Learning Task Experiments in the TREC 2011 Legal Track

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

January 25, 2012

Abstract

The Learning Task of the TREC 2011 Legal Track investigated the effectiveness of e-Discovery search techniques at selecting training examples and learning from them to estimate the probability of relevance of every document in a collection. The task specified 3 test topics, each of which included a one-sentence request for documents to produce from a target collection of 685,592 e-mail messages and attachments. In this paper, we describe the experimental approaches used and report the scores that each achieved.

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 [4] and NTCIR [8]) 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 2011 Legal Track.

2 Learning Task

The Learning Task of the TREC 2011 Legal Track investigated the effectiveness of e-Discovery search techniques at selecting training examples and learning from them to estimate the probability of relevance of every document in a collection.

This is the sixth year of the TREC Legal Track and second year of the Learning Task. We have participated in the 6 years of the Legal Track to date (2006-2011). (We also helped with coordinating the Legal Track in 3 of these years (2007-2009) as described in [21], [10] and [6]; however, we were not part of the coordination of this year's track.)

The Learning Task used the same document collection as last year (described below). Like last year, this task requires the system to estimate of the probability of relevance of each e-mail or attachment for each

[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.

test topic. New this year was the opportunity for the system to specify in advance a few hundred documents (in batches of 100) and be given the relevance assessments for those documents (last year the set of example documents was the same for all groups).

The document collection 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.

To test the systems, there were 3 production requests, herein called "topics", numbered 401 to 403. Each topic included a one-sentence request for documents to produce for each topic. A topic authority (assessor) also produced a "coding manual" for each topic with more details of what was considered relevant or not.

Please see the task guidelines [22] for more details on the task and track. [9], [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], [19].

2.1 Indexing

To index the collection, we processed the 685,592 .txt files in the same way as last year, as follows:

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 in 2006-2009. As in those 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. 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 Baseline Runs (No Training Examples)

Participants could submit up to 3 baseline runs, also known as the "first interim submissions", which did not make use of any example judgments from the topic authority. Our 3 baseline runs are described in the next 3 sections.

2.2.1 Boolean Run - otL11BT1

The submitted experimental otL11BT1 run was a Boolean-based run.

We created the following short Boolean queries (1 or 2 words or phrases) based on reading the topic statements and coding manuals:

```
401: FT_TEXT CONTAINS 'enrononline%'
402: FT_TEXT CONTAINS 'over-the-counter'|'OTC'
403: FT_TEXT CONTAINS 'environmental'|'environment%'
```

Unlike last year, no linguistic expansion from English inflectional stemming was applied for our Boolean queries. The "%" character however indicated wildcard expansion, e.g. 'environment%' would also match 'environmentalist'. In the topic 403 query, the function of separately listing 'environmental' (which would already be matched by 'environment%') was to increase the weight on that particular term for relevance calculations. The hyphenated 'over-the-counter' phrase would also match non-hyphenated variations such as 'over the counter'.

It was hoped that these simple queries would have good recall of the relevant documents. With the final judgments we can now see that the first one had 0.59 precision but just 0.21 recall, the second one had just 0.18 precision and 0.12 recall, and the third one had 0.04 precision and 0.60 recall, as listed in the P@B and R@B columns of Tables 4-6. (The B values, i.e. the number matches for the Boolean queries, were 8241, 5085 and 19435 for the 3 topics respectively.)

The matches were still relevance-ranked. (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 wildcard terms (e.g. "televis%"), all variants (e.g. "television", "televised", "televisions", etc.) were treated as occurrences of the same term for term frequency purposes, and inverse document frequency was based on the most common variant. For runs which used inflectional matching (which was the case for the Request and Feedback runs described later), these calculations were based on the stems of the terms.) For terms in phrases of Boolean queries, only occurrences of the term satisfying the phrase counted towards term frequency.)

To assign a probability of relevance to each matching document, we used the same experimental "general-purpose" probability formula as last year, which was to take 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 probability set to 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 401, the Boolean query had 8241 matches, but the sum of the probabilities from the (experimental) general-purpose formula was just 7319.

This experimental otL11BT1 run was submitted August 1, 2011.

