the task.

- **Complaint and Topics.** Context for the interactive task is provided by a mock complaint that sets forth the legal and factual basis for the hypothetical lawsuit that motivates the discovery requests at the heart of the exercise. Associated with the complaint are document requests that specify the categories of documents which must be located and produced. For purposes of the interactive task, each of these document requests serves as a separate topic. The goal of a team participating in a given topic is to retrieve all, and only, documents relevant to that topic (as defined by the "Topic Authority;" see below). A team may choose to participate in as many, or as few, topics as they wish;

track coordinators reserve the right, however, to limit the number of teams that participate in a given topic in order to avoid placing an undue burden on the Topic Authority assigned to the topic.

- **The Topic Authority.** A key role in the task is played by the "Topic Authority." The Topic Authority plays the role of a senior attorney who is charged with overseeing a client's response to a request for production and who, in that capacity, must certify to the court that their client's response to the request is complete and correct (commensurate with a reasonable and good-faith effort). In keeping with that role, it is the Topic Authority who, taking into account considerations of genuine subject-matter relevance as well as pragmatic considerations of legal strategy and tactics, holds ultimate responsibility for deciding what is and is not relevant to a target topic (or, in real-world terms, what is and is not responsive to a document request). The Topic Authority's role, then, is to be the source for the authoritative conception of responsiveness that each participating team, in the role of a hired cohort of manual reviewers or of a vendor of document-retrieval services, will be asked to replicate across the full document collection. Each topic has a single Topic Authority, and each Topic Authority has responsibility for a single topic.

- **Interaction with the Topic Authority.** If it is the Topic Authority who defines the target (i.e., who determines what should and should not be considered relevant to a topic), it is essential that provision be made for teams to be able to interact with the Topic Authority in order to gain a better understanding of the Topic Authority's conception of relevance. In the interactive task, this provision takes the following form. Each team can ask, for each topic for which it plans to submit results, for up to 10 hours of a Topic Authority's time for purposes of clarifying a topic. A team can call upon a Topic Authority at any point in the exercise, from the kickoff of the task to the deadline for the submission of results; in the final weeks of the exercise, however, the amount of time a team may ask of a Topic Authority may be restricted (in order to avoid an undue burden on the Topic Authority in the final weeks of the exercise). How a team makes use of the Topic Authority's time is largely unrestricted: a team can ask the Topic Authority to pass judgment on exemplar documents; a team can submit questions to the Topic Authority by email; a team can arrange for conference calls to discuss aspects of the topic. Teams are not permitted, however, to engage in in-person interaction with Topic Authorities (on TREC-related matters) in order to avoid giving any possible unfair advantage to teams with the resources for travel. A further constraint that is placed on communication between the teams and their designated Topic Authorities is introduced in order to minimize the sharing of information developed by one team with another; while the Topic Authorities are instructed to be free in sharing the information they have about their topics, they are asked to avoid volunteering to one team specific information that was developed only in the course of interaction with another team.

- **Participant submissions.** Each team's final deliverable is 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 of arriving at their result sets, but the final deliverable is a single binary classification (relevant/not relevant) of the full population of documents.

- **Effectiveness Metrics.** Given the nature of the submissions (sets of documents identified as relevant to a topic), we look to set-based metrics to gauge effectiveness. In the interactive task, the metrics used are Recall, Precision, and, as a summary measure of effectiveness, F_1.

- **Sampling and Estimation.** As in previous years, in order to obtain estimates of effectiveness scores, we use stratified sampling and a two-stage sample assessment protocol. Further specifics are as follows.

  - **Sampling.** The sets of documents submitted by the participants in a topic allow for a straightforward submission-based stratification of the document collection: one stratum will contain the documents all participants submitted as relevant, another stratum will contain the documents no participant submitted as relevant, and other strata will be defined for each of the other possible submission combinations. If, for example, there are 5 teams participating in a topic, the collection will be partitioned into $2^5 = 32$ strata. In creating samples, strata will be represented largely

in keeping with their full-population proportions. In order to ensure that a sufficient number of documents are drawn from all strata, however, some small strata may be over-represented, and some large strata under-represented, relative to their full-population proportions. Selection within a stratum is simple random selection without replacement.

