Overview of the TREC 2010 Legal Track
Notebook Draft 2010.10.25

Gordon V. Cormack
Maura R. Grossman
Bruce Hedin
Douglas W. Oard

Abstract

The TREC 2010 Legal Track consisted of two distinct tasks: the *learning task*, in which participants were required to estimate the probability of relevance for each document, and the *interactive task*, in which participants were required to identify all relevant documents using a human-in-the-loop process. 2010 is the fifth year of the legal track, the third year of the interactive task within the legal track, and the first year of the learning task. We provide a brief overview of the two tasks, and results that were available as of October 25, 2010. More detailed results will be available to TREC participants during the conference at the web address `http://plg1.uwaterloo.ca/trec-assess` .

1 Introduction

We are concerned with the document selection and review component of the *e-discovery* process, for which the objective is to identify as nearly as practicable all documents from a collection that are responsive to a *discovery request* in civil litigation, while minimizing the number of unresponsive documents that are identified by the method. The learning and interactive tasks represent two different e-discovery scenarios:

- The learning task represents the scenario in which preliminary search and assessment has yielded a set of documents that are coded as relevant or not; this *seed set* is used as input to a process involving humans or technology to estimate the *probability* that each of the remaining documents in the collection is relevant.
- The interactive task represents the process of using humans and technology, in consultation with a *Topic Authority*, to identify as well as possible *all relevant documents* in the collection, while simultaneously minimizing the number of false positives.

2 Document Collection

The document collection for both tasks was derived from the EDRM Enron Dataset, version 2, prepared by ZL Technologies in consultation with the Legal Track coordinators, and hosted by EDRM. The EDRM dataset consists of 1.3 million email messages captured by the Federal Energy Review Commission (FERC) from Enron, in the course of its investigation of Enron's collapse. ZL acquired the dataset from Lougheed Systems (formerly Aspen Systems) who captured and maintain the dataset on behalf of FERC. The EDRM dataset is available in two formats: EDRM XML and PST. The EDRM XML version contains a text rendering of each email message and attachment, as well as the original native format. The PST version contains the same messages, in a Microsoft proprietary format used by many commercial tools.

Both versions of the dataset approach 100GB in size, presenting an obstacle to participants. Furthermore, there are a large number of duplicate email messages in the dataset, that were captured more than once

by Lougheed. For TREC 2010, a list of 470,000 distinct messages were identified as canonical; all other messages duplicate one of the canonical messages. These messages contain about 200,000 attachment files; together these 470,000 messages plus 200,000 attachments form the 670,000 documents of the TREC 2010 Legal Track collection, used for both the learning and interactive tasks. Text and native versions of these documents were made available to participants, along with a mapping from the EDRM XML and PST files to their canonical counterparts in the TREC collection.

3 Relevance Assessments

In order to measure the efficacy of TREC participant efforts in the two tasks, it is necessary to compare their results to a *gold standard* indicating whether or not each document in the collection is relevant to a particular discovery request. The learning task used eight distinct discovery requests, while the interactive task used four. Ideally, a gold standard would indicate the relevance of each document to each topic, a total of eight million judgments.

It is impractical to use human assessors to render these eight million assessments. Instead, a sample of documents was identified for each topic, and assessors were asked to code only the documents in the sample as relevant or not. For the learning task, 78,000 human assessments were used; for the interactive task, 50,000 human assessments were used.

The learning task assessments were rendered by individual volunteers, primarily, but not exclusively, law students. For each document and topic, three assessments were rendered, and the majority opinion was taken to be the gold standard. The interactive task assessments were assessed by professional review companies. Ten percent of the documents for each topic were assigned to more than one reviewer; the agreement among these redundant assessments was used to estimate and correct for assessor error. In both cases, individual assessors were asked to review documents in batches of 500, and reviewed one or more batches.

The assessors used a new Web-based platform developed by the coordinators to view scanned documents and to record their relevance judgments. To avoid problems with local rendering software on each assessor's workstation, the assessors made their judgments based on pdf-formatted versions documents, as opposed to the original native format documents.

