Learning Task Experiments in the TREC 2010 Legal Track

Stephen Tomlinson
OpenText
Ottawa, Ontario, Canada
stomlins@opentext.com
http://www.opentext.com/

February 28, 2011

Abstract

The Learning Task of the TREC 2010 Legal Track investigated the effectiveness of e-Discovery search techniques at learning from examples to estimate the probability of relevance of every document in a collection. The task specified 8 test topics, each of which included a one-sentence request for documents to produce and several examples of relevant and non-relevant items from a new target collection of 685,592 e-mail messages and attachments. For our participation, we produced three retrieval sets to compare experimental feedback-based, topic-based and Boolean-based techniques. In this paper, we describe the experimental approaches and report the scores that each achieved on various set-based and rank-based measures. We report not just the mean scores of the experimental approaches but also the scores on each of the 8 individual test topics and the largest per-topic impacts of the techniques for several measures. Of the three experimental approaches compared, the experimental feedback-based approach had the highest score in the rank-based F_1@R measure and set-based F_1@K measure for a majority of the test topics.

1 Introduction

OpenText Search Server®, eDOCS Edition (formerly known as Open Text eDOCS SearchServer™) is a toolkit for developing enterprise search and retrieval applications. The eDOCS SearchServer kernel is also embedded in various components of the OpenText eDOCS Suite[1].

The eDOCS SearchServer kernel works in Unicode internally [7] and supports most of the world's major character sets and languages. The major conferences in text retrieval experimentation (TREC [12], CLEF [5] and NTCIR [9]) have provided judged test collections for objective experimentation with the SearchServer kernel in more than a dozen languages.

This paper describes experimental work with the eDOCS SearchServer kernel (experimental post-6.0 builds) conducted in part by participating in the Learning Task of the TREC 2010 Legal Track.

2 Learning Task

The Learning Task of the TREC 2010 Legal Track investigated the effectiveness of e-Discovery search techniques at learning from examples to estimate the probability of relevance of every document in a collection.

The Learning task was a successor task to the Ad Hoc, Relevance Feedback and Batch tasks of past Legal Tracks. We have participated in the 5 years of the Legal Track to date (2006-2010). (We also helped with

[1] OpenText, Open Text eDOCS SearchServer and Open Text eDOCS Suite are trademarks or registered trademarks of Open Text Corporation in the United States of America, Canada, the European Union and/or other countries. This list of trademarks is not exhaustive. Other trademarks, registered trademarks, product names, company names, brands and service names mentioned herein are property of Open Text Corporation or other respective owners.

coordinating the Legal Track in 3 of these years (2007-2009) as described in [19], [10] and [6]; however, we were not part of the coordination of this year's track.)

Compared to last year's Batch Task, the Learning Task used a new document collection (based on Enron e-mail instead of scanned documents from tobacco companies). The requirement to estimate the probability of relevance (of each e-mail or attachment) was also new this year.

The new document collection this year was called the "EDRM Enron Email Data Set v2" collection which consisted of 685,592 e-mail messages and attachments (approximately 4GB of text) from 159 mailbox directories. (By our count, there were 146 different employee mailboxes, with a few large mailboxes split into multiple directories.) We just used the "Deduplicated text-only" version of this collection available in a compressed file called edrmv2txt-v2.tar.bz2. (For binary attachments, this version contained the text extracted by a 3rd-party tool, which was of variable quality.) Uncompressed, the collection contained 685,592 .txt files, totaling 3,991,162,863 bytes. The document id was the part of the filename before the .txt suffix. Each attachment to a message was in a separate .txt file, numbered .1, .2, and so on. For example, container message "3.129461.NC5X5LNTR5XI1CBA3P4QVXG4YOWV5J0NB.txt" had 2 attachments called "3.129461.NC5X5LNTR5XI1CBA3P4QVXG4YOWV5J0NB.1.txt" and "3.129461.NC5X5LNTR5XI1CBA3P4QVXG4YOWV5J0NB.2.txt"; these were 3 of the 685,592 "documents" in the collection.

(The document set in 2010 was a substantial revision of the TREC 2009 Enron collection used in the Interactive task of the preceding year.)

To test the systems, there were 8 production requests, herein called "topics", numbered 200 to 207. 7 of these (numbers 201-207) were taken from the previous year's Interactive Task, which shared a (fictitious) 16-page background complaint regarding "securities fraud". The 8th topic (number 200) was a new one regarding real estate, with no background complaint, though it had 7 sentences of guidelines on what was responsive or not. Each topic included a one-sentence request for documents to produce for each topic. Furthermore, for each topic, several example relevant and non-relevant "seed" documents from the collection were provided (as described further below).

Please see the task guidelines [21] and track overview paper [4] for more details on the task and track. [1] and [2] have more background on e-Discovery in general. Also, background on our past participations in the track are in [15], [16], [17], [18].

2.1 Indexing

To index the collection, we processed the 685,592 .txt files as follows:

Firstly, for each container message (i.e. non-attachment messages, which were identified as those having just 2 dots in the document id instead of 3 dots), we discarded lines which appeared to be "noise" lines, which were those starting with "X-SDOC: ", "X-ZLID: zl-edrm-enron-v2-" or "EDRM Enron Email Data Set has been produced in EML, PST and NSF format by ZL Technologies, Inc". (Note that, for these experiments, we did not bother to take advantage of any of the structure of the e-mail messages. In particular, the "Date:", "From:", "To:" and "Subject:" lines were just treated as plain text like any line of the body of the email.)

Then for each message (including attachments), we added a "<record>" tag before each message, followed by the document id inside "<tid>..</tid>" tags, followed by the message text converted to an XML-safe form (e.g. special characters such as "&" were converted to XML entities such as "&"), followed by a closing "</record>" tag. The output of the re-formatting of the .txt files of each subdirectory was sent to one file, resulting in 228 .xml files (as some of the 159 mailbox directories had more than one subdirectory), but still comprising 685,592 records.