2.2.2 Request Run - otL11FT1

The submitted experimental otL11FT1 run was just based on the terms in the one-sentence request (topic statement) after manually removing what seemed to be common instruction words (e.g. "please", "produce", "documents"). The resulting queries were as follows:

```
401: FT_TEXT CONTAINS 'design'|'development'|'operation'|'marketing'|'enrononline'|
     'online'|'service'|'offered'|'provided'|'used'|'subsidiaries'|'predecessors'|
     'successors'|'interest'|'purchase'|'sale'|'trading'|'exchange'|'financial'|
     'instruments'|'products'|'limited'|'derivative'|'instruments'|'commodities'|
     'futures'|'swaps'
402: FT_TEXT CONTAINS 'purchase'|'sale'|'trading'|'exchange'|'counter'|'derivatives'|
     'actual'|'contemplated'|'financial'|'instruments'|'products'|'legal'|'illegal'|
```

```
     'permitted'|'prohibited'|'proposed'|'rule'|'regulation'|'law'|'standard'|
     'proscription'|'domestic'|'foreign'
403: FT_TEXT CONTAINS 'environmental'|'impact'|'activity'|'activities'|'undertaken'|
     'limited'|'measures'|'taken'|'conform'|'comply'|'avoid'|'circumvent'|'influence'|
     'proposed'|'rule'|'regulation'|'law'|'standard'|'proscription'|'governing'|
     'environmental'|'emmissions'|'spills'|'pollution'|'noise'|'animal'|'habitats'
```

Note that linguistic expansion from English inflectional stemming was also applied for these queries, e.g. the search for 'design' also matched 'designs'.

The probability formula was the same as for the short Boolean query run, i.e. take the raw relevance() score (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. Non-matches were all assigned 0.0001.

With the final judgments, we can get an idea of how these queries fared compared to the short Boolean queries by cutting off the ranked list at the same number of matches as the corresponding Boolean query ("depth B"). As listed in the P@B and R@B columns of Tables 4-6, we see that the request-based query had a lower precision and recall at depth B for the first and third queries, but a little higher for the second query.

2.2.3 Fusion Run - otL11HT1

The submitted experimental otL11HT1 run was a fusion run which just assigned the probability to each document by summing half of the probability assigned in the otL11BT1 run and half of the probability assigned in the otL11FT1 run.

2.3 Selection of 100 Training Examples

Unlike last year, this year we could choose our own training example documents for each topic (up to 1000 per topic, but only 100 at a time).

Our approach was to start by drawing 100 samples from the Boolean run otL11BT1 by selecting the documents of the following ranks for each topic:

```
1, 2, 3, 4, 5, 6, 7, 8, 9, 10,
11, 12, 13, 14, 15, 16, 17, 18,
20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100,
150, 200, 250, 300, 350, 400, 450, 500, 550,
600, 650, 700, 750, 800, 850, 900, 950, 1000,
1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000, 5500,
6000, 6500, 7000, 7500, 8000, 8500, 9000, 9500, 10000,
15000, 20000, 25000, 30000, 35000, 40000, 45000, 50000, 55000,
60000, 65000, 70000, 75000, 80000, 85000, 90000, 95000, 100000,
150000, 200000, 250000, 300000, 350000, 400000, 450000, 500000, 550000, 600000, 650000
```

We favored documents of earlier ranks in hopes of getting a lot of example relevant documents, but we also sampled from the rest of the run (which included the entire target document set) so that we could compute an estimate of the number of relevant documents in the document set for each topic (albeit a coarse estimate) which would be useful for estimating the probabilities of relevance.

We used the Boolean run as our base instead of the Request run because the short Boolean queries tended to retrieve short documents at the top of the results, whereas the long queries of the Request run tended to retrieve a lot of longer documents at the top of the list. We suspect that our learning method works better with shorter training documents.

On August 2, 2011, we submitted our request for the judgments for these 100 documents for each topic. We received the 100 relevance assessments for topic 402 on August 4, for topic 403 on August 5, and for topic 401 on August 8.

Of the 100 documents for each topic, it turned out that 58 were judged relevant for topic 401, 23 were judged relevant for topic 402, and 18 were judged relevant for topic 403. Hence we now had some example relevant documents to work with for each topic.