Assessors were provided with orientation and detailed guidelines created by a Topic Authority. For the learning task, assessors were given 10 examples each of relevant and a non-relevant documents for their particular topic. The review platform included a "seek assistance" link which assessors were encouraged to use to request that the Topic Authority resolve any uncertainty that may have arisen.

In reviewing their bins, assessors were instructed to make a relevance judgment of relevant (R), not relevant (N), or broken (B) for every document in their bins. The latter code reflects the fact that a small percentage of documents from the EDRM dataset are malformed and therefore cannot be assessed.

4 Learning Task

The learning task models the use of automated or semi-automated methods to guide review strategy for a multi-stage document review effort, organized as follows:

1. **Preliminary search and assessment.** The responding party analyzes the production request. Using ad hoc methods the team identifies a *seed set* of potentially responsive documents, and assesses each as responsive or not.

2. **Learning by example.** A learning method is used to rank the documents in the collection from most to least likely to be responsive to the production request, and to estimate this likelihood for each document. The input to the learning method consists of the seed set, the assessments for the seed set, and the unranked collection; the output is a ranked list consisting of the document identifier and a probability of responsiveness for each document in the collection.

	topic								avg		
run	200	201	202	203	204	205	206	207	actual	est	acc
otL10rvlT	39.8	85.5	100.2	100.5	85.2	84.6	98.9	86.6	85.1	98.8	86.1
xrceLogA	77.9	93.8	99.4	92.2	73.8	71.8	74.9	92.0	84.4	88.8	95.0
xrceCalA	77.9	93.8	99.4	92.2	73.8	71.8	74.9	92.0	84.4	82.7	97.8
DUTHsdtA	90.6	85.0	97.8	103.4	98.0	82.2	88.0	18.7	82.9	90.1	92.0
DUTHsdeA	90.6	85.0	97.8	103.4	98.0	82.2	88.0	18.7	82.9	82.4	99.3
DUTHlrgA	90.6	85.0	97.8	103.4	98.0	82.2	88.0	18.7	82.9	96.3	86.1
otL10FT	97.9	94.9	98.8	105.6	68.8	84.9	88.3	21.7	82.6	96.6	85.5
tcd1	67.2	61.6	98.2	77.9	76.2	57.5	97.5	87.4	77.9	55.8	71.6
rmitindA	72.9	85.8	96.7	102.5	79.2	87.7	78.3	19.8	77.8	53.3	68.5
otL10bT	52.4	82.1	102.7	108.3	49.5	65.2	97.6	51.1	76.1	99.2	76.1
xrceNoRA	83.2	66.1	85.5	95.9	35.2	54.4	76.5	92.0	73.6	73.2	99.4
BckExtA	78.9	75.1	90.0	38.6	67.5	72.7	85.5	80.9	73.6	49.7	67.5
BckBigA	80.7	75.1	89.9	36.1	67.4	72.7	85.5	80.9	73.5	49.5	67.3
rmitmlsT	66.6	61.6	104.3	83.8	45.6	57.8	57.9	16.3	61.7	70.7	87.3
BckLitA	44.1	76.9	88.2	77.0	42.5	74.6	57.7	11.7	59.0	49.6	88.4
rmitmlfT	68.7	59.9	90.7	52.2	47.5	56.1	58.6	15.7	56.1	67.1	83.6
ITD	-	45.6	67.5	20.7	41.6	35.2	29.7	74.7	44.9	61.8	72.6
URSK70T	51.0	17.7	18.6	23.8	62.2	22.4	87.6	22.6	38.2	91.0	42.0
URSK35T	51.3	27.7	18.4	27.6	40.5	30.3	90.3	18.3	38.0	93.2	40.8
URSLSIT	51.0	19.2	21.3	23.8	50.6	22.4	87.6	22.5	37.3	83.5	44.7

Table 1: Recall at 30% cut.