- **First-Pass Assessment.** As in previous years, the contents of each sample will be randomly assigned to "bins" of approximately 500 documents and these bins will be distributed to teams of manual assessors. Each assessor, equipped with detailed assessment guidelines and provided with access to the Topic Authority, will assess each document in his or her bin for relevance to his or her assigned topic. For the 2010 exercise, in order to improve the accuracy of the first-pass review, and thereby reduce the burden placed on the appeals phase, we plan (i) to introduce a number of enhancements to the training and support given the first-pass assessors and (ii) to introduce some amount of overlap in the documents assigned to different assessors.
- **Appeal and Adjudication.** No matter how rigorous the quality control regimen of the first-pass assessment, it is to be expected that some errors will remain in the sets of assessments that are the output of the first-pass review. As a corrective measure, the interactive task features an additional appeal/adjudication phase, whereby teams are given the opportunity to review the results of the first-pass assessment and appeal, to the Topic Authority, any assessments they believe are incorrect. The Topic Authority then renders a final judgment on all appealed assessments. In 2010, an appeals/adjudication phase will again be a a key element of the evaluation protocol. We will, however, be introducing some modifications to the appeals procedures in the interest of improving the efficiency and effectiveness of the adjudication phase (see below).

- **Reporting.** The final step in the interactive task, as in other TREC tasks, is reporting on and discussing the results for the task. The initial discussion takes place in two venues: first, at the TREC annual meeting in November and, second, in papers submitted for inclusion in the TREC 2010 proceedings. Of course, the discussion does not end there, but continues on in other conferences and publications to which the Legal Track's participants contribute.

3 New to the 2010 Task

The defining features of the interactive task will remain largely unchanged in 2010. We are, however, introducing some modifications designed to make the task more efficient and meaningful to the communities that participate in the Legal Track. The modifications are the following.

1. **Number of Topics & Number of Teams per Topic.** In 2009, we featured seven topics and, in order to avoid overburdening the Topic Authority, set the maximum number of teams that could participate in each topic at four. Our experience from the 2009 exercise indicates (a) that we can increase the maximum number of participants for each topic, without unduly overburdening the Topic Authority, and (b) that we can improve the efficiency of administering the exercise by limiting the number of topics.

   For the 2010 exercise, therefore, we will limit the number of topics to three responsiveness topics and one privilege topic (see below) and will allow a greater number of teams to participate in each topic. We believe that this will provide richer data on each topic that is featured and aid in the efficiency with which the task is run.

2. **Privilege Topic.** In our discussions with 2009 participants and 2009 Topic Authorities, we found that there was considerable interest in including a privilege topic (i.e., a topic that covered documents that are subject to a claim of attorney-client privilege, work-product, or other any other applicable privilege) in the interactive task. We therefore will do so in 2010.

   The rules for the privilege topic will be the same as the rules for the responsiveness topics: there will be a single Topic Authority designated for the topic, teams will be permitted to interact with the

Topic Authority, and so on. For purposes of this exercise, the scope of the privilege topic will not be restricted to what is also responsive to the other topics featured in the task; rather, any document in the collection that meets the Topic Authority's definition of privilege or work product, regardless of whether or not it is relevant to a responsiveness topic, will count as privileged. For the precise formulation of the privilege topic, see the 2010 mock complaint and associated topics.

3. **Provision for Entirely Manual Submission.** Those looking at the results of the Legal Track have often asked whether or not we have had participation from a team that adopted a traditional linear manual review as its approach to identifying responsive documents. The answer, at least until this year, has been No, and one reason for this is likely the fact that to complete a manual review of the full test collection, in the time allowed to prepare a submission, would require the commitment of more resources than any potential participant was prepared to commit.

