Overview of the TREC 2011 Legal Track

Maura R. Grossman, *Wachtell, Lipton, Rosen & Katz*, New York, NY (USA)*
Gordon V. Cormack, *University of Waterloo*, Waterloo, ON (Canada)
Bruce Hedin, *H5*, San Francisco, CA (USA)
Douglas W. Oard, *University of Maryland*, College Park, MD (USA)†

Abstract

The TREC 2011 Legal Track consisted of a single task: the learning task, which captured elements of both the TREC 2010 learning and interactive tasks. Participants were required to rank the entire corpus of 685,592 documents by their estimate of the probability of responsiveness to each of three topics, and also to provide a quantitative estimate of that probability. Participants were permitted to request up to 1,000 responsiveness determinations from a Topic Authority for each topic. Participants elected either to use only these responsiveness determinations in preparing automatic submissions, or to augment these determinations with their own manual review in preparing technology-assisted submissions. We provide an overview of the task and a summary of the results. More detailed results are available in the Appendix to the TREC 2011 Proceedings.

1 Introduction

We are concerned with the identification of responsive documents as part 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 request for production in civil litigation, while minimizing the number of unresponsive documents that are identified.

The learning task models the scenario in which a senior attorney – the *Topic Authority* – is charged with interpreting the request for production, communicating that interpretation to a review team, and producing responsive documents to the requesting party. TREC participants play the role of the review team.

At the outset, the Topic Authority reviews the request and a sample of potentially responsive documents, and prepares a set of coding guidelines. The production request and the guidelines are provided to participants, and an initial kick-off call allows interested participants to ask the Topic Authority questions about his or her interpretation of the request for production.

Over the course of several weeks, each participant is entitled to request feedback from the Topic Authority on a number of documents from the collection. This feedback consists of a simple binary *responsiveness determination:* participants are informed whether the Topic Authority determines each document to be responsive or not. No other communication with the Topic Authority is permitted.

Teams from ten different organizations participated in the 2011 Legal Track; the names of the teams, as well as the prefix used to label each team's results, are shown in Table 1.[1]

2 Document Collection

The document collection used for the TREC 2011 Legal Track was identical to that used for TREC 2010. It was derived from the EDRM Enron Dataset, version 2, prepared by ZL Technologies in consultation with the 2010 Legal

*The views expressed herein are solely those of the author and should not be attributed to her firm or its clients.

†This work has been supported in part by the National Science Foundation under grant IIS-1065250. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

[1]Track coordinator Gordon V. Cormack did not participate in the University of Waterloo's TREC 2011 Legal Track effort.

Participating Organization	Run Prefix
Beijing University of Posts and Telecommunications	pri
Helioid	HEL
Indian Statistical Institute	ISI
OpenText	ot
Recommind	rec
TCDI	tcd
University of Melbourne	mlb
University of South Florida	USF
University of Waterloo	UW
Ursinus College	URS

Table 1: Organizations participating in the TREC 2011 Legal Track.

Track coordinators, and hosted by EDRM [1]. The EDRM dataset consists of 1.3 million email messages captured by the Federal Energy Regulatory Commission ("FERC") from Enron, in the course of its investigation of Enron [2]. ZL independently acquired the dataset from Lockheed 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 some participants. Furthermore, there are a large number of duplicate email messages in the dataset that were captured more than once by Lockheed. For TREC, a list of 455,449 distinct messages were identified as canonical; all other messages duplicate one of the canonical messages. These messages contain 230,143 attachment files; together, these messages plus attachments form the 685,592 documents of the TREC 2010/2011 Legal Track collection. 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 Responsiveness Assessments

In order to measure the efficacy of TREC participants' efforts, it is necessary to compare their results to a *gold standard* indicating whether or not each document in the collection is responsive to a particular discovery request. The learning task had three distinct *topics*, each representing a distinct request for production.

Ideally, a gold standard would indicate the responsiveness of each document to each topic. Because it would be impractical to use human assessors to render these two million assessments, a sample of documents was identified for each topic, and assessors were asked to code only the documents in the sample as responsive or not. Since errors in the gold standard can have substantial impact on evaluation, redundant independent assessments were made for the majority of the sampled documents, and disagreements were adjudicated by the Topic Authority.

A total of 16,999 documents – about 5,600 per topic – were selected and assessed to form the gold standard. The documents that were selected met one or more of the following four criteria:

1. All documents that were identified by the Track coordinators to be potentially responsive in the course of developing the topics before the start of the task;
2. All documents submitted by any team for responsiveness determination;
3. All documents ranked among the 100 most probably responsive by any submission;
4. A uniform random sample of the remaining documents.

11,612 documents (referred to as the *100 stratum*) were selected according to one or more of the first three criteria; 5,387 documents (referred to as the *1000 stratum*) were sampled according to the fourth. All documents in the 100

stratum were assessed, regardless of whether or not a responsiveness determination had been previously rendered by the Topic Authority. Each document in the 1000 stratum was given to two assessors; that is, each sampled document was assessed twice.

The learning task assessments were rendered by four professional review companies, who volunteered their services *pro bono* (although, to our knowledge, the reviewers themselves were paid for their services). Three of the companies used a Web-based platform developed by the Track coordinators to view scanned documents and to record their responsiveness judgments. To avoid problems with local rendering software on each assessor's workstation, the assessors made their judgments based on pdf-formatted versions of the documents, as opposed to the original native format documents. The fourth review company downloaded the pdf documents and conducted the review on their own platform. All review companies were asked to employ their established commercial practice, including their quality assurance procedures.

Assessors were provided with orientation and detailed guidelines created by a Topic Authority. The review platform included a "seek assistance" link which assessors were encouraged to use to request that the Topic Authority resolve any uncertainties. Assessors were instructed to make a responsiveness judgment of responsive ("R"), not responsive ("N"), or broken ("B") for every document assigned to them for review. The latter code reflects the fact that a small percentage of documents from the EDRM dataset are malformed and therefore could not be assessed.

Once the preliminary assessments were complete, quality assurance was conducted by having the Topic Authority adjudicate conflicting assessments, which occurred in one of two cases:

1. For documents selected according to criterion 2 above, the Topic Authority's initial responsiveness determination and the assessor's responsiveness judgment differed; or

2. For documents selected according to criterion 4 above (i.e., the 1000 stratum), the two assessors' judgments differed.

The Topic Authority adjudicated all conflicting documents together, with no indication of which documents had been subject to a previous responsiveness determination, or what that determination had been.

The gold standard consists of:

- The assessor's judgment, for documents without conflicting assessments; and,

- The Topic Authority's final judgment, for documents with conflicting assessments.

The gold standard, along with the toolkit used for the evaluation, may be found on the web:
`http://plg1.uwaterloo.ca/trec11-assess`.

4 The 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. **Initial 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 the likelihood of responsiveness 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.

   The two learning objectives – ranking and estimating the likelihood of responsiveness – 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, from most likely to least likely to be responsive, thereby 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 could be answered by first answering the more detailed question,

> *What is the probability of each document in the set being responsive?*

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 task were required to provide an estimate of the probability of responsiveness for each document in the collection. This probability estimate serves a dual purpose:

1. The documents may be sorted by this probability in order to rank them from the most likely to the least likely to be responsive. It stands to reason that the set of documents ranked 1 through c is likely to contain more of the responsive documents (i.e., to achieve higher recall, precision, and F_1) than some other set of c documents [13].

2. The probabilities of the top-ranked c documents may be summed to yield an estimate of the number of responsive documents that they contain, Rel_c. Furthermore, the probabilities of *all* documents may be summed to yield an estimate of the number of responsive documents in the collection, Rel. From these estimates we may derive estimates of recall ($\frac{Rel_c}{Rel}$), precision ($\frac{Rel_c}{c}$), and $F_1(\frac{2}{\frac{Rel}{Rel_c}+\frac{c}{Rel_c}})$. These estimates, if they are accurate, may be used to inform the selection of the cutoff value c to account for the tradeoffs among recall, precision, effort, and size of production.