	topic								
run	200	201	202	203	204	205	206	207	avg
xrceLogA	74.6	94.7	99.4	95.1	76.2	75.3	83.1	95.3	86.7
DUTHlrgA	83.3	85.7	95.6	98.4	85.4	87.3	90.0	61.7	85.9
DUTHsdeA	82.8	86.1	95.1	97.0	83.3	87.1	90.1	62.5	85.5
otL10FT	95.0	91.4	96.8	98.5	79.9	87.4	87.8	42.0	84.8
otL10rvlT	57.3	85.6	86.2	94.6	80.5	82.9	97.0	90.0	84.3
xrceCalA	73.1	72.1	99.0	88.8	78.8	73.6	83.1	94.3	82.9
BckBigA	77.2	87.1	92.9	67.8	79.2	76.7	82.7	88.2	81.5
rmitindA	81.6	87.2	96.3	98.7	86.1	88.4	85.8	26.8	81.4
BckExtA	75.4	87.1	92.9	66.1	79.3	76.7	82.7	88.3	81.1
xrceNoRA	78.0	84.0	90.6	87.3	56.6	69.1	76.4	89.8	79.0
otL10bT	70.8	83.8	90.9	97.5	62.7	68.7	97.8	57.0	78.6
tcd1	67.7	74.3	81.1	79.5	77.8	71.4	88.2	84.1	78.0
BckLitA	74.2	86.1	91.1	88.8	67.5	75.3	82.0	31.0	74.5
rmitmlsT	63.4	71.8	95.3	91.2	67.6	70.2	69.3	27.7	69.6
rmitmlfT	71.8	71.8	94.4	72.9	72.1	73.3	70.2	27.9	69.3
ITD	-	53.4	73.4	28.3	57.7	48.1	45.8	76.8	61.9
DUTHsdtA	56.3	57.5	68.1	71.8	52.3	56.6	57.3	54.8	59.3
URSK70T	61.2	48.6	50.6	49.1	69.1	51.1	89.2	47.0	58.2
URSLSIT	61.2	44.3	48.7	49.1	68.8	51.1	89.2	37.8	56.3
URSK35T	63.5	46.2	44.5	49.5	58.0	52.4	87.3	44.6	55.8

Table 2: AUC: Area under the receiver operating characteristic curve.

The two components of learning by example – ranking and estimation – may be accomplished by the same method or by different methods. Either may be automated or manual. For example, ranking may be done using an information retrieval method or by human review using a five-point scale. Estimation may be done in the course of ranking or, for example, by sampling and reviewing documents at representative ranks.

3. **Review process.** A review process may be conducted, with strategy guided by the ranked list. One possible strategy is to review documents in order, thus discovering as many responsive documents as possible for a given amount of effort. Another possible strategy is triage: to review only mid-ranked documents, deeming, without further review, the top-ranked ones to be responsive, and the bottom-ranked ones to be non-responsive.

   Review strategy may be guided not only by the order of the ranked list, as outlined above, but also by the estimated effectiveness of various alternatives. Consider the strategy of reviewing the top-ranked documents. Where should a *cut* be made so that documents above the cut are reviewed and documents below are not? For triage, where should the two necessary cuts be made?

Practically every review strategy decision boils down to the question,

Of this particular set of documents, how many are responsive and how many are not?

This question itself can be reduced to,

What is the probability of each document in the set being relevant?

Given an answer to the second question, the answer to the first is simply the sum of the probabilities. For this reason, participants in the learning track were required to provide an estimate of the probability of relevance for each document in the collection. Using these estimates the documents were ranked from most likely to least likely relevant. At each rank, the estimated number of relevant documents – the sum of the probabilities up to that rank – was computed, and used to estimate recall, precision and F_1.

The *accuracy* of the estimate is defined to be

$$accuracy = 100\% \times \frac{\min(estimate,\, true\, value)}{\max(estimate,\, true\, value)}\,. \tag{1}$$

Eight participating groups submitted 20 runs. Table 1 summarizes the results of those when the cutoff is set at 200,000 documents, or 30% of the collection. The second through ninth columns show the recall achieved when the top 30% of documents returned from each system were considered. The tenth column shows the average over all eight topics. The next column shows the estimate of this average, computed by summing the submitted probability estimates. The last column shows the accuracy of the estimate, using the method shown above.

Table 2 shows the Area Under the Receiver Operating Characteristic Curve (AUC) summary measure for each topic, as well as the average. AUC is a summary measure, based on signal detection theory, which captures the effectiveness of ranking over all cutoff levels.[1]

5 Interactive 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 2009 task guidelines. For purposes of the current overview, we briefly summarize the key features of the task.