The reason for converting the collection to this XML format was that we could then index it with the same scripts we had used for the IIT CDIP collection of the previous 4 years. As in past years, for each record, we indexed from the "</tid>" tag to the "</record>" tag. Any tags themselves were indexed (we just didn't bother to discard them; a minor side effect is that this meant the term "record" matched every document). Entities (e.g. "&") were converted back to the character they represented (e.g. "&").

We did not use a stopword list, though unlike the past few years, we did not bother to index punctuation characters this year. The index supported both searching on just the surface forms of the words and also

searching on inflections from English lexical stemming. The documents were assumed to be in the Windows-1252 character set when converted to Unicode. Words were normalized to upper-case and any accents were dropped.

2.2 Training Examples

The training examples (also known as "training judgments", "training qrels" or "seed documents") for all 8 test topics were provided in a file called "seed.csv". For 7 of the topics (201-207), these were based on documents judged in the previous year's Interactive task, when the organizers were able to find the documents in the new collection (there apparently was no straightforward mapping from the document ids used in one collection to the other).

We reported some issues with the seed.csv file to the track mailing list (July 31, 2010), most notably that some messages were listed multiple times, with different judgments. We discarded lines with duplicate document ids as follows:

First, some lines were obviously "broken" in that they included an md5 sum in the first column (which was supposed to just be the case for attachments) but they did not include the attachment number (e.g. ".1", ".2", etc.) in the 4th column. So we dicarded these lines (there were 97 such lines).

Then, for any other case of a duplicated document id for a topic, we just kept the first judgment. This caused 25 more lines to be removed for topic 200, 10 more lines to be removed for topic 202, and 4 or fewer more lines to be removed for each of the other topics.

We also re-formatted the remaining lines to the traditional 4-column qrels format used by the trec_eval utility, which is also readable by the l07_eval utility used by the track in the previous 3 years. Our output of all this processing was a new file called "qrelsL10.seed".

The following list shows, for each of the 8 topics, the count of the number of relevance judgments in qrelsL10.seed, the number judged relevant, and the number judged non-relevant. Note that the "count" may exceed the sum of the "rel" and the "non" because some documents were "gray" (had a label of -1 or -2, indicating that the assessor didn't render a judgment for some reason):

```
Topic 200: count=822, rel=205, non=617
Topic 201: count=724, rel=168, non=516
Topic 202: count=1430, rel=990, non=393
Topic 203: count=1024, rel=65, non=878
Topic 204: count=1219, rel=59, non=1122
Topic 205: count=1951, rel=330, non=1499
Topic 206: count=352, rel=18, non=324
Topic 207: count=583, rel=80, non=492
```

As can be seen from the above list, the number of example relevant documents (after our arbitrary processing to remove duplicate messages) ranged from 18 (for topic 206) to 990 (for topic 202).

2.3 Feedback Run - otL10FT

The submitted experimental otL10FT run was a pure feedback run that did not make any use of the topic statements. Instead, the feedback technique was just based on the set of documents that were previously judged relevant (the "feedback set").

In the case of topic 202, which had a large number of example relevant documents (990 as mentioned earlier), we reduced the number in the feedback set to 305 by taking a random sample. This was just to guard against potential buffer overflow in some later steps.

Then documents of 10,000 bytes or more (in the XML formatting described earlier) were excluded from the feedback set in hopes of reducing the percentage of input text that was not relevant. (This step seemed to have been helpful last year for the otL09F run.) The resulting number of relevant documents, in order by topic, were 197, 115, 193, 36, 50, 232, 15 and 75.

Also, we created an alternate feedback set that further excluded documents of 1000 bytes or more. Its resulting number of relevant documents, in order by topic, were 96, 24, 24, 6, 6, 12, 6 and 26.

The documents from each feedback set for each topic were used as the input to the SearchServer IS_ABOUT predicate which created a vector query from the highest weighted terms (based on a tf.idf calculation after appending the input documents together). English inflections were enabled, and stems in more than 5% of the collection's documents were omitted.

To decide which result set to use for each topic (the one based on documents less than 1000 bytes in size, or the one based on documents less than 10,000 bytes in size), we scored each result set using the full qrelsL10.seed judgments with the l07_eval utility, and picked the result set for each topic with the higher score in the induced average precison measure (indAP), which is denoted ":mapJudged:" in l07_eval; (indAP is like the traditional average precisiion measure, except that unjudged documents are omitted). It turned out that the more restrictive feedback set scored higher for the first 5 topics (200-204) and the less restrictive set scored higher for the latter 3 topics (205-207).

To assign a probability of relevance to each result, our main input was the relevance() function score from SearchServer, which is a score between 0 and 1000 for each document (though most documents score between 0 and 500). These scores were found to have some value for predicting query difficulty in the TREC 2004 Robust Track [20]. Roughly speaking, the relevance() score is higher when there are rare terms (i.e. terms of high inverse document frequency) that match, which intuitively is a good reason to be more confident of the relevance of the document.

(The relevance ranking approach was the same for all runs, and also the same as in past years. The relevance function dampened the term frequency and adjusted for document length in a manner similar to Okapi [11] and dampened the inverse document frequency using an approximation of the logarithm. For runs which used inflectional matching (which was the case for all 3 submitted runs this year), these calculations were based on the stems of the terms.)

Our experimental probability formula for these experiments was to take the raw relevance() score (which, again, was usually between 0 and 500), multiply it by 0.002, square it, divide by 0.75, and enforce a max of 0.75 and min of 0.0001.

Furthermore, for this feedback run, the seed relevant documents (from qrelsL10.seed) were moved to the front and assigned probability 0.75. And any documents unmatched by the feedback query were appended to the end with the min probability of 0.0001.