Task participants were therefore required to submit, for each document in the collection and for each topic, an estimate of the probability that the document was responsive to the topic. The participants' objectives in supplying these estimates were twofold:

1. To yield a good ranking of documents: for any given cutoff c, the number of responsive documents among the top-ranked c documents should be as large as possible.

2. To yield good effectiveness estimates: for any given cutoff c, the estimate Rel_c should be as close as possible to the actual number of responsive documents, so that the estimates of recall, precision, and F_1 at cutoff c are also as close as possible to their true values.

These objectives are consistent with the requirement in civil litigation to produce as nearly as practicable all and only the documents that are responsive to the request for production, independent of their evidentiary value.

4.1 Submission Phases

For each topic, teams were required to submit an *initial* set of probability estimates prior to requesting any responsiveness determinations from the Topic Authority. Following the initial submission, teams were entitled to receive up to 100 responsiveness determinations before being required to submit an *interim* set of probability estimates. After submitting the first interim results, teams were entitled to receive up to 200 further responsiveness determinations

Topic	Resp.	Nonresp.	Total
401	1,040	1,460	2,500
402	238	1,864	2,102
403	245	1,954	2,199

Table 2: Number of Topic Authority Relevance Determinations for mopup runs.

before submitting a second interim set of results. Thereafter, teams were entitled to receive up to 700 additional responsiveness determinations. In total, each team was allowed to request at most 1,000 responsiveness determinations per topic, subject to submitting the required initial and interim results.

Each team was required to submit a *final* set of probability estimates once it had received all the responsiveness determinations requested by the team. In a final *mopup* phase, all responsiveness determinations requested by all teams were distributed to all teams, who had the opportunity to submit a *mopup* set of probability estimates. Thus, the final submission used only relevance determinations for documents specified by the submitting team, while the mopup submission used relevance determinations for documents specified by all teams. Table 2 shows the total number of responsiveness determinations given to the teams for the mopup phase.

In this Overview, we report results for the *final* and *mopup* submissions. The run identifiers for the various phases may be distinguished by their final symbol: final submissions end in "F"; and mopup submissions end in "M". In the Appendix to the proceedings [15], we provide all results, including those for the initial and interim submissions,[2] whose run identifiers end in "1", "2", and "3".

4.2 Participation Categories

Participants were asked to declare each run to be *automatic* or *technology-assisted*. Automatic runs were allowed to use manual query formulation, but human review of the document collection (other than that provided by TREC via responsiveness determinations) was not permitted. Technology-assisted runs were allowed to avail themselves of any amount of human review. Participants were asked to state the number of hours spent – configuring the system, searching the dataset, reviewing documents, and analyzing the results – as summarized in Table 3. The participation category of a run is specified by the penultimate character in its name: "A" for automatic; and "T" for technology-assisted. For example, the run named "gggxxxAF" is a final run, automatic participation, by the group whose run prefix is ggg, with the letters xxx chosen by the submitting team to distinguish among its submissions.

4.3 Topics

The learning task used three topics: 401, 402 and 403.

- Topic 401 (Topic Authority: Kevin F. Brady, Eckert, Seamans, Cherin & Mellott LLC.)
  All documents or communications that describe, discuss, refer to, report on, or relate to the design, development, operation, or marketing of enrononline, or any other online service offered, provided, or used by the Company (or any of its subsidiaries, predecessors, or successors-in-interest), for the purchase, sale, trading, or exchange of financial or other instruments or products, including but not limited to, derivative instruments, commodities, futures, and swaps.

- Topic 402 (Topic Authority: Brendan M. Schulman, Kramer Levin Naftalis & Frankel LLP.)
  All documents or communications that describe, discuss, refer to, report on, or relate to whether the purchase, sale, trading, or exchange of over-the-counter derivatives, or any other actual or contemplated financial instruments or products, is, was, would be, or will be legal or illegal, or permitted or prohibited, under any existing or proposed rule(s), regulation(s), law(s), standard(s), or other proscription(s), whether domestic or foreign.

[2] Due to logistical challenges, the Track coordinators were unable to enforce the initial and interim submission requirements, with the consequence that these results are incomplete.

Run	Setup	Search	Review	Analysis	Total
HELclrAM	2	1	0	0	3
HELq20rAM	2	1	0	0	3
ISIFuSAM	10	10	0	5	25
ISIFUSAM	10	5	0	10	25
ISILRFTF	40	10	10	10	70
ISILrFTF	40	10	10	10	70
ISIRoTAM	10	10	0	10	30
ISIROTAM	20	10	0	10	40
ISIROTTF	40	10	10	10	70
ISIRoTTF	40	10	10	10	70
ISITrFAM	10	10	0	5	25
ISITRFAM	20	10	0	15	45
ISITrFTF	40	10	10	10	70
ISITRFTF	40	10	20	10	80
mlbclsAF	10	1	0	3	14
mlblrnTF	10	1	10	3	24
mlblrnTM	10	0	0	5	15
otL11BTM	10	1	2	1	14
otL11FTM	10	1	0	0	11
otL11HTM	10	1	2	1	14
priindAM	5	4	5	4	18
rec03TF	20	15	500	120	655
rec04TM	20	30	150	30	230
tcdAF	40	1	0	20	61
URS205AM	400	28	20	20	468
USFDSETF	20	20	48	48	136
USFEOLTF	80	5	36	24	145
USFMOPTF	10	5	4	7	26
UWABASA4	7	0	0	1	8
UWABASAF	40	0	0	8	48
UWABASAM	40	0	0	8	48
UWALINA4	40	0	0	8	48
UWALINAF	40	0	0	8	48
UWALINAM	40	0	0	8	48
UWASNAA4	40	0	0	8	48
UWASNAAF	40	0	0	8	48
UWASNAAM	40	0	0	8	48

Table 3: Self-reported effort (in hours) to configure the participants' systems, search for documents, review documents, and analyze results prior to submission.

Topic	Number of Responsive Documents	95% Confidence Interval
401	20,017	(14,595–25,439)
402	3,012	(1,436–4,588)
403	1,239	(166–2,312)

Table 4: Estimated number of responsive documents for each topic.

- Topic 403 (Topic Authority: Robert Singleton, Squire, Sanders & Dempsey (US) LLP.)
  All documents or communications that describe, discuss, refer to, report on, or relate to the environmental impact of any activity or activities undertaken by the Company, including but not limited to, any measures taken to conform to, comply with, avoid, circumvent, or influence any existing or proposed rule(s), regulation(s), law(s), standard(s), or other proscription(s), such as those governing environmental emissions, spills, pollution, noise, and/or animal habitats.

In contrast to the interactive tasks of TREC 2008 through 2010, for 2011 the Track coordinators did not compose a new mock complaint to provide context for the three topics. Topics 401 and 402 were cast as supplemental requests relating to the 2009 complaint [3], while topic 403 was cast as a supplemental request relating to the 2010 complaint [4].

Table 4 shows the estimated number of responsive documents for each topic, with 95% confidence intervals, calculated using 100 bootstrap samples to estimate the standard error of measurement [7].

5 Evaluation

Each submission was evaluated according to how well it met the objectives of the task: *ranking* and *estimation*.

5.1 Ranking

For representative values of the cutoff value c, representing the number of top-ranked documents to be considered for production, Tables 5 through 7 show recall, precision, and F_1 for each run with respect to each topic. For any given combination of topic and cutoff, higher recall, precision, and F_1 indicate better ranking and hence greater retrieval effectiveness. It follows from the definitions of recall, precision, and F_1 that if one submission is superior to another for a given topic and cutoff, it will be superior on all three measures. The best measures for each combination of topic and cutoff – that is, the best measures in each column – are shown in bold font.

Each row of Tables 5 through 7 illustrates the recall-precision tradeoff inherent in the choice of cutoff. At low cutoff values, precision is generally high while recall is low. At high cutoff values, recall is high while precision is low. F_1 is low at both low and high cutoff values, and peaks somewhere in between.