[1] `www.ncbi.nlm.nih.gov/pmc/articles/PMC1065080/`

- **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). For the first time this year, one "topic" call for privilege-review rather than for production of documents based on their content.

- **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. 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. One 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.

- **Submissions.** Each team's final deliverable is a binary classification of the full population for relevance to each target topic in which it has chosen to participate.

- **Effectiveness Measures.** 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.** 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 contains the documents all participants submitted as relevant, another stratum contains 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 that submitted results for a topic, the collection will be partitioned into $2^5 = 32$ strata. In creating samples, strata are 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.

- **Binning.** For purposes of assessment, the contents of each sample is randomly assigned to "bins" of approximately 500 documents and these bins are then distributed to teams of manual assessors. About 10% of each bin was also assigned to another assessor in order to support assessor consistency studies and to provide a basis for non-uniform double sampling.
- **Appeal.** Once the first-pass assessors complete their work, we will provide each team with the full set of first-pass assessments for each topic for which they submitted results and invite them to appeal any assessments they believed had been made in error (i.e., out of keeping with the Topic Authority's conception of relevance). Unlike the procedure used in 2008, we will not guarantee that the Topic Authority will examine every appealed document; rather, appeals will be used to guide the double sampling process.
- **Double Sampling.** No matter how rigorous the quality control regimen of the first-pass assessment, some differences of opinion will surely remain between the assessments for the initial sample and the judgments that the Topic Authority would have made on those same documents . We therefore will then sample from this initial sample in a manner designed to allow us to estimate systematic differences between decisions on relevance made by each first-pass assessor and by the Topic Authority and we present that second sample to the Topic Authority for assessment [2].
- **Estimation.** Once all samples have been assessed, we are able to compute estimates both of the full-population yield of relevant documents for each topic and of each participant's effectiveness scores (recall, precision, F_1) for each topic. For further detail on the estimation procedures followed in the Interactive task, see the appendix to the Overview of the TREC 2008 Legal Track [1].

5.1 Topics

The Interactive task topics and the corresponding Topic Authorities were:

- **Topic 301.** All documents or communications that describe, discuss, refer to, report on, or relate to onshore or offshore oil and gas drilling or extraction activities, whether past, present or future, actual, anticipated, possible or potential, including, but not limited to, all business and other plans relating thereto, all anticipated revenues therefrom, and all risk calculations or risk management analyses in connection therewith.
  - **Topic Authority:** Mira Edelman (Hughes Hubbard & Reed).
- **Topic 302.** All documents or communications that describe, discuss, refer to, report on, or relate to actual, anticipated, possible or potential responses to oil and gas spills, blowouts or releases, or pipeline eruptions, whether past, present or future, including, but not limited to, any assessment, evaluation, remediation or repair activities, contingency plans and/or environmental disaster, recovery or clean-up efforts.
  - **Topic Authority:** John F. Curran (Stroz Friedberg).
- **Topic 303.** All documents or communications that describe, discuss, refer to, report on, or relate to activities, plans or efforts (whether past, present or future) aimed, intended or directed at lobbying public or other officials regarding any actual, pending, anticipated, possible or potential legislation, including but not limited to, activities aimed, intended or directed at influencing or affecting any actual, pending, anticipated, possible or potential rule, regulation, standard, policy, law or amendment thereto.

- **Topic Authority:** Howard J. C. Nicols (Squire, Sanders, & Dempsey).

- **Topic 304.** All documents or communications that describe, discuss, refer to, report on, or relate to any intentions, plans, efforts, or activities involving the alteration, destruction, retention, lack of retention, deletion, or shredding of documents or other evidence, whether in hard-copy or electronic form.

  - **Topic Authority:** Michael Roman Geske (Aphelion Legal Solutions).

Topic 304 was the new "privilege topic." Teams were invited to rank topics in preference order, and topics were then assigned to teams in a manner that respected those preferences to the degree possible.