The resulting sum of the probabilities for each topic was 10990, 5752, 7478, 4404, 4466, 15743, 6346 and 8721, which corresponds to the predicted number of relevant documents for the topic. We compared this to the estimated number of relevant documents for the 7 Interactive topics of last year (the latter 7 topics, 201-207), which were 2454, 9514, 1831, 3242, 33614, 26343 and 26420, which was a factor in deciding on the 0.002 and squaring parts of the formula. While the fit was not all that close, the collection was said to be quite different this year, with duplicates removed, and new messages added, so we did not attempt to fit individual topics any more closely.

The max probability of 0.75 was chosen based on looking at the past assessor agreement on relevant documents. The min probability of 0.0001 seemed small enough to not affect the overall estimates much.

2.4 Request Run - otL10rvlT

The submitted experimental otL10rvlT run did not make any use of the training examples; instead, it just used the one-sentence request. Common instruction words (e.g. "please", "produce", "documents") were manually removed. For example, for topic 202, for which the request text was "All documents or communications that describe, discuss, refer to, report on, or relate to the Company's engagement in transactions that the Company characterized as compliant with FAS 140 (or its predecessor FAS 125).", the WHERE clause of the corresponding SearchSQL statement was

```
WHERE FT_TEXT CONTAINS 'engagement'|'transactions'|'compliant'|'FAS'|'140'|
        'predecessor'|'FAS'|'125'
```

Linguistic expansion from English inflectional stemming was applied, e.g. the search for 'engagement' also matched 'engagements'.

The experimental probability formula was the same as for the other runs, i.e. start with the raw relevance() score (which was usually between 0 and 500), multiply by 0.002, square it, divide by 0.75, and enforce a max of 0.75 and min of 0.0001. As mentioned earlier, the density of relevant documents on last year's collection was a factor in settling on this (experimental) general-purpose probability formula, so one could argue there was a modest feedback influence in this sense, though again, this run did not make any use of the provided training examples, and there was no tuning of the probabilities for individual topics.

Any documents unmatched by the request query were appended to the end with the min probability of 0.0001.

This otL10rvlT run is meant to be considered as a baseline run representing what can be done by a mostly-automatic approach without using the training examples.

2.5 Boolean Run - otL10bT

The submitted experimental otL10bT run was a Boolean-based run.

There was no organizer-provided reference Boolean query for the 8 topics (unlike for the Ad Hoc tasks of past Legal Tracks, which included a mock Boolean negotiation for each topic). So we attempted to quickly create our own Boolean query for each topic. The approach was to look at some of the example relevant documents matched by the feedback run (described earlier), with the feedback terms highlighted in those documents. We then manually picked keywords or phrases that seemed like they might be particularly useful for identifying relevant documents (which may or may not have included terms in the feedback query). Of course, this Boolean term selection was subjective, and we had a bias to making the Boolean query fairly short, and we did not want to spend a lot of time fine-tuning the query. One can perhaps think of this run as a hastily produced manual or interactive run.

The resulting Boolean queries were as follows:

```
200: FT_TEXT CONTAINS 'house'|'rental'|'apartment'|'condo'
201: FT_TEXT CONTAINS 'pre-pay'|'swap'
202: FT_TEXT CONTAINS 'FAS'|'transaction'|'swap'|'trust'|'Transferor'|'Transferee'
203: FT_TEXT CONTAINS 'forecast'|'earnings'|'profit'|'quarter'|'balance sheet'
204: FT_TEXT CONTAINS 'retention'|'compliance'|'preserve'|'discard'|'destroy'|'delete'|
     'clean'|'eliminate'|'shred'|'schedule'|'period'|'documents'|'file'|'policy'|'e-mail'
205: FT_TEXT CONTAINS 'electricity'|'electric'|'loads'|'hydro'|'generator'|'power'
206: FT_TEXT CONTAINS 'analyst'|'credit'|'rating'|'grade'
207: FT_TEXT CONTAINS 'football'|'Eric Bass'
```

Note that linguistic expansion from English inflectional stemming was also applied, e.g. the search for "house" also matched "housing", "houses", "housed" and "house's". Also, the hyphenated search for "e-mail" also matched non-hyphenated forms such as "email", "emails", "emailed", "emailing" and "email's" (along with hyphenated forms such as "e-mails", "e-mailing", "e-mailed", and so on).

Note also that the matches were still relevance-ranked. (For terms in phrases of Boolean queries, only occurrences of the term satisfying the phrase counted towards term frequency.) The same experimental probability formula was used as for the other runs, i.e. start with the raw relevance() score (which was usually between 0 and 500), multiply it by 0.002, square it, divide by 0.75, and enforce a max of 0.75 and min of 0.0001.

Any documents unmatched by the Boolean query were appended to the end with the min probability of 0.0001. (It was required to submit all of the documents for each topic.)

Note that the sum of the probabilities was typically different than the number of matches for the Boolean query. e.g. for topic 200, the Boolean query had 17,917 matches, but the sum of the probabilities from the (experimental) general-purpose formula was just 3430.

3 Results

We submitted our 3 experimental runs (otL10FT, otL10rvlT and otL10bT) by the August 25, 2010 deadline. The task organizers then had a sample of the test collection judged for relevance as the basis for estimating the various scores, such as recall, precision and F_1. The details presumably will be in the track overview paper [4], but our understanding from the discussion on the track mailing list is that it proceeded as follows.

The test collection was divided into 4 strata, called stratum 100, stratum 1000, stratum 10000, and stratum 1000000. Our understanding is that any document that was ranked in the top-100 by any participant submission was in stratum 100 (and this stratum was completely judged). Any remaining document that was ranked in the top-1000 by any participant submission was in stratum 1000, but only 5-10% of this stratum was judged. And so on for the other 2 strata.

The task organizers produced a preliminary set of judgments (qrels.t10legallearn.prelim) on October 12, 2010, in time for the October 25 notebook paper deadline and November 16-19 conference. The final set of judgments (qrels.t10legallearn) were released January 19, 2011. In this paper, we just use the final set of judgments.