We converted the 100 examples for each topic into a "qrels" file compatible with the l07_eval utility. For our Boolean run (otL11BT1), it estimated the following scores:

```
otL11BT1.eval::K:              401 8241.0000
otL11BT1.eval::K:              402 5085.0000
otL11BT1.eval::K:              403 19435.0000

otL11BT1.eval::est_K-Prec:     401 0.5561
otL11BT1.eval::est_K-Prec:     402 0.2460
otL11BT1.eval::est_K-Prec:     403 0.0890

otL11BT1.eval::est_K-Recall:   401 1.0000
otL11BT1.eval::est_K-Recall:   402 1.0000
otL11BT1.eval::est_K-Recall:   403 1.0000

otL11BT1.eval::est_K-F1:       401 0.7148
otL11BT1.eval::est_K-F1:       402 0.3949
otL11BT1.eval::est_K-F1:       403 0.1635
```

All of all example relevant documents came from the documents that matched the Boolean query (leading to the overly high estimates of 100% recall) presumably because the set was in arbitrary order after depth B. More thought should be put into the sampling approach in the future.

Our estimates of the number of relevant documents for each topic were also off compared to the full official judgments, especially for the first topic:

```
our 100::est_rel:       401 4449.0000
our 100::est_rel:       402 1230.0000
our 100::est_rel:       403 1335.0000

prelim: :est_rel:       401 30852.8914
prelim: :est_rel:       402 1920.0998
prelim: :est_rel:       403 1238.7119

final: :est_rel:       401 20016.8646
final: :est_rel:       402 3012.1996
final: :est_rel:       403 1238.7119
```

2.4 Runs Using 100 Training Examples

The "second interim submissions" incorporated learning from the 100 example judgments for each topic.

2.4.1 Feedback Run using 100 Examples - otL11FT2

The submitted experimental otL11FT2 run was a pure feedback run that did not make any use of the topic statements. The input was just the example relevant documents, which were the 100 example judged documents after discarding non-relevant documents and discarding documents of 10,000 bytes or more (in the XML formatting described earlier) in hopes of reducing the percentage of input text that was not relevant.

These example relevant documents for each topic were 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.

Our probability formula was updated to take advantage of our estimates of the number of relevant documents for each topic. The updated formula was to take the relevance() score (usually between 0 and 500), multiply by 0.002, apply the exponent x, divide by 0.98, and enforce a max of 0.98 and min of 0.0001. Documents not matched by the feedback query were also assigned 0.0001. The known relevant documents were moved to the front and assigned 0.99, while the known non-relevant documents were assigned 0.01. Exponent x was 2.237, 3.339, 3.301 for the 3 topics respectively, chosen to make the probabilities sum to the estimated number of relevant documents based on our sample of 100 documents per topic.

The reason for assigning 0.99 to the known relevant documents instead of 1.00 (and 0.01 for known non-relevant documents instead of 0.00) was that we were advised that it was possible the assessors might change their assessment in the final judgments (e.g. to fix an error).

This experimental otL11FT2 run was submitted August 27, 2011. At this point, we were eligible again to request another 100 judgments for each topic, but the deadline for runs using up to 1000 example judgments was August 28, and even if the judgments came back instantaneously we wouldn't have had time to incorporate them.

2.4.2 Boolean Run using 100 Examples - otL11BT2

The submitted experimental otL11BT2 run was the same as the baseline otL11BT1 run except that the known relevant documents were moved to the front and the known non-relevant documents were moved down (from our example 100 judgments per topic). The probabilities were assigned using the same approach as for otL11FT2 except that the exponents worked out to 3.14, 1.61 and 8.54 for the 3 topics respectively.

2.4.3 Fusion Run using 100 Examples - otL11HT2

The submitted experimental otL11HT2 run was a fusion run which just assigned the probability to each document by summing half of the probability assigned in the otL11BT2 run and half of the probability assigned in the otL11FT2 run.

2.5 Mop-up Runs

On August 30, 2011, the track organizers released all of the example judgments for each topic that had been requested by any participating group. "Mop-up" runs using all of these examples were due September 6.

Here are the counts of the number of example judgments, judged relevant and judged non-relevant for each topic in the mop-up examples:

```
Topic 401: count=2500, rel=1040, non=1460
Topic 402: count=2102, rel= 238, non=1864
Topic 403: count=2199, rel= 245, non=1954
```

The extra examples weren't of much help for improving our estimates of the number of relevant documents for each topic because we did not have information on how these examples were sampled, so we continued to target the same numbers of relevant documents as we did with just 100 training examples per topic.