   We would, however, like to encourage participation from such teams. Therefore, in order to lower at least one of the barriers to participation, we will, for the 2010 exercise, permit any entirely manual approach to review, and be evaluated on, a subset of the full test collection. The subset will consist of 50,000 email messages randomly selected from the full test collection; the subset will be defined by the coordinators prior to the start of the exercise. In other regards, a manual team will participate on the same terms as the other teams in the exercise (same topics, same Topic Authorities, same timetable, and so on). For evaluation purposes, should a manual team participate in a topic, the collection will be stratified and sampled so as to allow the estimation of scores both on the subset (for the manual team) and on the full collection (for the other teams).

   Note that the list of messages in the manual subset will be disclosed only to those participants who will be preparing entirely manual submissions; we do this in order to avoid the possibility that non-manual participants, being aware of the contents of the manual subset, might focus more of their efforts on that part of the collection, thereby biasing the results we obtain. We therefore ask that any participant wishing to prepare an entirely manual submission notify us of their intentions to do so; we will then release the list of messages in the subset to them.

4. **Reporting on Resources Expended.** Given that retrieval approaches can vary with respect to the amount of resources required and given that the 2008 and 2009 interactive tasks found that the retrieval approaches evaluated varied with respect to effectiveness, a natural follow-up question is: to what extent do the observed differences in effectiveness correlate with differences in the amount of resources required?

   In the 2009 exercise, we took a step toward answering that question by asking that participants *optionally* report the amount of person-hours expended in preparing their submissions. We found, however, that, with the reporting optional, only two of the eleven participants chose to provide data on hours expended. In 2010, we will make that reporting *mandatory*. More specifically, we will require that participants report, upon submission of their result sets, the following information:

   - the approximate number of person-hours expended in preparing their submission;
   - the number of documents that were manually reviewed in preparing their submission.

   We recognize, of course, that input costs cannot be derived simply by counting up the number of hours expended or the number of documents reviewed. We recognize, further, that the price of a product or service is a function of a number of factors independent of input costs and that, therefore, data on the resources required by a particular approach are not necessarily indicators of how that approach will be priced in the marketplace. The data we collect, therefore, are simply a gauge of how resource-intensive a particular approach is and not of the price that the market will attach to a given approach.

5. **Substratifying the "All-N" Stratum.** Utilizing, as we do, a submission-based stratification scheme, we typically find that a very high proportion of the collection (often 0.9 or greater) falls in the "All-N" stratum, the stratum defined for documents no participating team has deemed relevant. Sampling

effectively from this stratum is difficult, as the stratum is large and the frequency of relevant items in it low (and even more so when, as is often the case, the frequency of relevant items is low even in the unstratified collection). While difficult, it is also important, however, in that we want some gauge of the number of relevant documents missed by all participants.

A partial remedy would be to partition the All-N stratum into two substrata: one containing documents with a greater likelihood of being found relevant and the other containing documents less likely to be found relevant. Sampling from the former substratum should be more effective than sampling from the un-partitioned All-N stratum. This device is only a partial remedy, of course, in that the sampling challenges remain, probably in a still more acute form, in the substratum holding the documents less likely to be found relevant. Nevertheless, insofar as the device allows us to gauge more effectively the number of relevant documents in a salient part of the All-N stratum, it will be a step forward.

In order to take this step, we will be making the following provision for the 2010 interactive task. Any team not participating in a given topic, but registered for at least one of the tasks in the Legal Track, is invited to submit a list of documents potentially relevant to the topic. The purpose of such lists will be to compile a pool of potentially relevant documents, the intersection of which with the All-N stratum will define a substratum of documents that, though not identified as relevant by any of the active participants in the topic, have some characteristics that caused at least one Legal Track participant to put them in the potentially relevant category. A team can submit just one such list for any topic in which it is not a participant, and the deadline for submitting the list is the same as it is for evaluation submissions.

We expect that, by utilizing input from non-participant teams in this way, we will be able to obtain a clearer picture of what we know and what we do not know about the All-N stratum.

6. **Modifications to Appeals Procedures.** As noted above, a key component of the interactive task, is an appeals process whereby teams have the opportunity to ask the Topic Authority to adjudicate any first-pass assessments they believe are incorrect (i.e., inconsistent with the Topic Authority's conception of relevance). While the appeals mechanism, as implemented in the 2008 and 2009 exercises, has been reasonably effective at correcting first-pass errors (see, for example, the pre- to post-adjudication changes in scores in the 2009 task [5]), it suffers from the drawback that it is (a) time consuming and (b) dependent on the extent to which participants in a topic choose to utilize the appeals mechanism.