Based on the final judgments in qrels.t10legallearn, we produced our own counts of the number of documents in each stratum, the number judged in each stratum, and the ratio (which is the probability of each document in that stratum being chosen for judging), which are listed here:

```
Topic 200:
 stratum 100: count=918, judged=918, prob=1.000000000000
 stratum 1000: count=8707, judged=600, prob=0.068910072356
 stratum 10000: count=74838, judged=600, prob=0.008017317406
 stratum 1000000: count=601129, judged=602, prob=0.001001448940
Topic 201:
 stratum 100: count=1050, judged=1050, prob=1.000000000000
 stratum 1000: count=7271, judged=556, prob=0.076468161188
 stratum 10000: count=73807, judged=556, prob=0.007533160811
 stratum 1000000: count=603464, judged=558, prob=0.000924661620
Topic 202:
 stratum 100: count=1146, judged=1146, prob=1.000000000000
 stratum 1000: count=5439, judged=524, prob=0.096341239198
 stratum 10000: count=64236, judged=524, prob=0.008157419516
 stratum 1000000: count=614771, judged=526, prob=0.000855603143
Topic 203:
 stratum 100: count=1201, judged=1201, prob=1.000000000000
 stratum 1000: count=8586, judged=506, prob=0.058933146983
 stratum 10000: count=80978, judged=506, prob=0.006248610734
 stratum 1000000: count=594827, judged=507, prob=0.000852348666
Topic 204:
 stratum 100: count=964, judged=964, prob=1.000000000000
 stratum 1000: count=9995, judged=585, prob=0.058529264632
 stratum 10000: count=84066, judged=585, prob=0.006958818072
 stratum 1000000: count=590567, judged=586, prob=0.000992266754
Topic 205:
 stratum 100: count=1172, judged=1172, prob=1.000000000000
 stratum 1000: count=6428, judged=516, prob=0.080273802116
 stratum 10000: count=45913, judged=516, prob=0.011238647006
 stratum 1000000: count=632079, judged=516, prob=0.000816353652
Topic 206:
 stratum 100: count=1077, judged=1077, prob=1.000000000000
 stratum 1000: count=9367, judged=547, prob=0.058396498345
```

```
 stratum 10000: count=77038, judged=547, prob=0.007100392014
 stratum 1000000: count=598110, judged=549, prob=0.000917891358
Topic 207:
 stratum 100: count=976, judged=976, prob=1.000000000000
 stratum 1000: count=6714, judged=581, prob=0.086535597259
 stratum 10000: count=84580, judged=581, prob=0.006869236226
 stratum 1000000: count=593322, judged=582, prob=0.000980917613
```

The counts for each topic should sum to 685,592 (the number of documents in the collection). The number of judged documents added to 2720 for each topic.

3.1 L07 vs. L10 measures

The task organizers developed a new approach to estimating scores from the samples, based on individually estimating the precision on each stratum and then extrapolating over the entire stratum. We call this approach the "L10" approach, in contrast to the "L07" approach that was used the previous 3 years (for which we led the design when helping to coordinate the task). The L07 approach essentially assigned a fixed weight to each judged document based on the reciprocal of its probability of being judged. We reported on the track mailing list (Oct 18, 2010) that the new L10 approach had some anomalies, such as that if set D1 was a strict superset of set D2, it could still estimate recall(D1) to be less than recall(D2). Furthermore, the recall of a set could be estimated to be greater than 100%. (These particular anomalies could not happen with the L07 approach.) In this paper, we generally just report the L07-based scores (re-using scripts that we had set up in past years).

To compute the L07-based scores, we created a qrelsL10.probs file (for use with the l07_eval scoring utility) by taking the judged documents from qrels.t10legallearn and assigning them the probability as listed in the previous section.

To get an idea of how similar or different the L07 and L10 approaches are, Table 1 compares the estimated F_1@K score from each approach for the experimental otL10FT run. For the "K" value, i.e. the retrieval depth at which to estimate F_1, we use the "Cutoff Estimate" reported by the task organizers in otL10FT.sum, which was the depth at which the (L10) F_1 would be expected to be maximized if the run's probabilities of relevance were accurate. The L10 F_1@K score was reported in otL10FT.sum as "F1". We see in Table 1 that the F_1@K scores are almost the same for all 8 topics, suggesting that both estimation approaches are likely to lead to similar conclusions.

Note: The detailed L07 formulas for estimating the number of relevant and non-relevant documents for each topic, and also for estimating precision and recall, were reported in the 2007 Ad Hoc task section of [19], and the detailed formulas for estimating F_1 were reported in the 2008 Ad Hoc task section of [10]. The l07_eval software used to compute the L07 evaluation measures is online at `http://trec.nist.gov/data/legal09.html` .

3.2 L07 measures

For each topic, we report a table of set-based scores and a table of rank-based scores. Of the various measures, probably the most informative set-based measure is F_1@K and most informative rank-based measure is F_1@R.

F_1@R is easy to interpret because depth R (where R is the estimated number of relevant documents) is the special depth at which recall, precision and F_1 all have the same value. For F_1@R, just the relative probabilities of relevance matter (i.e., the ranking) not the absolute probabilities.

F_1@K is a more challenging measure because the system has to choose the depth K at which to be evaluated. This K value is implied by the absolute probabilities of relevance calculated by the system. Normally the best K value is approximately the same as R (which, of course, is not known to the system in advance). If F_1@K is substantially less than F_1@R, then the system likely substantially overestimated or underestimated the probabilities of relevance. The F_1 measure requires both high precision and high recall to achieve a high score. Overestimating the probabilities leads to too high a K value and typically lowers

Topic	K	L07 F_1@K	L10 F_1@K
200	19097	0.080	0.082
201	9305	0.068	0.068
202	9578	0.316	0.332
203	7553	0.261	0.258
204	11780	0.099	0.095
205	26615	0.409	0.406
206	12628	0.059	0.062
207	15588	0.175	0.169

Table 1: Comparison of the Estimated F_1@K Scores of the Experimental otL10FT Run using the l07_eval Approach of 2007-2009 ("L07") and New Approach of 2010 ("L10")

precision substantially, which in turn lowers F_1. Underestimating the probabilities leads to too low a K value and typically lowers recall substantially, which in turn lowers F_1.