2.5.1 Feedback Run using 2000+ Mop-up Examples - otL11FTM

The submitted experimental otL11FTM run was produced in the same way as the otL11FT2 run except that we had a lot more example relevant documents to start with. The exponents in the probability formulas worked out to 2.72, 3.92, 3.48 for the 3 topics respectively.

2.5.2 Boolean Run using 2000+ Mop-up Examples - otL11BTM

The submitted experimental otL11BTM run was produced in the same way as the otL11BT2 run except that we had a lot more example relevant documents to move to the front. The exponents in the probability formulas worked out to 3.86, 1.83 and 9.07 for the 3 topics respectively.

2.5.3 Fusion Run using 2000+ Mop-up Examples - otL11HTM

The submitted experimental otL11HTM run was a fusion run which just assigned the probability to each document by summing half of the probability assigned in the otL11BTM run and half of the probability assigned in the otL11FTM run.

2.6 Preliminary and Final Judgments

On October 16, 2011, preliminary relevance assessments were released. Here are the counts of the number of judged documents, relevant documents and non-relevant documents for each topic in these preliminary assessments:

```
Topic 401: count=5871, rel=2585, non=3286
Topic 402: count=5583, rel= 843, non=4740
Topic 403: count=5545, rel= 534, non=5011
```

Here are the estimated number of relevant documents for each topic based on the preliminary judgments:

```
prelim: :est_rel:       401 30852.8914
prelim: :est_rel:       402 1920.0998
prelim: :est_rel:       403 1238.7119
```

At about the time of the conference (November 15-18, 2011), the final relevance assessments were released. Here are the counts of the number of judged documents, relevant documents and non-relevant documents for each topic in the final assessments:

```
Topic 401: count=5871, rel=2621, non=3250
Topic 402: count=5583, rel= 858, non=4725
Topic 403: count=5545, rel= 534, non=5011
```

As you can see, some of the judgments changed for the first two topics. The third topic did not have any changes between the preliminary and final judgments.

Here are the estimated number of relevant documents for each topic based on the final judgments:

```
final: :est_rel:       401 20016.8646
final: :est_rel:       402 3012.1996
final: :est_rel:       403 1238.7119
```

Run	K	Recall	Precision	F_1	Num. Judged
otL11FT2	7624	0.241	0.675	**0.355**	1919 (1771r, 148n, 0g)
otL11HT1	9422	0.206	0.550	0.300	2289 (1955r, 334n, 0g)
otL11BT1	7502	0.178	0.601	0.275	2016 (1825r, 191n, 0g)
otL11HTM	6176	0.138	0.605	0.225	1963 (1898r, 65n, 0g)
otL11BTM	6652	0.139	0.554	0.223	1989 (1920r, 69n, 0g)
otL11HT2	6857	0.131	0.531	0.211	1910 (1759r, 151n, 0g)
otL11BT2	6792	0.131	0.529	0.209	1893 (1742r, 151n, 0g)
otL11FT1	38669	0.284	0.161	0.206	1739 (1333r, 406n, 0g)
otL11FTM	1644	0.067	**0.968**	0.124	1374 (1330r, 44n, 0g)
mopup-rels	1040	0.050	0.961	0.095	1040 (999r, 41n, 0g)
fullsetL11	685592	**1.000**	0.029	0.057	5871 (2621r, 3250n, 0g)
mopup-nons	1460	0.005	0.070	0.009	1460 (102r, 1358n, 0g)

Table 1: Set-based Scores for Topic 401 (20016.9 Est. Relevant Documents)

Run	K	Recall	Precision	F_1	Num. Judged
otL11FT2	1770	0.158	0.506	**0.241**	584 (477r, 107n, 0g)
otL11HTM	628	0.115	0.309	0.168	401 (346r, 55n, 0g)
otL11BTM	1740	0.121	0.163	0.139	443 (364r, 79n, 0g)
otL11FTM	242	0.074	**0.926**	0.138	242 (224r, 18n, 0g)
mopup-rels	238	0.073	0.924	0.135	238 (220r, 18n, 0g)
otL11HT2	2707	0.099	0.101	0.100	454 (299r, 155n, 0g)
otL11BT2	2769	0.091	0.094	0.093	417 (275r, 142n, 0g)
otL11BT1	2507	0.090	0.092	0.091	429 (270r, 159n, 0g)
otL11FT1	23289	0.319	0.044	0.078	1276 (601r, 675n, 0g)
otL11HT1	15286	0.186	0.035	0.059	1167 (559r, 608n, 0g)
mopup-nons	1864	0.014	0.023	0.018	1864 (43r, 1821n, 0g)
fullsetL11	685592	**1.000**	0.004	0.009	5583 (858r, 4725n, 0g)