For the purpose of guiding review strategy, recall conveys completeness as a function of cutoff much more directly than the other measures, answering the question, "If we were to examine the top-ranked c documents, what fraction of the responsive ones would be found?" Precision provides a measure of the efficiency with which a review can be conducted; F_1 sheds no additional light. Precision and F_1 are in fact mathematically redundant, as they may be calculated from recall, given c and Rel. *Gain curves*, shown in Figures 1 through 3, plot recall as a function of cutoff for each of the three topics. Gain curves allow the reader to see at a glance the absolute and relative effectiveness of the submissions, at various cutoff levels.

5.2 Estimation

The recall-precision-F_1 tables and gain curves detailed in the previous section indicate the effectiveness of participants' approaches at various cutoff levels. During the course of an actual review effort, if the gain curves were known, it would be a simple matter to pick the value of c that best captured the desired tradeoff between effort and recall. But the gain curves presented here are not known; they are the result of an extensive evaluation effort that could not reasonably

Cutoff (# docs)	2,000			5,000			20,000			50,000			100,000			200,000		
Run	R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1
HELclrAM	4	42	7	6	24	10	20	20	20	33	14	19	50	10	17	65	7	12
HELq20rAM	4	42	8	6	25	10	20	20	20	33	14	19	50	10	17	65	7	12
ISIFuSAM	10	98	17	8	32	13	11	11	11	10	4	6	11	2	4	25	3	5
ISILRFTF	4	35	7	15	61	24	17	17	17	16	6	9	18	4	6	33	3	6
ISIRoTAM	5	50	9	5	21	8	8	8	8	8	3	4	8	2	3	25	3	5
ISIROTTF	10	98	18	10	38	15	12	12	12	11	5	7	12	2	4	27	3	5
ISITrFAM	7	70	13	13	54	21	**53**	**51**	**52**	65	25	36	69	14	23	71	7	13
ISITRFTF	7	68	12	15	65	25	42	43	42	56	22	32	69	14	23	71	7	13
mlbclsAF	2	17	3	6	23	9	18	18	18	29	12	17	45	9	15	53	5	10
mlblrnTF	3	33	6	9	35	14	14	14	14	14	6	8	13	3	4	18	2	3
mlblrnTM	**10**	**99**	**18**	21	88	34	41	43	42	**70**	**29**	**41**	76	16	26	89	9	17
otL11BTM	6	66	12	15	61	24	25	25	25	23	9	13	24	5	8	34	3	6
otL11FTM	7	74	13	**22**	**99**	**36**	37	39	38	66	26	37	79	15	26	95	9	17
otL11HTM	7	69	13	16	67	26	39	41	40	66	26	37	80	15	26	**96**	**9**	**17**
priindAM	5	50	9	5	20	8	5	5	5	10	4	6	18	4	6	28	3	5
rec03TF	7	68	12	18	75	29	50	51	50	66	26	37	**82**	**16**	**27**	89	9	17
rec04TM	8	77	14	18	75	29	48	48	48	60	23	33	**82**	**16**	**27**	91	9	17
tcdAF	9	99	17	6	23	9	11	11	11	29	12	17	42	9	14	77	8	14
URS205AM	4	41	7	17	68	27	39	37	38	52	21	30	54	11	18	58	6	11
USFDSETF	3	27	5	3	13	5	49	48	48	54	21	30	58	12	19	65	7	12
USFEOLTF	9	86	16	17	62	27	30	29	30	36	14	21	44	9	15	51	5	9
USFMOPTF	6	61	11	9	37	15	43	40	41	57	22	32	60	12	20	61	6	11
UWABASAF	9	97	17	9	37	15	24	24	24	46	18	26	54	11	18	58	6	11
UWABASAM	4	45	8	15	62	24	30	31	31	50	20	28	62	12	21	74	7	13
UWALINAF	1	12	2	5	20	8	10	10	10	21	9	12	42	8	14	67	7	12
UWALINAM	10	89	17	20	82	33	41	43	42	64	25	36	77	16	26	83	9	16
UWASNAAF	6	64	12	8	31	12	22	21	22	38	15	22	38	8	13	38	4	7
UWASNAAM	6	65	12	9	37	15	18	18	18	23	9	13	31	6	10	68	7	12

Table 5: Topic 401 Recall (%), Precision (%), and F_1 at representative document review cutoffs. The best result for each cutoff is shown in bold.

be conducted within the context of a single document review. Instead, the task required participants to estimate their own gain curves in the form of a probability estimate for each document and topic.

Tables 8 through 10 show the participants' estimates of recall for each combination of cutoff and topic. For comparison, the gold-standard estimates (from Tables 5 through 7) are shown, as well as the difference. A positive difference indicates that the participant's estimate was too high; a negative difference indicates that the participant's estimate was too low. All estimates and differences are rounded to the closest integer, so the rounded integer difference shown in the table is not always equal to the difference between the rounded integer estimates.

5.3 Summary measures

No single measure can fully characterize how well a system ranks documents and estimates the probability of responsiveness. Nevertheless, it is useful to have summary measures that roughly capture the effectiveness of the various approaches. As measures of ranking effectiveness, without regard to the accuracy of the participant's estimates, we use *Hypothetical* F_1 and *Area Under the Receiver Operating Characteristic Curve ("AUC").* As a measure to combine ranking effectiveness and estimation accuracy, we use *Actual* F_1. These summary results are shown in Tables 11 through 13.

Note from Tables 5 through 7 that F_1 depends on the cutoff c. Each submission, for each topic, has 685,592

possible F_1 scores – one for each possible value of c. *Hypothetical* F_1 is simply the best of these 685,592 possible scores, calculated by enumerating all possible values of c and calculating F_1 for each. It is called "Hypothetical" because it is achieved only when the optimal cutoff is used, and there is no way to determine this cutoff from the submission itself (i.e., without the gold standard). *Hypothetical* F_1 is the F_1 score that could have been achieved, had this optimal value c been known to the participants when submitting their results.

Actual F_1 relies on the submitted probability estimates to choose c. As with Hypothetical F_1, all possible values of c are enumerated, but instead of choosing after the fact the one that maximizes F_1, we choose c that maximizes the participant's estimate of F_1, which is known beforehand. *Actual* F_1 is the actual value of F_1 (i.e., computed using the gold standard) based on c chosen to maximize the estimated value of F_1 (i.e., computed using the participant's probability estimates). Thus, Actual F_1 is a summary measure that captures both the effectiveness of the ranking and the accuracy of the estimates, and, as such, provides the most informative gauge of the effectiveness of an approach at meeting the retrieval challenge in a real-world scenario (when an after-the-fact gold standard would not be available). Both ranking and estimation must be good in order to achieve a high *Actual* F_1 score.

AUC is a summary measure for ranking effectiveness (regardless of estimation accuracy) derived from signal detection theory [8]. Although its name implies a geometric quantity, AUC has a particularly simple probabilistic meaning: AUC is the probability that a responsive document will be ranked higher than a non-responsive document. It is easily estimated by enumerating all pairs of responsive and non-responsive documents and computing the fraction of pairs for which the relevant document has a higher rank.

6 Discussion

The TREC 2011 Legal Track evaluated the efficacy of various review techniques and tools chosen and implemented by the participating teams. Some participants may have conducted an all-out effort to achieve the best possible results, while others may have conducted experiments to illuminate selected aspects of document review technology. It is inappropriate – and forbidden by the TREC participation agreement – to claim that the results presented here show that one participant's system or approach is generally better than another's. It is also inappropriate to compare the results of TREC 2011 with the results of past TREC Legal Track exercises, as the test conditions as well as the particular techniques and tools employed by the participating teams are not directly comparable. One TREC 2011 Legal Track participant was barred from future participation in TREC for advertising such invalid comparisons.

One may see from the results presented in this Overview that some particular techniques and tools achieved good results in this exercise, and therefore show promise that they might also achieve good results in other document review efforts. The efficacy of the participants' efforts are characterized by the quality of ranking and the accuracy of recall estimates. Efficacy must be interpreted in light of effort, which is characterized by the number of relevance determinations sought from the Topic Authority, as well as by the amount of manual effort employed by the participating team (see Table 3).