In the tables of set-based scores, we include not just the 3 experimental submissions (otL10FT, otL10rvlT, otL10bT), but also 3 reference runs as follows:

- "seeds-rel" is the set of relevant seed documents.
- "seeds-non" is the set of non-relevant seed documents.
- "fullsetL10" is the set consisting of the entire document collection.

The tables of set-based measures have the following columns:

- "K": For the 3 submitted runs, K came from the organizer-provided "Cutoff Estimate" as described earlier, i.e. the cutoff at which the (L10) F_1 would be expected to be maximized if the run's probabilities of relevance were accurate. (Note that the K value is a property of the run and is computable before the relevance judgments are known.) For the 3 reference runs, K is the size of the set.
- "R@K": The estimated recall at depth K. Recall is the estimated number of relevant documents retrieved (at depth K) divided by the estimated total number of relevant documents (in the entire collection). (From ":est_K-Recall:" in l07_eval output.)
- "P@K": The estimated precision at depth K. Precision is the estimated number of relevant documents retrieved (at depth K) divided by the sum of the estimated number of relevant and non-relevant documents (at depth K). (From ":est_K-Prec:" in l07_eval output.)
- "F_1@K": The estimated F_1 score at depth K. F_1 is 2*Precision*Recall/(Precision+Recall) or 0 if both Precision and Recall are 0. (Note that this F_1 formula is only applicable for individual topics; the mean F_1 across topics may differ from plugging the mean precision and recall into the formula.) (From ":est_K-F1:" in l07_eval output.)
- "Num. Judged@K" is the actual number of judged documents in the top-K, followed in parentheses by the actual number of judged relevant (r), non-relevant (n) and gray (g) documents. Note that because not all documents were drawn for judging with the same probability, the estimated numbers of relevant and non-relevant documents in a result set are not in general exactly proportional to the drawn numbers. (From ":K-jg_ret:", ":K-rel_ret:", ":K-nonrel_ret:" and ":K-gray_ret:" in l07_eval output respectively.)

The table caption reports the estimated number of relevant documents for the topic.
The tables of rank-based measures have the following columns:

Run	K	R@K	P@K	F_1@K	Num. Judged@K
otL10bT	8929	0.442	0.130	**0.201**	580 (208r, 372n, 0g)
seeds-rel	205	0.047	**0.610**	0.087	142 (106r, 36n, 0g)
otL10FT	19097	0.343	0.045	0.080	731 (217r, 514n, 0g)
seeds-non	617	0.023	0.280	0.043	47 (18r, 29n, 0g)
otL10rvlT	21711	0.043	0.005	0.009	551 (54r, 497n, 0g)
fullsetL10	685592	**1.000**	0.004	0.007	2720 (261r, 2459n, 0g)

Table 2: Set-based Scores for Topic 200 (2543.5 Est. Relevant Documents)

Run	P@B	R@B	F_1@R	indAP	GS10J	First 10 Ret
otL10FT	0.047	0.342	**0.179**	**0.549**	**1.000**	RNRRNRRNRR
otL10bT	**0.077**	**0.447**	0.161	0.373	**1.000**	RNNRRRRNRR
otL10rvlT	0.006	0.041	0.024	0.115	0.681	NNNNNRNNNN

Table 3: Rank-based Scores for Topic 200 (B=17917, R=2544)

- "P@B" and "R@B": Estimated Precision and Recall at Depth B (where B is the number of documents matching our experimental Boolean query (used for otL10bT), which is listed in the table caption). (From ":est_PB:" and ":est_RB:" in l07_eval output respectively.)
- "F_1@R": Estimated F_1 at Depth R (where R is the estimated number of relevant documents, which is listed in the table caption). (From ":est_R-F1:" in l07_eval output.)
- "indAP": Induced Average Precision (the popular "average precision" after discarding unjudged documents; the sampling probabilities are not used for this measure, i.e. indAP is not infAP or statAP). (From ":mapJudged:" in l07_eval output.)
- "GS10J": Generalized Success@10 on Judged Documents (1.08^{1-r} where r is the rank of the first relevant document, only counting judged documents, or zero if no relevant document is retrieved). GS10J is a robustness measure which exposes the downside of blind feedback techniques [13]. "Generalized Success@10" was originally introduced as "First Relevant Score" (FRS) in [14]. Intuitively, GS10J is a predictor of the percentage of topics for which a relevant document is returned in the first 10 rows. (From ":GS10J:" in l07_eval output.)
- "First 10 Ret": The judgments of the top-10 ranked documents of the run. 'R' indicates judged relevant. 'N' indicates judged non-relevant. (From ":relstring:" in l07_eval output.)
- "S1J": Success of the First Judged Document (in table of mean scores only).

The highest scores of each measure are in bold; however, see Table 20 for which mean differences may be statistically significant.

Run	K	R@K	P@K	F_1@K	Num. Judged@K
otL10bT	16217	0.728	0.083	**0.149**	909 (315r, 594n, 0g)
seeds-rel	168	0.045	**0.519**	0.082	150 (72r, 78n, 0g)
otL10FT	9305	0.194	0.041	0.068	562 (233r, 329n, 0g)
otL10rvlT	39234	0.648	0.031	0.059	1229 (285r, 944n, 0g)
fullsetL10	685592	**1.000**	0.003	0.005	2720 (384r, 2336n, 0g)
seeds-non	516	0.001	0.004	0.001	8 (1r, 7n, 0g)

Table 4: Set-based Scores for Topic 201 (1885.9 Est. Relevant Documents)

Run	P@B	R@B	F_1@R	indAP	GS10J	First 10 Ret
otL10bT	**0.058**	**0.827**	**0.209**	**0.525**	**1.000**	RRRRRRRRNR
otL10FT	0.034	0.425	0.107	0.444	**1.000**	RNRRRRNNNR
otL10rvlT	0.043	0.600	0.102	0.289	**1.000**	RRRRRRNNRN