Table 2: Set-based Scores for Topic 402 (3012.2 Est. Relevant Documents)

Run	K	Recall	Precision	F_1	Num. Judged
otL11HTM	923	0.270	0.689	**0.389**	486 (335r, 151n, 0g)
otL11FTM	304	0.215	0.905	0.347	294 (266r, 28n, 0g)
mopup-rels	245	0.195	**0.984**	0.325	245 (241r, 4n, 0g)
otL11BTM	1808	0.281	0.381	0.323	561 (348r, 213n, 0g)
otL11FT2	3236	0.519	0.149	0.231	807 (291r, 516n, 0g)
otL11HT2	2860	0.220	0.123	0.158	815 (273r, 542n, 0g)
otL11BT2	3017	0.208	0.101	0.136	795 (258r, 537n, 0g)
otL11BT1	12739	0.585	0.057	0.104	1388 (372r, 1016n, 0g)
otL11HT1	11370	0.586	0.054	0.100	1368 (373r, 995n, 0g)
otL11FT1	14442	0.533	0.040	0.074	1078 (308r, 770n, 0g)
mopup-nons	1954	0.019	0.012	0.015	1954 (24r, 1930n, 0g)
fullsetL11	685592	**1.000**	0.002	0.004	5545 (534r, 5011n, 0g)

Table 3: Set-based Scores for Topic 403 (1238.7 Est. Relevant Documents)

Run	Retrieved	P@B	R@B	F_1@R	R@ret	indAP	GS10J	First 10 Ret
otL11HTM	685592	0.651	0.212	**0.383**	1.000	0.940	1.000	RRRRRNRRRR
otL11FTM	685592	**0.691**	**0.255**	0.371	1.000	**0.946**	1.000	RRRRRRRRRR
otL11HT2	685592	0.591	0.205	0.306	1.000	0.916	1.000	RRRRRRRRRR
otL11FT2	685592	0.675	0.242	0.306	1.000	0.916	1.000	RRRRRRRRRR
otL11HT1	685592	0.550	0.204	0.273	1.000	0.880	1.000	RRRRRRRRRR
otL11BTM	685592	0.610	0.213	0.246	1.000	0.908	1.000	RRRRRRRRRR
otL11BT1	685592	0.593	0.206	0.239	1.000	0.869	1.000	RRRRRRRRRR
otL11BT2	685592	0.593	0.206	0.239	1.000	0.870	1.000	RRRRRRRRRR
otL11FT1	685592	0.392	0.120	0.196	1.000	0.713	0.926	NRRRRNNRRR

Table 4: Rank-based Scores for Topic 401 (B=8241, R=20017)

Run	Retrieved	P@B	R@B	F_1@R	R@ret	indAP	GS10J	First 10 Ret
otL11FTM	685592	**0.231**	**0.318**	**0.210**	1.000	**0.780**	1.000	RRRRRRRRRR
otL11FT2	685592	0.217	0.297	0.194	1.000	0.731	1.000	RRRRRRRRRR
otL11HT1	685592	0.127	0.156	0.156	1.000	0.604	1.000	RRRRRRRRRR
otL11HTM	685592	0.121	0.175	0.137	1.000	0.702	1.000	RRRRRRRRRR
otL11BTM	685592	0.094	0.144	0.134	1.000	0.609	1.000	RRRRRRRRRR
otL11FT1	685592	0.097	0.123	0.113	1.000	0.597	1.000	RRRRRRRRRR
otL11HT2	685592	0.106	0.153	0.101	1.000	0.615	1.000	RRRRRRRRRR
otL11BT1	685592	0.075	0.115	0.098	1.000	0.419	1.000	RRNRRRRRNR
otL11BT2	685592	0.074	0.113	0.097	1.000	0.435	1.000	RRRRRRRRRR

Table 5: Rank-based Scores for Topic 402 (B=5085, R=3013)