The quality of ranking is illustrated in Tables 5 through 7, the gain curves in Figures 1 through 3, and the Hypothetical F_1 and AUC summary measures shown in Tables 11 through 13. The gain curves convey at a glance the tradeoff between recall and cutoff. Figure 1 shows that, for Topic 401, four submissions (otL11FTM, rec03TF, UWALINAM, and mlblrnTM) achieve about 70% recall when only the top-ranked 75,000 documents (11% of the collection) are considered. Assuming that each of these 75,000 documents is reviewed by a human, examining only the top 11% represents a nine-fold saving in review effort, compared to a manual review of the entire collection. Somewhat higher recall may be achieved with more effort, but it is unclear whether improvements in recall measures above 70% are meaningful, given the inherent uncertainties arising from sampling and human assessment of responsiveness [9, 10]. Figure 2 shows that, for Topic 402, one run (rec03TF) achieves a recall of more than 70% when only the top-ranked 20,000 documents (3% of the collection) are considered. Another run (UWALINAM) achieves similar recall when about 40,000 documents (6% of the collection) are considered. It is worth noting that the former is a technology-assisted final run, while the latter is an automatic mopup run. Figure 3 shows that two runs (rec03TF and UWALINAM) achieve 70% recall when only the top-ranked 5,000 documents (less than 1% of the collection) are considered. For all three topics, the achievement of 70% recall at the cutoffs noted above reveals relatively low levels of precision; nonetheless, even at these levels of precision, the savings gained by reviewing the top c documents (rather than the entire collection) would be substantial. The recall, precision, and F_1 measures in Tables 5 through 7 more

Cutoff (# docs)	2,000			5,000			20,000			50,000			100,000			200,000		
Run	R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1
HELclrAM	6	9	7	8	5	6	18	3	5	60	4	7	62	2	4	63	1	2
HELq20rAM	6	9	7	8	5	6	18	3	5	60	4	7	62	2	4	63	1	2
mlbclsAF	3	4	4	4	2	3	13	2	3	46	3	5	74	2	4	88	1	3
mlblrnTM	10	15	12	13	8	10	15	2	4	16	1	2	29	1	2	31	0	1
otL11BTM	12	19	15	14	9	11	14	2	4	15	1	2	15	0	1	32	0	1
otL11FTM	17	26	21	34	21	26	50	8	14	64	4	8	75	2	4	87	1	3
otL11HTM	13	20	16	17	10	13	52	8	14	64	4	8	76	2	5	100	1	3
priindAM	7	11	9	7	4	5	19	3	5	20	1	2	22	1	1	49	1	1
rec03TF	**51**	**77**	**61**	57	35	44	**75**	**12**	**21**	**88**	**6**	**10**	88	3	5	88	1	3
rec04TM	38	57	46	**58**	**37**	**45**	72	12	20	83	5	9	86	3	5	86	1	3
tcdAF	2	3	2	7	4	5	32	5	8	62	4	7	76	2	5	100	1	3
URS205AM	7	11	9	9	5	7	25	4	7	36	2	4	49	1	3	49	1	1
UWABASAF	11	16	13	14	8	10	33	5	9	37	2	4	37	1	2	37	1	1
UWABASAM	15	22	18	18	11	14	37	6	10	52	3	6	65	2	4	64	1	2
UWALINAF	4	6	5	6	4	5	24	4	6	30	2	3	46	1	3	85	1	3
UWALINAM	14	22	17	18	11	13	34	5	9	86	5	10	**99**	**3**	**6**	**100**	**2**	**3**
UWASNAAF	8	12	10	11	6	8	29	4	8	27	2	3	31	1	2	63	1	2
UWASNAAM	11	17	13	13	8	10	30	5	8	32	2	4	65	2	4	64	1	2

Table 6: Topic 402 Recall (%), Precision (%), and F_1 at representative document review cutoffs. The best result for each cutoff is shown in bold.

precisely quantify these observations, and the Hypothetical F_1 and AUC measures in Tables 11 through 13 provide a rough estimate of overall ranking effectiveness.

In practice, a high-quality ranking offers the promise of a review effort that examines only a fraction of the collection – whether 11%, 3%, or 1% – while still achieving substantial recall. To achieve this promise, it is essential to determine, at review time, exactly what fraction of the collection must be reviewed to achieve this end: Is it 1%, 3%, 11%, or some other number? Tables 8 through 13 show the participants' estimates of the recall they thought they achieved for various fractions of the collection. The results are not encouraging. Most runs for most topics dramatically overestimated recall at all cutoff levels. Such an overestimate might lead the manager of a review effort to terminate the review prematurely, due to the false belief that a high level of recall had been achieved. Two participants (Recommind and the University of Waterloo) underestimated recall by a relatively small amount for Topics 401 and 402, and by a much larger amount for Topic 403. Overall, while teams occasionally achieved Actual F_1 scores that came close to the Hypothetical scores (e.g., on Topic 401, one team (Recommind) achieved an Actual F_1 score of 54%, which is reasonably close to their corresponding Hypothetical F_1 score of 58%), no team was able to estimate recall consistently enough to achieve, for all topics, Actual F_1 scores near the Hypothetical F_1 scores that could have been achieved, were their estimates accurate. Overall, consistent recall estimation continues to be a challenge worthy of investigation.

7 Conclusion

The 2011 TREC Legal Track was the sixth since the Track's inception in 2006, and the third that has used a collection based on Enron email (see [5, 14, 12, 11, 6]). From 2008 through 2011, the results show that the technology-assisted review efforts of several participants achieve recall scores that are about as high as might reasonably be measured using current evaluation methodologies. These efforts require human review of only a fraction of the entire collection, with the consequence that they are far more cost-effective than manual review. There is still plenty of room for improvement in the efficiency and effectiveness of technology-assisted review efforts, and, in particular, the accuracy of intra-review recall estimation tools, so as to support a reasonable decision that "enough is enough" and to declare the review

Cutoff (# docs)	2,000			5,000			20,000			50,000			100,000			200,000		
Run	R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1
HELclrAM	11	7	8	16	4	7	28	2	3	62	2	3	64	1	2	66	0	1
HELq20rAM	11	7	9	17	4	7	28	2	3	62	2	3	64	1	2	66	0	1
ISIFUSAM	26	16	20	28	7	11	28	2	3	29	1	1	36	0	1	38	0	0
ISILrFTF	5	3	4	6	2	2	7	0	1	9	0	0	23	0	1	27	0	0
ISIROTAM	20	12	15	20	5	8	21	1	2	22	1	1	31	0	1	33	0	0
ISIRoTTF	6	4	5	6	2	3	8	0	1	9	0	0	23	0	1	27	0	0
ISITRFAM	21	13	16	52	12	20	63	4	7	96	2	5	99	1	2	99	1	1
ISITrFTF	8	5	6	12	3	5	47	3	5	49	1	2	60	1	1	63	0	1
mlbclsAF	2	1	1	2	0	1	3	0	0	4	0	0	4	0	0	4	0	0
mlblrnTM	8	5	6	8	2	3	8	1	1	9	0	0	10	0	0	16	0	0
otL11BTM	28	18	22	32	8	13	64	4	8	65	2	3	67	1	2	68	0	1
otL11FTM	60	37	46	66	17	27	68	4	8	98	2	5	98	1	2	98	1	1
otL11HTM	30	18	23	62	15	24	69	4	8	**100**	**2**	**5**	**100**	**1**	**2**	**100**	**1**	**1**
priindAM	20	12	15	20	5	8	20	1	2	20	0	1	23	0	1	25	0	0
rec03TF	31	19	23	**95**	**26**	**41**	**97**	**7**	**12**	99	3	5	99	1	2	**100**	**1**	**1**
rec04TM	27	17	21	87	23	36	92	7	13	93	3	5	94	1	2	96	1	1
tcdAF	13	8	10	22	5	9	31	2	4	37	1	2	70	1	2	99	1	1
URS205AM	17	11	13	18	4	7	50	3	6	50	1	3	49	1	1	49	0	1
UWABASAF	17	11	13	22	6	9	59	4	7	62	2	3	61	1	1	62	0	1
UWABASAM	**60**	**44**	**51**	62	18	28	62	4	7	65	2	3	64	1	2	66	0	1
UWALINAF	1	1	1	2	0	1	5	0	1	38	1	2	46	1	1	55	0	1
UWALINAM	50	36	42	76	20	31	82	5	9	88	2	4	95	1	2	98	1	1
UWASNAAF	14	9	11	46	11	18	53	3	6	53	1	3	59	1	1	64	0	1
UWASNAAM	45	28	35	52	12	20	60	4	7	60	2	3	62	1	2	66	0	1

Table 7: Topic 403 Recall (%), Precision (%), and F_1 at representative document review cutoffs. The best result for each cutoff is shown in bold.