Table 5: Rank-based Scores for Topic 201 (B=25561, R=1886)

Run	K	R@K	P@K	F_1@K	Num. Judged@K
otL10FT	9578	0.436	0.247	**0.316**	1033 (895r, 138n, 0g)
otL10bT	17225	0.456	0.171	0.249	1209 (865r, 344n, 0g)
seeds-rel	990	0.144	**0.827**	0.245	376 (345r, 31n, 0g)
otL10rvlT	8299	0.155	0.110	0.129	651 (441r, 210n, 0g)
fullsetL10	685592	**1.000**	0.009	0.018	2720 (996r, 1724n, 0g)
seeds-non	393	0.000	0.000	0.000	9 (0r, 9n, 0g)

Table 6: Set-based Scores for Topic 202 (6312.4 Est. Relevant Documents)

Run	P@B	R@B	F_1@R	indAP	GS10J	First 10 Ret
otL10FT	**0.069**	**0.675**	**0.358**	**0.904**	**1.000**	RRRRNRRRRR
otL10bT	0.057	0.627	0.291	0.748	0.429	NNNNNNNNNN
otL10rvlT	0.059	0.584	0.109	0.543	**1.000**	RNRRNNRRRR

Table 7: Rank-based Scores for Topic 202 (B=66770, R=6313)

Run	K	R@K	P@K	F_1@K	Num. Judged@K
otL10FT	7553	0.432	0.187	**0.261**	835 (409r, 426n, 0g)
otL10bT	11005	0.352	0.127	0.187	860 (381r, 479n, 0g)
otL10rvlT	44131	0.554	0.042	0.078	1018 (394r, 624n, 0g)
seeds-rel	65	0.012	**0.569**	0.023	65 (37r, 28n, 0g)
fullsetL10	685592	**1.000**	0.005	0.009	2720 (481r, 2239n, 0g)
seeds-non	878	0.001	0.005	0.002	15 (4r, 11n, 0g)

Table 8: Set-based Scores for Topic 203 (3124.6 Est. Relevant Documents)

Run	P@B	R@B	F_1@R	indAP	GS10J	First 10 Ret
otL10FT	**0.071**	**0.829**	**0.303**	**0.568**	0.857	NNRRRRNRRR
otL10bT	0.057	0.798	0.266	0.523	**0.926**	NRRNNRNNRN
otL10rvlT	0.044	0.541	0.111	0.312	0.270	NNNNNNNNNN

Table 9: Rank-based Scores for Topic 203 (B=41609, R=3125)

Run	K	R@K	P@K	F_1@K	Num. Judged@K
otL10FT	11780	0.143	0.076	**0.099**	631 (292r, 339n, 0g)
otL10rvlT	17481	0.128	0.049	0.070	586 (252r, 334n, 0g)
otL10bT	21896	0.141	0.043	0.065	689 (306r, 383n, 0g)
fullsetL10	685592	**1.000**	0.009	0.018	2720 (475r, 2245n, 0g)
seeds-rel	59	0.006	**0.678**	0.013	59 (40r, 19n, 0g)
seeds-non	1122	0.000	0.005	0.000	9 (1r, 8n, 0g)

Table 10: Set-based Scores for Topic 204 (6361.8 Est. Relevant Documents)

Run	P@B	R@B	F_1@R	indAP	GS10J	First 10 Ret
otL10FT	0.021	0.680	**0.108**	0.551	**1.000**	RRRRRRRRRR
otL10rvlT	**0.027**	**0.817**	0.074	0.407	**1.000**	RRRRRRNRRN
otL10bT	0.016	0.494	0.064	**0.603**	**1.000**	RRRRRRRRRR

Table 11: Rank-based Scores for Topic 204 (B=203212, R=6362)

Run	K	R@K	P@K	F_1@K	Num. Judged@K
otL10FT	26615	0.285	0.722	**0.409**	1425 (1214r, 211n, 0g)
otL10bT	24535	0.234	0.634	0.341	1285 (1109r, 176n, 0g)
otL10rvlT	25801	0.206	0.480	0.288	1076 (877r, 199n, 0g)
fullsetL10	685592	**1.000**	0.099	0.180	2720 (1366r, 1354n, 0g)
seeds-rel	330	0.004	**0.837**	0.008	188 (158r, 30n, 0g)
seeds-non	1499	0.000	0.000	0.000	4 (0r, 4n, 0g)

Table 12: Set-based Scores for Topic 205 (67938.0 Est. Relevant Documents)

Run	P@B	R@B	F_1@R	indAP	GS10J	First 10 Ret
otL10FT	**0.428**	**0.636**	**0.482**	**0.883**	**1.000**	RRRRNRRRRR
otL10bT	0.377	0.549	0.442	0.881	**1.000**	RRRRRRRRRR
otL10rvlT	0.406	0.555	0.384	0.817	0.926	NRRRRRRRRR

Table 13: Rank-based Scores for Topic 205 (B=99233, R=67938)

Run	K	R@K	P@K	F_1@K	Num. Judged@K
otL10FT	12628	0.470	0.032	**0.059**	674 (117r, 557n, 0g)
otL10rvlT	28169	0.778	0.026	0.050	760 (131r, 629n, 0g)
otL10bT	39235	0.941	0.021	0.040	823 (132r, 691n, 0g)
seeds-rel	18	0.002	**0.111**	0.004	18 (2r, 16n, 0g)
fullsetL10	685592	**1.000**	0.001	0.003	2720 (135r, 2585n, 0g)
seeds-non	324	0.000	0.000	0.000	1 (0r, 1n, 0g)

Table 14: Set-based Scores for Topic 206 (866.2 Est. Relevant Documents)