Run	Retrieved	P@B	R@B	F_1@R	R@ret	indAP	GS10J	First 10 Ret
otL11BTM	685592	0.041	0.638	**0.384**	1.000	0.729	1.000	RRRRRRRRRR
otL11FTM	685592	**0.044**	0.674	0.347	1.000	**0.796**	1.000	RRRRRRRRRR
otL11HTM	685592	0.042	**0.684**	0.329	1.000	0.760	1.000	RRRRRRRRRR
otL11BT2	685592	0.038	0.595	0.222	1.000	0.345	1.000	RRRRNRRRRR
otL11BT1	685592	0.038	0.595	0.214	1.000	0.286	0.681	NNNNNRRNNR
otL11FT2	685592	0.038	0.625	0.205	1.000	0.419	1.000	RRRRRRRRRR
otL11HT2	685592	0.041	0.628	0.190	1.000	0.389	0.926	NRRRRRRRRR
otL11HT1	685592	0.037	0.600	0.149	1.000	0.295	0.857	NNRRNNNNRN
otL11FT1	685592	0.034	0.558	0.100	1.000	0.269	0.857	NNRRNNNNRN

Table 6: Rank-based Scores for Topic 403 (B=19435, R=1239)

3 Results

After all the participants submitted their experimental runs (due September 6, 2011), 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 [5], but our understanding is that it proceeded as follows.

The test collection was divided into 2 strata, called stratum 100 and stratum 1000. 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 almost completely judged. The remaining documents were in the 'stratum 1000' which was uniformly sampled (e.g. approx 1 in every 353 documents was sampled for topic 403, as per below).

The task organizers produced a preliminary set of judgments (qrels.t11legallearn.prelim) on October 16, 2011, in time for the October 24 notebook paper deadline. The final set of judgments (qrels.t11legallearn) were released during the conference (November 15-18). In this paper, we just use the final set of judgments.

Based on the final judgments in qrels.t11legallearn, 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 401:
 stratum 100: count=4309, judged=4308, prob=0.999767927593
 stratum 1000: count=681283, judged=1563, prob=0.002294200795
Topic 402:
 stratum 100: count=3689, judged=3689, prob=1.000000000000
 stratum 1000: count=681903, judged=1894, prob=0.002777521143
Topic 403:
 stratum 100: count=3615, judged=3615, prob=1.000000000000
 stratum 1000: count=681977, judged=1930, prob=0.002830007464
```

The counts for each topic should sum to 685,592 (the number of documents in the collection). The number of judged documents was between 5500 and 6000 for each topic.

3.1 Score Coarseness Issue

The aforementioned sampling into 2 strata led to a coarseness issue with the scores. For example, for the otL11FTM run, if you cutoff the ranked list for topic 403 at 1341 items, the estimated recall is 30%, the estimated precision is 41%, and the estimated F_1 is 35%. But if you cutoff the ranked list just one document later, i.e. at 1342 items, the estimated recall jumps to 59%, the estimated precision jumps to 58%, and the estimated F_1 jumps to 58%.

Why does retrieving just one more document make such a difference to the estimated scores? For topic 403, as shown in the previous section, the sampling rate of the 2nd stratum was approximately 1 in 353.4 (the number 353.4 comes from dividing the listed number of documents in stratum, 681977, by the listed number of judged documents in the stratum, 1930). Of the 534 documents judged relevant for topic 403, just 2 of them were in the 2nd stratum (the other 532 were in the 1st stratum, which was fully judged). The total estimated number of relevant documents for topic 403 is 1239 (which comes from 532 plus 2 times 353.4). In the case of run otL11FTM, the document ranked at position 1342 for topic 403 was one of the relevant documents from that 2nd stratum, hence including it added approximately 353.4 to the estimated number of relevant documents retrieved, which compared to the total estimate of 1239 relevant documents leads to the aforementioned jump in recall. (At rank 1341, the estimated recall comes from 376/1239 and the estimated precision comes from 376/(376+532). At rank 1342, the estimated recall comes from 729/1239 and the estimated precision comes from 729/(729+532).)

Even without increasing the number of judgments, the coarseness issue could be reduced a lot with a different sampling strategy, like the one used in the 2009 Batch Task [6], for which (simplifying a bit) the sampling rate was p(d)=C/hiRank(d), where hiRank(d) was the highest rank at which any submission ranked

document d (where 1 is the highest possible rank), and C was chosen so that the probabilities summed to the number of judgments that could be done. With this approach, the impact of 1 document at rank 1342 on recall would probably have been less than 0.10, instead of 0.29.

With a small number of strata, it's hard to not have dry spots where the coarseness is high (e.g. for the topic 403 sampling, the driest spot would be documents of hiRank just past 100 where the weight of 1 document might be 353 which swamps all the previous judgments; by rank 10000 the 1/353 sampling rate is probably fine.) Using the non-uniform sampling formula p(d)=C/hiRank(d) is an easy way to treat all potential cutoffs equally.

3.2 L07 vs. L10 measures

In 2010, 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 qrelsL11.probs file (for use with the l07_eval scoring utility) by taking the judged documents from qrels.t11legallearn and assigning them the probability as listed in the previous section. Last year we reported some example comparisons of F_1@K scores from the two approaches and found that they were almost the same, suggesting that both estimation approaches are likely to lead to similar conclusions [19].

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 [21], 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` .

In this paper, we actually used the following simpler versions of the L07 estimators, which are analyzed in more detail in [20]:

```
estRel@k = sum(d is Rel)(1/p(d))
estNon@k = sum(d is Non)(1/p(d))
recall@k = estRel@k / estRel@D
prec@k   = estRel@k / (estRel@k + estNon@k)
```

To summarize, the L07 estimators have the following advantages over the L10 estimators:

- recall is non-decreasing as more results are retrieved (monotonicity property)
- recall is always in the 0..1 range
- the qrels only need to list the few thousand judged documents, not all 680,000+ documents
- they work naturally with non-uniform probability formulas such as the p(d) formula recommended earlier (no need to define strata).

3.3 L07 measures

For each topic, we report a table of set-based scores (Tables 1-3) and a table of rank-based scores (Tables 4-6). 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 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 9 experimental submissions, but also 3 reference runs as follows:

- "mopup-rels" is the set of example relevant documents (released August 30, 2011) for use in the "Mop-up" runs. Some of the example relevant documents ended up being judged non-relevant in the final judgments, as per Tables 1-3.
- "mopup-nons" is the set of example non-relevant documents (released August 30, 2011) for use in the "Mop-up" runs. Some of the example non-relevant documents ended up being judged relevant in the final judgments, as per Tables 1-3.
- "fullsetL11" is the set consisting of the entire document collection.

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

- "K": For the 9 submitted runs, K came from the organizer-provided "Cutoff Estimate", 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:

- "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 otL11BT1), 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.)
- "R@ret": Estimated Recall of the full retrieval set.
- "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.)

4 Conclusions

The Learning Task of the TREC 2011 Legal Track investigated the effectiveness of e-Discovery search techniques at selecting training examples and learning from them to estimate the probability of relevance of every document in a collection. The task specified 3 test topics, each of which included a one-sentence request for documents to produce from a target collection of 685,592 e-mail messages and attachments. For our participation, we produced nine retrieval sets to compare experimental feedback-based, topic-based and Boolean-based techniques. In this paper, we described the experimental approaches used and reported the scores that each achieved on each topic on various set-based and rank-based measures. Generally speaking, approaches based on relevance feedback were found to outperform the other approaches. We also identified a coarseness issue with the estimated scores and recommended a sampling approach and associated estimators for addressing this issue.

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

[5] Maura R. Grossman, Gordon V. Cormack, Bruce Hedin and Douglas W. Oard. Overview of the TREC 2011 Legal Track. (To appear in) Proceedings of TREC 2011.

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

[9] Douglas W. Oard, Jason R. Baron, Bruce Hedin, David D. Lewis and Stephen Tomlinson. Evaluation of Information Retrieval for E-Discovery. Artificial Intelligence and Law, Volume 18, Number 4, pp. 347-386, 2010.

[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. Learning Task Experiments in the TREC 2010 Legal Track. Proceedings of TREC 2010.

[20] Stephen Tomlinson and Bruce Hedin. Measuring Effectiveness in the TREC Legal Track. In M. Lupu, K. Mayer, J. Tait and A. Trippe (editors), Current Challenges in Patent Information Retrieval. Springer, 2011.

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

[22] TREC 2011 Legal Track – Learning Task Guidelines. `http://http://plg.uwaterloo.ca/~gvcormac/legal11/treclegal11.html`