   For the 2010 interactive task, in the interest of improving the efficiency and effectiveness of the appeals process, we plan to introduce some modifications to the appeal/adjudication protocol. We are still working out the specifics of the modifications; we can say, however, the proposed changes will include (in some shape or form) both of the following:

   - having samples of appealed assessments adjudicated (in order to estimate the effects of appeals with greater efficiency);
   - having some assessments adjudicated by the Topic Authority, regardless of whether or not they were appealed.

   Once our proposed modifications are finalized, we will circulate for further discussion.

7. **Unit of Assessment.** In 2010, as in 2009, we will be using a collection of Enron emails as the test collection for the interactive task (see below for more on this year's test collection). In evaluating the effectiveness of approaches to assessing the relevance of email messages, one must decide whether one wants to assess effectiveness at the *message* level (i.e.,treat the parent email together with all of its attachments as the unit of assessment) or to assess effectiveness at the *document* level (i.e., treat each of the components of an email message (the parent email and each child attachment) as a distinct unit of assessment.

   In 2009, after much discussion (in which some participants argued in favor of message-level assessment and others in favor of document-level assessment), we opted for an all-of-the-above approach that asked

participants to submit their results at the document level (in order to enable document-level analysis) from which we would, by rule, derive message-level values (which would serve as the primary basis for evaluation). We found that the approach worked reasonably well, but not without some uncertainty, among both participants and first-pass assessors, as to what sorts of information could be used, in arriving at document-specific assessments, from other documents that belonged to the same message.

For the 2010 exercise, we will again gather document-level relevance assessments from which we can derive message-level values. In the interest of eliminating uncertainty, we provide the following clarifications.

- An email attachment should be deemed relevant if it has content that meets the Topic Authority's definition of relevance; in making this determination, contextual information contained in associated documents (parent email or sister attachments) should be taken into account.
- A parent email should be deemed relevant either if, in itself, it has content that meets the definition of relevance or if any of its attachments meet that definition; again, contextual information contained in all components of the email message should be taken into account.
- A message will count as relevant if at least one of its component documents (parent email or attachments) has been found relevant.
- For purposes of scoring, the primary level will be the message-level; document-level analysis will be supplementary.

8. **Test Collection.** As Gordon Cormack announced in his May 26 email to the Legal Track listserv, we will be using a new version of the Enron collection for the 2010 interactive task (and learning task); see Gordon's email for additional information on the new version [3]. See below (Section 4.3) for information on how to acquire the data set.

4 Further Specifics

We conclude the 2010 Interactive Guidelines with a review of specific requirements for registering and executing the task. This section covers, in particular, the following: (i) task prerequisites and registration; (ii) target topics; (iii) submission of results; (iv) task schedule; and (v) where to go for additional information.

4.1 Prerequisites & 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 carefully review these task guidelines (along with supporting materials, such as the 2008 and 2009 Guidelines [1] [2], the 2008 Legal Track Overview [6], and, when available, the 2009 Legal Track Overview [5]) in order to understand the task requirements and expectations. If there are any questions regarding the task design, please do not hesitate to contact Bruce Hedin (email address given below).

Registration for the interactive task is a matter of three steps.

1. Register your intent to participate in TREC 2010; do so by contacting the TREC Program Manager at trec(at)nist.gov. Once your registration information has been processed, you will be added to the mailing list for TREC participants and will receive a "Welcome" message containing other details about participation in TREC.

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 Bruce Hedin, indicating that you will be participating in the interactive task. In the email, also specify (i) the number of topics for which you would like to submit results (for more on target topics, see Section 4.2) and (ii) a list of your preferred topics (together with some alternatives) in order of preference. We will endeavor to give all teams their preferred topics, but, in the interest of load-balancing among the Topic Authorities, we may have to ask some teams to take a topic other than their first preference. Email addresses for the Track Coordinators are as follows:

Gordon V. Cormack	gvcormac (at) uwaterloo.ca;
Maura R. Grossman	MRGrossman (at) wlrk.com;
Bruce Hedin	bhedin (at) h5.com;
Douglas W. Oard	oard (at) umd.edu.