Run	P@B	R@B	F_1@R	indAP	GS10J	First 10 Ret
otL10FT	0.007	0.673	**0.249**	0.311	0.857	NNRNNRNNNN
otL10rvlT	0.010	**1.000**	0.144	0.232	0.270	NNNNNNNNNN
otL10bT	**0.010**	0.980	0.114	**0.324**	**1.000**	RNRRRNRRRN

Table 15: Rank-based Scores for Topic 206 (B=88263, R=867)

Run	K	R@K	P@K	F_1@K	Num. Judged@K
otL10rvlT	3189	0.131	0.743	**0.222**	715 (659r, 56n, 0g)
otL10FT	15588	0.143	0.224	0.175	942 (720r, 222n, 0g)
otL10bT	4205	0.095	0.474	0.159	748 (611r, 137n, 0g)
fullsetL10	685592	**1.000**	0.030	0.059	2720 (1294r, 1426n, 0g)
seeds-rel	80	0.003	**0.775**	0.006	80 (62r, 18n, 0g)
seeds-non	492	0.000	0.000	0.000	2 (0r, 2n, 0g)

Table 16: Set-based Scores for Topic 207 (20929.2 Est. Relevant Documents)

Run	P@B	R@B	F_1@R	indAP	GS10J	First 10 Ret
otL10rvlT	**0.564**	**0.194**	**0.701**	**0.894**	**1.000**	RRRRRRRRRR
otL10FT	0.479	0.092	0.161	0.722	**1.000**	RNRNRRRRRR
otL10bT	0.529	0.150	0.154	0.729	**1.000**	RRRRRRRRRR

Table 17: Rank-based Scores for Topic 207 (B=5572, R=20930)

Run	Avg. K	R@K	P@K	F_1@K	Avg. Num. Judged@K
otL10FT	14018	0.306	0.197	**0.183**	854 (512r, 342n, 0g)
otL10bT	17906	0.423	0.210	0.174	888 (491r, 397n, 0g)
otL10rvlT	23502	0.330	0.186	0.113	823 (387r, 437n, 0g)
seeds-rel	239	0.033	**0.616**	0.059	135 (103r, 32n, 0g)
fullsetL10	685592	**1.000**	0.020	0.038	2720 (674r, 2046n, 0g)
seeds-non	730	0.003	0.037	0.006	12 (3r, 9n, 0g)

Table 18: Mean Set-based Scores of Experimental Learning Task Runs (Avg. 13745.2 Est. Relevant Documents)

Run	P@B	R@B	F_1@R	indAP	GS10J	S1J
otL10FT	0.144	0.544	**0.243**	**0.616**	**0.964**	**6/8**
otL10bT	**0.148**	**0.609**	0.213	0.588	0.919	**6/8**
otL10rvlT	0.145	0.541	0.206	0.451	0.768	4/8

Table 19: Mean Rank-based Scores of Experimental Learning Task Runs

Table 20: Impact of Experimental Techniques on Various Measures

Expt	ΔF_1@K	95% Conf	vs.	3 Extreme Diffs (Topic)
F-b	0.009	(−0.042, 0.061)	6-2-0	−0.12 (200), −0.08 (201), 0.07 (203)
F-rvl	0.070	(0.009, 0.132)	7-1-0	0.19 (202), 0.18 (203), −0.05 (207)
b-rvl	0.061	(0.001, 0.120)	5-3-0	0.19 (200), 0.12 (202), −0.06 (207)
	ΔF_1@R			3 Extreme Diffs (Topic)
F-b	0.031	(−0.017, 0.078)	7-1-0	0.14 (206), 0.07 (202), −0.10 (201)
F-rvl	0.037	(−0.137, 0.212)	7-1-0	−0.54 (207), 0.19 (203), 0.25 (202)
b-rvl	0.007	(−0.161, 0.174)	5-3-0	−0.55 (207), 0.15 (203), 0.18 (202)
	ΔP@B			3 Extreme Diffs (Topic)
F-b	−0.003	(−0.026, 0.020)	4-4-0	0.05 (205), −0.03 (200), −0.05 (207)
F-rvl	−0.000	(−0.028, 0.027)	4-4-0	−0.08 (207), 0.03 (203), 0.04 (200)
b-rvl	0.003	(−0.021, 0.027)	4-4-0	0.07 (200), −0.03 (205), −0.03 (207)
	ΔR@B			3 Extreme Diffs (Topic)
F-b	−0.065	(−0.207, 0.078)	4-4-0	−0.40 (201), −0.31 (206), 0.19 (204)
F-rvl	0.003	(−0.157, 0.163)	4-4-0	−0.33 (206), 0.29 (203), 0.30 (200)
b-rvl	0.068	(−0.092, 0.227)	4-4-0	0.41 (200), 0.26 (203), −0.32 (204)
	ΔindAP			3 Extreme Diffs (Topic)
F-b	0.028	(−0.038, 0.094)	4-4-0	0.18 (200), 0.16 (202), −0.08 (201)
F-rvl	0.165	(0.031, 0.299)	7-1-0	0.43 (200), 0.36 (202), −0.17 (207)
b-rvl	0.137	(0.038, 0.236)	7-1-0	0.26 (200), 0.24 (201), −0.17 (207)
	ΔGS10J			3 Extreme Diffs (Topic)
F-b	0.045	(−0.110, 0.200)	1-2-5	0.57 (202), −0.07 (203), −0.14 (206)
F-rvl	0.196	(0.008, 0.383)	4-0-4	0.59 (206), 0.59 (203), 0.00 (201)
b-rvl	0.151	(−0.144, 0.446)	4-1-3	0.73 (206), 0.66 (203), −0.57 (202)

We also include a comparision table (Table 20) to highlight the differences between the submitted runs in various measures. Its columns are as follows:

- "Expt" specifies the experiment (the codes of the two runs being compared are listed, indicating first run minus second run).
- "Δ" is the difference of the mean scores of the two runs being compared (the column heading says for which retrieval measure).
- "95% Conf" is an approximate 95% confidence interval for the mean difference (calculated from plus/minus twice the standard error of the mean difference; strictly speaking, for 8 topics, the multiplier should be greater than 2.0, but we did not update our scripts for this paper). If zero is not in the interval, the result is "statistically significant" (at the 5% level), i.e. the feature is unlikely to be of neutral impact (on average), though if the average difference is small (e.g. <0.020) it may still be too minor to be considered "significant" in the magnitude sense.
- "vs." is the number of topics on which the first run scored higher, lower and tied (respectively) compared to the second run. These numbers should always add to the number of topics.
- "3 Extreme Diffs (Topic)" lists 3 of the individual topic differences, each followed by the topic number in brackets. The first difference is the largest one of any topic (based on the absolute value). The third difference is the largest difference in the other direction (so the first and third differences give the *range* of differences observed in this experiment). The middle difference is the largest of the remaining differences (based on the absolute value).