5.1.1 Submissions

The Interactive task received submissions from twelve teams (seven commercial and five academic), who, collectively, submitted a total of 22 single-topic runs. The twelve teams that submitted results for evaluation are as follows (full name followed by two-letter team ID):

- A collaborative effort of Cleary Gottlieb Steen & Hamilton LLP and Backstop LLP (**CB**);
- Clearwell Systems (**CS**);
- Equivio (**EQ**);
- Integreon (**IN**);
- Indian Statistical Institute (**IS**);
- A collaborative effort of IT.com and Williams Mullen (**IT**);
- Los Alamos National Labs (**LA**);
- MailMeter (**MM**);
- U. South Florida (**SF**);
- U. Buffalo, State University of New York (**UB**);
- A collaborative effort of U. Melbourne and the Royal Melbourne Institute of Technology (**UM**); and
- U. Waterloo (**UW**).

It should be noted that Douglas Oard, a track coordinator who was on sabbatical at the University of Melbourne and RMIT during a part of this period, participated in that team's research.

Teams were invited to ask to participate in as many, or as few, topics as they chose. Given constraints on the number of teams for which a Topic Authority could take responsibility (typically, a maximum of eight teams), we indicated that we might not be able to give all teams all of their choices and asked teams to rank their topic selections in order of preference. Topics were assigned largely on a first-come-first-serve basis. For the 2010 task, it turned out that we were able to give all teams their preferred topics. Table 5.1.1 shows the number of runs submitted by each team for each topic; in the table, an empty cell represents no submissions for the given team-topic combination.

As can be seen from the table, in most cases, each team submitted, in accordance with the task guidelines, just one run for each topic it chose to be evaluated. In one case, however, a team asked for, and was given, permission to submit multiple runs for a single topic.

The Cleary-Backstop team (CB) wished, for Topic 303, to have two submissions evaluated, with each of the two submissions representing a different approach to taking into account "family" associations between documents. The Cleary-Backstop team wished, for Topic 304, to have four submissions evaluated, two

Team	Topics				Total Runs
	301	302	303	304	
CB			2	4	**6**
CS	1				**1**
EQ			1		**1**
IN		1		1	**2**
IS	1	1			**2**
IT	1		1		**2**
LA		1			**1**
MM		1			**1**
SF	1				**1**
UB			1		**1**
UM		1			**1**
UW	1	1	1		**3**
Total Runs	**5**	**6**	**6**	**5**	**22**

Table 3:

targeting just the conception of responsiveness (privilege) defined by the Topic Authority (and with each of the two representing a different approach to taking into account "family" associations between documents) and two targeting a broader notion of "potentially privileged" (and, again, with each of the two representing a different approach to taking into account "family" associations between documents).

5.1.2 Unit of Assessment

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. (Sometimes, in past discussions, the term *record* has been used as a synonym for *message*).

For the 2010 Interactive task we asked participants to submit their results only at the message level (by specifying the message if any of its components—the message itself or any of its attachments—is believed to be relevant). For teams that specified documents other than messages, we derived message-level values using the same process that was used in the 2008 Interactive task.

5.2 Results

The timing of the availability of this year's new document collection document collection and an unanticipated delay in finalizing Topic Authority assignments resulted in a late start for the Interactive task for the second year in a row. At the time of this writing (October 25, 2010), first-pass relevance judgments are not yet available for the four Interactive task topics. As first-pass relevance judgments become available we will distribute preliminary results (i.e., results not yet based on double sampling) to participants.

6 Conclusion

Evaluation of the TREC Legal Track tasks – learning and interactive – continues. For learning, it appears that automated and technology-assisted approaches can identify a small subset of the collection that contains the vast majority of relevant documents. For 2011, we propose to expand the role of relevance feedback in the

learning task, so that the ranking and probability estimates may be improved during the course of review. For 2010, the interactive task ran closer to schedule but still missed its deadlines, due to various logistical issues. At the time of writing, assessment was complete for topic 304 (privilege) and ongoing for the other three.

References

[1] Douglas W. Oard, Bruce Hedin, Stephen Tomlinson, and Jason R. Baron. Overview of the TREC 2008 Legal Track. In *The Seventeenth Text REtrieval Conference (TREC 2008)*, 2009.

[2] William Webber, Douglas W. Oard, Falk Scholer, and Bruce Hedin. Assessor error in stratified evaluation. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM)*, 2010. to appear.