Cutoff (# docs)	2,000			5,000			20,000			50,000			100,000			200,000		
Run	*est*	*R*	*err*	*est*	*R*	*err*	*est*	*R*	*err*	*est*	*R*	*err*	*est*	*R*	*err*	*est*	*R*	*err*
HELclrAM	1	4	-4	1	6	-5	5	20	-15	13	33	-20	27	50	-23	53	65	-12
HELq20rAM	0	4	-4	1	6	-5	5	20	-15	13	33	-20	26	50	-23	53	65	-12
ISIFuSAM	58	10	+49	65	8	+57	75	11	+64	81	10	+71	86	11	+75	91	25	+66
ISILRFTF	58	4	+55	65	15	+50	75	17	+58	81	16	+65	86	18	+69	91	33	+58
ISIRoTAM	58	5	+53	65	5	+60	75	8	+67	81	8	+74	86	8	+78	91	25	+66
ISIROTTF	58	10	+48	65	10	+55	75	12	+62	81	11	+70	86	12	+74	91	27	+64
ISITrFAM	58	7	+51	65	13	+52	75	53	+21	81	65	+16	86	69	+17	91	71	+20
ISITRFTF	58	7	+51	65	15	+50	75	42	+32	81	56	+25	86	69	+17	91	71	+20
mlbclsAF	27	2	+25	36	6	+30	51	18	+33	60	29	+31	68	45	+23	80	53	+27
mlblrnTF	9	3	+6	11	9	+3	16	14	+2	25	14	+11	38	13	+25	63	18	+46
mlblrnTM	6	10	-4	14	21	-8	34	41	-7	52	70	-18	66	76	-10	79	89	-10
otL11BTM	38	6	+32	76	15	+61	98	25	+74	98	23	+76	98	24	+75	99	34	+64
otL11FTM	28	7	+20	36	22	+14	61	37	+23	81	66	+16	94	79	+15	98	95	+4
otL11HTM	32	7	+26	55	16	+38	79	39	+41	90	66	+24	97	80	+17	99	96	+3
priindAM	49	5	+44	72	5	+67	100	5	+95	100	10	+90	100	18	+82	100	28	+72
rec03TF	8	7	+2	20	18	+2	43	50	-6	51	66	-15	59	82	-23	73	89	-16
rec04TM	7	8	-1	16	18	-2	37	48	-11	46	60	-14	58	82	-23	77	91	-14
tcdAF	1	9	-9	1	6	-4	5	11	-6	11	29	-18	21	42	-21	38	77	-39
URS205AM	7	4	+2	10	17	-7	21	39	-18	36	52	-16	54	54	+0	78	58	+20
USFDSETF	37	3	+34	64	3	+61	82	49	+33	84	54	+29	85	58	+27	88	65	+23
USFEOLTF	48	9	+39	70	17	+53	84	30	+54	85	36	+49	86	44	+42	88	51	+38
USFMOPTF	74	6	+68	84	9	+75	94	43	+52	95	57	+38	96	60	+36	96	61	+35
UWABASAF	0	9	-9	1	9	-8	4	24	-20	10	46	-36	19	54	-35	39	58	-20
UWABASAM	0	4	-4	1	15	-14	4	30	-27	10	50	-41	19	62	-42	38	74	-35
UWALINAF	1	1	-0	2	5	-4	6	10	-4	16	21	-5	30	42	-12	53	67	-14
UWALINAM	2	10	-8	4	20	-16	15	41	-26	34	64	-30	57	77	-20	84	83	+1
UWASNAAF	0	6	-6	1	8	-7	4	22	-18	10	38	-28	20	38	-18	39	38	+1
UWASNAAM	0	6	-6	1	9	-8	4	18	-14	10	23	-13	19	31	-12	38	68	-29

Table 8: Topic 401 Participant-Estimated Recall (%), Actual Recall (%), and Error in Estimate. "+" indicates an overestimate; "-" indicates an underestimate.

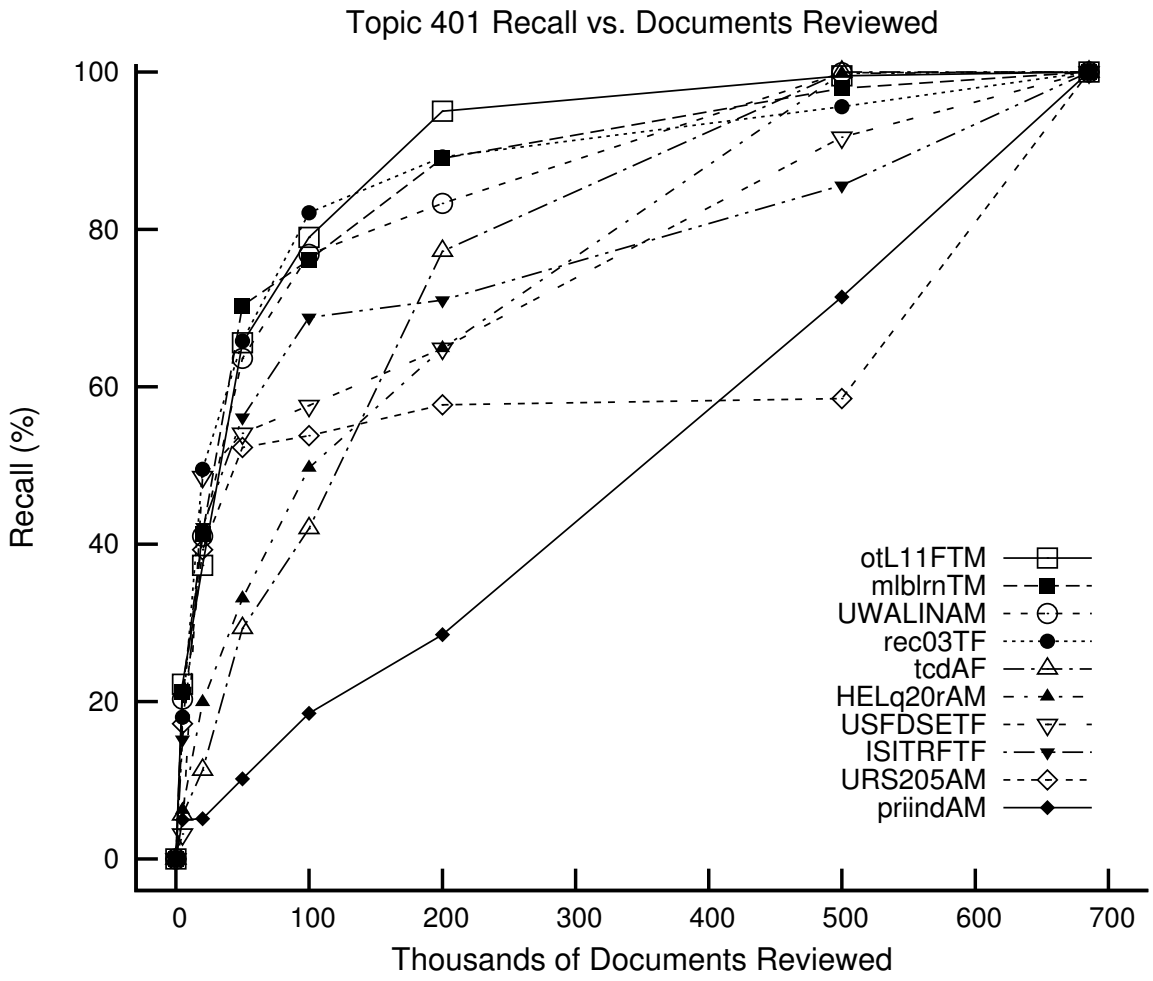

Figure 1: Topic 401 Gain Curves. For each participant, the run with the best AUC score is shown.