From the tables, we see that the experimental feedback-based approach (otL10FT) had the highest score in the rank-based F_1@R measure and set-based F_1@K measure for a majority of the test topics. Of course, the experiments described in this paper just scratch the surface of the research that should be possible with the standard set of test topics, documents and relevance judgments that have been created by the collaborative efforts of the task organizers and participants.

4 Assessor Consistency Study

The set-based tables in the previous section included "seeds-rel" and "seeds-non" entries which show the precision of the training example relevant and non-relevant documents respectively for each topic. On average, just 62% of the example relevant documents were judged relevant by this year's assessors (based on the estimated precision of "seeds-rel" for each topic, weighting each topic equally, as per Table 18). 4% of the example non-relevant documents were judged relevant by this year's assessors (based on the estimated precision of "seeds-non" for each topic, weighting each topic equally, as per Table 18). These consistency results are not much different than expectations, except perhaps for Topic 206, where just 2 of the 18 examples of relevant documents were considered relevant by this year's assessors (as per Table 14).

5 Conclusions

The Learning Task of the TREC 2010 Legal Track investigated the effectiveness of e-Discovery search techniques at learning from examples to estimate the probability of relevance of every document in a collection. The task specified 8 test topics, each of which included a one-sentence request for documents to produce and several examples of relevant and non-relevant items from a new target collection of 685,592 e-mail messages and attachments. For our participation, we produced three retrieval sets to compare experimental feedback-based, topic-based and Boolean-based techniques. In this paper, we described the experimental approaches and reported the scores that each achieved on various set-based and rank-based measures. We reported not just the mean scores of the experimental approaches but also the scores on each of the 8 individual test topics and the largest per-topic impacts of the techniques for several measures. Of the three experimental approaches compared, the experimental feedback-based approach had the highest score in the rank-based F_1@R measure and set-based F_1@K measure for a majority of the test topics. The test data gathered should enable us to further study ranking and probability estimation techniques.

References

[1] Jason R. Baron (Editor-in-Chief). The Sedona Conference® Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery. The Sedona Conference Journal, Volume VIII, pp. 189-223, 2007.

[2] Jason R. Baron. Toward A Federal Benchmarking Standard for Evaluating Information Retrieval Products Used in E-Discovery. The Sedona Conference Journal, Volume VI, pp. 237-246, 2005.

[3] Jason R. Baron, David D. Lewis and Douglas W. Oard. TREC-2006 Legal Track Overview. Proceedings of TREC 2006.

[4] Gordon V. Cormack, Maura R. Grossman, Bruce Hedin and Douglas W. Oard. Overview of the TREC 2010 Legal Track. (To appear in) Proceedings of TREC 2010.

[5] Cross-Language Evaluation Forum web site. `http://www.clef-campaign.org/`

[6] Bruce Hedin, Stephen Tomlinson, Jason R. Baron and Douglas W. Oard. Overview of the TREC 2009 Legal Track. Proceedings of TREC 2009.

[7] Andrew Hodgson. Converting the Fulcrum Search Engine to Unicode. Sixteenth International Unicode Conference, 2000.

[8] D. Lewis, G. Agam, S. Argamon, O. Frieder, D. Grossman, J. Heard. Building a Test Collection for Complex Document Information Processing. *SIGIR 2006*, pp. 665-666.

[9] NTCIR (NII-Test Collection for IR) Home Page. `http://research.nii.ac.jp/~ntcadm/index-en.html`

[10] Douglas W. Oard, Bruce Hedin, Stephen Tomlinson and Jason R. Baron. Overview of the TREC 2008 Legal Track. Proceedings of TREC 2008.

[11] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, M. Gatford. Okapi at TREC-3. Proceedings of TREC-3, 1995.

[12] Text REtrieval Conference (TREC) Home Page. `http://trec.nist.gov/`

[13] Stephen Tomlinson. Early Precision Measures: Implications from the Downside of Blind Feedback. *SIGIR 2006*, pp. 705-706.

[14] Stephen Tomlinson. European Ad Hoc Retrieval Experiments with Hummingbird SearchServer™ at CLEF 2005. Working Notes for the CLEF 2005 Workshop.

[15] Stephen Tomlinson. Experiments with the Negotiated Boolean Queries of the TREC 2006 Legal Discovery Track. Proceedings of TREC 2006.

[16] Stephen Tomlinson. Experiments with the Negotiated Boolean Queries of the TREC 2007 Legal Discovery Track. Proceedings of TREC 2007.

[17] Stephen Tomlinson. Experiments with the Negotiated Boolean Queries of the TREC 2008 Legal Track. Proceedings of TREC 2008.

[18] Stephen Tomlinson. Experiments with the Negotiated Boolean Queries of the TREC 2009 Legal Track. Proceedings of TREC 2009.

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

[20] Stephen Tomlinson. Robust, Web and Terabyte Retrieval with Hummingbird SearchServer™ at TREC 2004. Proceedings of TREC 2004.

[21] TREC 2010 Legal Track – Learning Task. Task Coordinators: Gordon V. Cormack and Maura R. Grossman. `http://plg.uwaterloo.ca/~gvcormac/legal10/`