In addition to the interactive task, the 2010 Legal Track also features a learning task; for additional information on this task, see the Learning Task Guidelines [4].

4.2 Target Topics

The mock complaint (Complaint "K") and the associated topics (301 – 304) will be released upon the official start of the exercise.

4.3 Acquisition of the Data Set

As the test collection for the 2010 interactive task, we will be using the set of documents that results from message-level de-duping of the Enron emails. The collection can be downloaded, in both text and native formats, from the following webpage.

http://durum0.uwaterloo.ca/trec/legal10/

The webpage noted above also provides the official list of document IDs that are in-scope for the exercise (i.e., are valid for submission). Please let track coordinators know if you have any questions regarding the test collection.

4.4 Submission of Results

Specifics of the procedures and format for submitting results will be provided to participants in a separate document. For present purposes, please note the following.

- Submissions must be made using Legal Track Interactive Task Results form, which will be accessible from the TREC Information for Active Participants page.
- Each team should prepare and submit a single file containing results for all topics for which a team is submitting results for evaluation.
- The submission file should list only documents found relevant to each target topic. Any document included in the submission list for a given topic will be assumed to have been found relevant to that topic; any document not included in the submission list for a topic will be assumed to have been found not relevant to the topic.
- Submissions must be made by the deadline noted below (Section 4.5); submissions made prior to the deadline will be accepted and are welcome.
- A team should submit one set of results for each of its target topics; it will be possible to obtain scores for alternative result sets once the evaluation samples have been selected, reviewed, and adjudicated.

4.5 Task Schedule

The timeline for the 2010 interactive task follows.

Guidelines frozen; topics released	07/07/10
Window for task execution	07/07/10 – 09/06/10
Window for task registration closes	07/15/10
Deadline for submitting results	09/06/10
First-pass review of evaluation samples	09/07/10 – 09/24/10
Preliminary (pre-adjudication) scores released	09/29/10
Appeal & adjudication	09/30/10 – 10/29/10
Final (post-adjudication) scores released	11/05/10
Conference papers due	10/25/10 (estimate)
TREC 2010	11/16/10 – 11/19/10
Final papers due	February 2011 (estimate)

4.6 Additional Information

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

Gordon V. Cormack	gvcormac (at) uwaterloo.ca;
Maura R. Grossman	MRGrossman (at) wlrk.com;
Bruce Hedin	bhedin (at) h5.com;
Douglas W. Oard	oard (at) umd.edu.

References

[1] J. R. Baron, B. Hedin, D. W. Oard, and S. Tomlinson. Final Interactive Task Guidelines, 2008. Available at http://trec-legal.umiacs.umd.edu/2008InteractiveGuidelines.pdf.

[2] J. R. Baron, B. Hedin, D. W. Oard, and S. Tomlinson. Final Interactive Task Guidelines, 2009. Available at http://trec-legal.umiacs.umd.edu/.

[3] G. V. Cormack. [Trec-legal] New dataset (same source) for TREC 2010 Legal Track. Email, May 26, 2010. Available at http://lists.umiacs.umd.edu/pipermail/trec-legal/2010/000465.html.

[4] G. V. Cormack and M. R. Grossman. Learning Task Guidelines, 2010. Available at http://plg.uwaterloo.ca/ gvcormac/legal10.

[5] B. Hedin, S. Tomlinson, J. R. Baron, and D. W. Oard. Overview of the TREC 2009 Legal Track. In *The Eighteenth Text REtrieval Conference (TREC 2009) Proceedings*, Forthcoming (July, 2010).

[6] D. W. Oard, B. Hedin, S. Tomlinson, and J. R. Baron. Overview of the TREC 2008 Legal Track. In *The Seventeenth Text REtrieval Conference (TREC 2008) Proceedings*, November 2008.