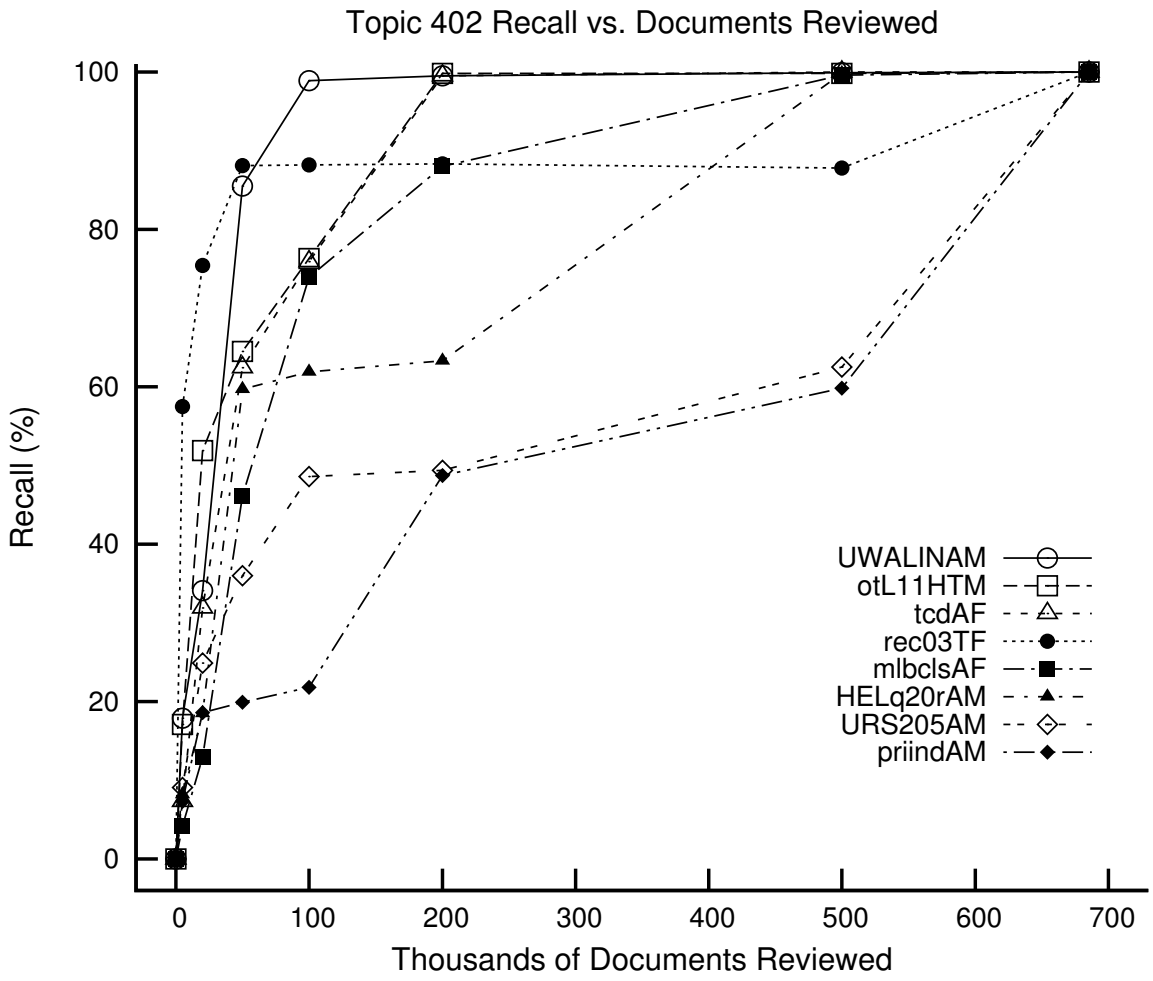

Figure 2: Topic 402 Gain Curves. For each participant, the run with the best AUC score is shown.

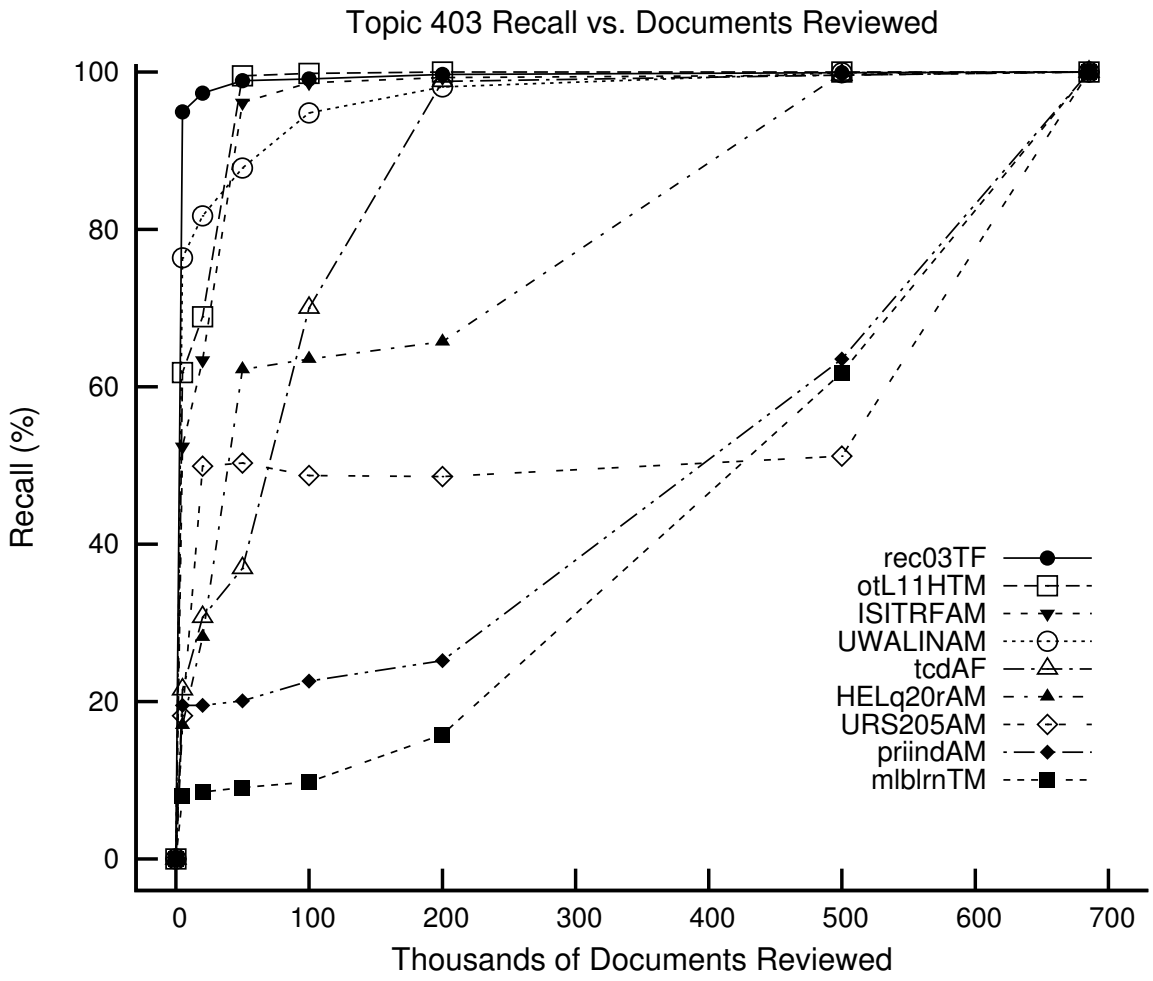

Figure 3: Topic 403 Gain Curves. For each participant, the run with the best AUC score is shown.

Cutoff (# docs)	2,000			5,000			20,000			50,000			100,000			200,000		
Run	*est*	*R*	*err*	*est*	*R*	*err*	*est*	*R*	*err*	*est*	*R*	*err*	*est*	*R*	*err*	*est*	*R*	*err*
HELclrAM	1	6	-6	2	8	-7	6	18	-13	14	60	-46	28	62	-34	55	63	-8
HELq20rAM	1	6	-5	1	8	-7	6	18	-13	14	60	-46	28	62	-34	55	63	-8
mlbclsAF	10	3	+7	24	4	+19	65	13	+53	78	46	+32	83	74	+9	88	88	+0
mlblrnTM	49	10	+39	51	13	+38	55	15	+40	61	16	+46	71	29	+42	84	31	+54
otL11BTM	62	12	+50	93	14	+79	93	14	+79	93	15	+79	94	15	+79	95	32	+62
otL11FTM	28	17	+11	38	34	+4	68	50	+18	87	64	+23	93	75	+18	95	87	+8
otL11HTM	43	13	+29	62	17	+45	80	52	+28	91	64	+27	95	76	+19	96	100	-4
priindAM	35	7	+27	62	7	+54	100	19	+81	100	20	+80	100	22	+78	100	49	+51
rec03TF	22	51	-30	23	57	-34	28	75	-47	36	88	-52	48	88	-41	68	88	-21
rec04TM	16	38	-22	18	58	-40	23	72	-50	31	83	-52	43	86	-42	65	86	-21
tcdAF	0	2	-2	1	7	-6	4	32	-27	10	62	-52	19	76	-57	36	100	-64
URS205AM	6	7	-2	10	9	+1	27	25	+2	46	36	+10	65	49	+17	84	49	+35
UWABASAF	0	11	-10	1	14	-13	4	33	-29	10	37	-27	20	37	-17	39	37	+2
UWABASAM	0	15	-14	1	18	-17	4	37	-33	11	52	-41	21	65	-44	40	64	-24
UWALINAF	1	4	-2	3	6	-3	13	24	-10	31	30	+1	53	46	+8	77	85	-8
UWALINAM	7	14	-7	16	18	-2	45	34	+11	72	86	-14	88	99	-11	97	100	-3
UWASNAAF	0	8	-8	1	11	-10	4	29	-25	10	27	-17	20	31	-11	39	63	-24
UWASNAAM	0	11	-11	1	13	-12	5	30	-25	12	32	-21	22	65	-43	42	64	-22

Table 9: Topic 402 Participant-Estimated Recall (%) and Error in Estimate. "+" indicates an overestimate; "-" indicates an underestimate.

complete. Commensurate with improvements in review efficiency and effectiveness is the need for improved external evaluation methodologies that address the limitations of those used in the TREC Legal Track and similar efforts. How can we construct a gold standard with reasonable effort, or, in the alternative, measure review effectiveness without a gold standard? How best can we measure recall and precision values that are beyond the limit of what can be measured with reference to a single assessor? How can we better control for the amount of effort expended in conducting document review?

The TREC 2011 coordinators determined that it would be not be worthwhile to pursue these research objectives further using the Enron email collection, and endeavored to build a new collection for TREC 2012 and beyond. At the time of writing, the collection was not available, and as a result the TREC Legal Track will not be run in 2012. Work on preparing the collection continues. When complete, the collection will be made available to interested researchers subject to a usage agreement. Further evaluation efforts – whether under the auspices of TREC or a different organization – will be able to use this collection.

Interested researchers may obtain the Enron collection, the Tobacco collection used in the TREC Legal Track from 2006 through 2009, as well as the submissions and evaluation results for the six years of the TREC Legal Track. These collections may be used to reproduce the results reported here, or to conduct new experiments to address the many outstanding research questions that remain.

Acknowledgments

The Legal Track would not have been possible without the efforts of a great many people. Our sincere thanks go to the Topic Authorities (Kevin F. Brady, Brendan M. Schulman, and Robert Singleton) for their diligent efforts in interpreting the topics, providing participating teams with assessments, and adjudicating the results of the sample review. We are grateful to the assessors (and, especially, to the four firms that made the assessors available: ACT Litigation Services, Inc.; Business Intelligence Associates, Inc. ("BIA"); Daegis; and IE Discovery, Inc.) for conducting the first-pass review of the evaluation samples. Finally, we are indebted to Ellen Voorhees and Ian Soboroff of NIST for their patient and thoughtful guidance in sorting out the many issues that arise in conducting an evaluation of this sort.

Cutoff (# docs)	2,000			5,000			20,000			50,000			100,000			200,000		
Run	*est*	*R*	*err*	*est*	*R*	*err*	*est*	*R*	*err*	*est*	*R*	*err*	*est*	*R*	*err*	*est*	*R*	*err*
HELclrAM	1	11	-10	1	16	-15	6	28	-23	13	62	-49	27	64	-37	53	66	-12
HELq20rAM	1	11	-11	1	17	-16	5	28	-23	13	62	-49	27	64	-37	53	66	-13
ISIFUSAM	58	26	+32	65	28	+37	75	28	+47	81	29	+52	86	36	+50	91	38	+54
ISILrFTF	58	5	+53	65	6	+59	75	7	+67	81	9	+72	86	23	+63	91	27	+64
ISIROTAM	58	20	+39	65	20	+45	75	21	+53	81	22	+59	86	31	+55	91	33	+58
ISIRoTTF	58	6	+52	65	6	+58	75	8	+66	81	9	+72	86	23	+63	91	27	+64
ISITRFAM	58	21	+37	65	52	+12	75	63	+11	81	96	-15	86	99	-12	91	99	-8
ISITrFTF	58	8	+51	65	12	+53	75	47	+28	81	49	+32	86	60	+26	91	63	+28
mlbclsAF	0	2	-1	1	2	-1	4	3	+1	9	4	+5	17	4	+13	34	4	+29
mlblrnTM	2	8	-6	3	8	-5	9	8	+0	20	9	+11	36	10	+26	61	16	+46
otL11BTM	53	28	+25	84	32	+52	94	64	+29	94	65	+29	94	67	+27	95	68	+26
otL11FTM	30	60	-31	42	66	-24	71	68	+3	87	98	-10	94	98	-4	95	98	-3
otL11HTM	40	30	+10	59	62	-3	81	69	+12	91	100	-8	95	100	-4	96	100	-4
priindAM	37	20	+18	59	20	+40	100	20	+80	100	20	+80	100	23	+77	100	25	+75
rec03TF	7	31	-24	8	95	-87	12	97	-85	19	99	-80	30	99	-69	50	100	-49
rec04TM	4	27	-23	6	87	-82	11	92	-81	20	93	-73	34	94	-60	61	96	-35
tcdAF	0	13	-13	1	22	-20	4	31	-26	10	37	-27	19	70	-51	35	99	-64
URS205AM	4	17	-13	7	18	-11	18	50	-31	34	50	-16	53	49	+5	77	49	+29
UWABASAF	0	17	-17	1	22	-22	4	59	-55	10	62	-51	20	61	-41	40	62	-22
UWABASAM	0	60	-59	1	62	-61	4	62	-58	10	65	-55	20	64	-45	39	66	-27
UWALINAF	2	1	+1	4	2	+3	14	5	+9	28	38	-9	46	46	-1	68	55	+13
UWALINAM	5	50	-45	12	76	-65	34	82	-48	59	88	-29	80	95	-15	95	98	-3
UWASNAAF	0	14	-14	1	46	-45	4	53	-49	10	53	-43	20	59	-39	40	64	-24
UWASNAAM	0	45	-45	1	52	-51	4	60	-56	10	60	-50	20	62	-42	39	66	-27

Table 10: Topic 403 Participant-Estimated Recall (%) and Error in Estimate. "+" indicates an overestimate; "-" indicates an underestimate.

Run	Hypothetical F_1 (%)	Actual F_1 (%)	AUC (%)
HELclrAM	23 (12-34)	10 (7-12)	78 (72-84)
HELq20rAM	23 (12-34)	10 (7-12)	78 (72-84)
ISIFuSAM	20 (8-32)	0 (0-0)	33 (25-41)
ISILRFTF	25 (14-36)	0 (0-0)	39 (29-49)
ISIRoTAM	10 (4-16)	0 (0-0)	32 (24-40)
ISIROTTF	26 (12-39)	0 (0-0)	34 (26-43)
ISITrFAM	56 (44-67)	0 (0-0)	76 (65-87)
ISITRFTF	45 (33-56)	0 (0-0)	77 (67-86)
mlbclsAF	20 (12-27)	3 (2-4)	67 (58-76)
mlblrnTF	15 (1-30)	6 (4-8)	42 (35-49)
mlblrnTM	47 (37-57)	45 (33-58)	**92 (86-99)**
otL11BTM	35 (24-46)	24 (12-36)	41 (32-50)
otL11FTM	45 (35-55)	12 (9-15)	**92 (88-97)**
otL11HTM	44 (33-56)	25 (13-36)	**92 (87-97)**
priindAM	10 (7-12)	9 (7-11)	49 (42-56)
rec03TF	57 (45-69)	52 (40-65)	89 (83-96)
rec04TM	**58 (46-70)**	**54 (41-67)**	89 (82-96)
tcdAF	20 (12-28)	6 (4-7)	80 (74-87)
URS205AM	39 (28-51)	36 (25-47)	58 (45-70)
USFDSETF	50 (38-62)	7 (0-14)	77 (68-86)
USFEOLTF	36 (24-49)	36 (24-47)	66 (58-75)
USFMOPTF	48 (37-58)	12 (5-20)	76 (67-85)
UWABASAF	28 (20-36)	7 (5-9)	70 (60-79)
UWABASAM	33 (22-44)	7 (5-9)	77 (69-85)
UWALINAF	16 (10-23)	9 (6-11)	76 (69-83)
UWALINAM	42 (31-53)	20 (14-25)	90 (83-97)
UWASNAAF	25 (16-35)	7 (5-9)	67 (59-75)
UWASNAAM	21 (12-31)	7 (5-9)	72 (65-79)

Table 11: Topic 401 summary results: Hypothetical F_1, Actual F_1, and Area Under Receiver Operating Characteristic Curve ("AUC"), as percentages with 95% confidence intervals.

Run	Hypothetical F_1 (%)	Actual F_1 (%)	AUC (%)
HELclrAM	8 (3-13)	2 (1-2)	83 (70-96)
HELq20rAM	8 (3-13)	2 (1-2)	83 (70-96)
mlbclsAF	6 (3-10)	3 (3-4)	89 (82-96)
mlblrnTM	13 (6-19)	8 (2-14)	53 (38-68)
otL11BTM	18 (7-29)	15 (9-21)	52 (37-68)
otL11FTM	33 (13-53)	14 (4-24)	91 (83-100)
otL11HTM	21 (8-34)	19 (8-30)	93 (87-99)
priindAM	14 (4-23)	6 (5-8)	56 (38-74)
rec03TF	**72 (47-96)**	**53 (30-76)**	90 (71-100)
rec04TM	54 (32-76)	45 (20-69)	87 (67-100)
tcdAF	11 (3-18)	1 (0-1)	92 (87-98)
URS205AM	10 (3-17)	6 (0-14)	57 (38-75)
UWABASAF	13 (7-20)	1 (1-2)	64 (45-84)
UWABASAM	19 (9-30)	1 (0-2)	75 (54-96)
UWALINAF	7 (2-13)	3 (1-5)	82 (72-92)
UWALINAM	19 (10-28)	11 (4-18)	**97 (94-100)**
UWASNAAF	19 (3-35)	1 (1-2)	66 (46-85)
UWASNAAM	17 (6-28)	1 (0-2)	74 (54-95)

Table 12: Topic 402 summary results: Hypothetical F_1, Actual F_1, and Area Under Receiver Operating Characteristic Curve ("AUC"), as percentages with 95% confidence intervals.

Run	Hypothetical F_1 (%)	Actual F_1 (%)	AUC (%)
HELclrAM	9 (5-13)	1 (0-1)	83 (65-100)
HELq20rAM	9 (5-14)	1 (0-1)	83 (65-100)
ISIFUSAM	34 (12-56)	1 (0-4)	61 (38-84)
ISILrFTF	5 (2-8)	0 (0-0)	52 (38-67)
ISIROTAM	32 (8-57)	1 (0-3)	58 (38-78)
ISIRoTTF	9 (1-17)	1 (0-2)	53 (38-67)
ISITRFAM	25 (10-40)	1 (0-3)	97 (94-99)
ISITrFTF	14 (0-29)	1 (0-2)	77 (57-96)
mlbclsAF	3 (0-6)	0 (0-1)	18 (9-26)
mlblrnTM	13 (2-25)	6 (0-13)	39 (24-55)
otL11BTM	34 (12-56)	23 (15-31)	81 (60-100)
otL11FTM	59 (29-89)	**34 (10-59)**	95 (92-98)
otL11HTM	38 (16-60)	31 (16-46)	98 (94-100)
priindAM	32 (8-57)	12 (9-16)	51 (32-70)
rec03TF	57 (27-87)	25 (15-34)	**100 (97-100)**
rec04TM	57 (31-82)	6 (0-12)	99 (94-100)
tcdAF	11 (6-15)	0 (0-1)	92 (86-98)
URS205AM	24 (8-40)	4 (0-10)	58 (30-85)
UWABASAF	15 (8-21)	0 (0-1)	81 (67-96)
UWABASAM	52 (9-95)	0 (0-1)	78 (52-100)
UWALINAF	4 (0-10)	1 (0-2)	78 (68-89)
UWALINAM	44 (8-80)	5 (0-10)	97 (94-100)
UWASNAAF	24 (4-43)	0 (0-1)	78 (60-97)
UWASNAAM	**63 (21-100)**	0 (0-1)	78 (52-100)

Table 13: Topic 403 summary results: Hypothetical F_1, Actual F_1, and Area Under Receiver Operating Characteristic Curve ("AUC"), as percentages with 95% confidence intervals.

References

[1] EDRM Data Set Project. Available at http://edrm.net/projects/dataset.

[2] Information released in Enron investigation. Available at http://www.ferc.gov/industries/electric/indus-act/wec/enron/info-release.asp.

[3] TREC-2009 Legal Track – Complaint J, 2009.
Available at http://trec-legal.umiacs.umd.edu/topics/LT09_Complaint_J_final.pdf.

[4] TREC-2010 Legal Track – Complaint K, 2010.
Available at http://trec-legal.umiacs.umd.edu/topics/LT10_Complaint_K_final-corrected.pdf.

[5] Jason R. Baron, David D. Lewis, and Douglas W. Oard. TREC 2006 Legal Track overview. In *Proc. 15th Text REtrieval Conference*, 2006.

[6] Gordon V. Cormack, Maura R. Grossman, Bruce Hedin, and Douglas W. Oard. Overview of the TREC 2010 Legal Track. In *Proc. 19th Text REtrieval Conference*, 2010.

[7] Gordon V. Cormack and Thomas R. Lynam. Statistical precision of information retrieval evaluation. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 533–540. ACM, 2006.

[8] Tom Fawcett. An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[9] Maura R. Grossman and Gordon V. Cormack. Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. *Richmond Journal of Law and Technology*, XVII(3), 2011.

[10] Maura R. Grossman and Gordon V. Cormack. Inconsistent responsiveness determination in document review: Difference of opinion or human errror? *Pace Law Review*, 32(2), 2012.

[11] Bruce Hedin, Stephen Tomlinson, Jason R. Baron, and Douglas W. Oard. Overview of the TREC 2009 Legal Track. In *Proc. 18th Text REtrieval Conference*, 2009.

[12] Douglas W. Oard, Bruce Hedin, Stephen Tomlinson, and Jason R. Baron. Overview of the TREC 2008 Legal Track. In *Proc. 17th Text REtrieval Conference*, 2008.

[13] Stephen E. Robertson. The probability ranking principle in IR. *Journal of Documentation*, 33(4):294–304, 1977.

[14] Stephen Tomlinson, Douglas W. Oard, Jason R. Baron, and Paul Thompson. Overview of the TREC 2007 Legal Track. In *Proc. 16th Text REtrieval Conference*, 2008.

[15] Ellen. M. Voorhees and Lori. P. Buckland (eds.). *Proc. 20th Text REtrieval Conference (TREC 2011) Proceedings*. NIST SP 500-295, 2011.