Experiments with the Negotiated Boolean Queries of the TREC 2008 Legal Track

Stephen Tomlinson
Open Text Corporation
Ottawa, Ontario, Canada
stomlins@opentext.com
http://www.opentext.com/

February 4, 2009

Abstract

We analyze the results of several experimental runs submitted for the TREC 2008 Legal Track. In the Ad Hoc task, we found that rank-based merging of vector results with the reference Boolean results produced a statistically significant increase in mean F_1@K and Recall@B compared to just using the reference Boolean results. In the Relevance Feedback task, we found that the investigated relevance feedback technique, when merged with the reference Boolean results, produced some substantial increases in Recall@B_r without any substantial decreases on individual topics.

1 Introduction

Open Text eDOCS SearchServer^{TM} is a toolkit for developing enterprise search and retrieval applications. The eDOCS SearchServer kernel is also embedded in various components of the Open Text eDOCS Suite[1].

The eDOCS SearchServer kernel works in Unicode internally [5] and supports most of the world's major character sets and languages. The major conferences in text retrieval experimentation (TREC [10], CLEF [4] and NTCIR [7]) 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 Ad Hoc and Relevance Feedback tasks of the TREC 2008 Legal Track.

2 Ad Hoc Task

The Ad Hoc task of the TREC 2008 Legal Track models e-Discovery searching on a fixed document collection using queries (production requests) that the system has not seen before. This is the 3rd year of the Ad Hoc task in the TREC Legal Track (and we have participated all 3 years). (We also have helped with coordinating the task the past couple years as described in [15] and [8]).

As in the past couple years, the document collection to be searched was the IIT Complex Document Information Processing (IIT CDIP) test collection [6]. It contained 6,910,192 metadata records from US tobacco companies; 6,794,895 of the records included document text of varying quality from an optical

[1] 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.

character reader. Uncompressed, the collection was 61,251,357,065 bytes (57.0 GB). The average record size (including metadata markup and the ocr document) was 8864 bytes.

In e-Discovery (also known as eDiscovery, electronic discovery or legal discovery), the goal is to return all documents responsive (relevant) to a production request, without returning any non-responsive documents.

For the Ad Hoc task of the TREC Legal Track, the organizers created 45 new production requests, herein called topics, numbered from 102 to 151. Each topic included a "request text" (a natural language description of the request, typically one-sentence), a "defendant query" (an initial Boolean query proposed by the defendant), a "plaintiff query" (a rejoinder Boolean query from the plaintiff) and a "final negotiated query" (the final Boolean query from the negotiations). (Examples appear below.) The final Boolean query was known to always match between 100 and 100,000 records. During the assessing phase, some topics were dropped, leaving 26 topics.

[16] and [8] have more details on the track and task, and [1] and [2] have more background on e-Discovery in general. Also, background on last year's track is in [14] and [15].

2.1 Indexing

Our index included both the metadata and the ocr document of each record. We indexed from the "</tid>" tag to the "</record>" tag, which meant both the metadata and the ocr document were in the FT_TEXT column. Any tags themselves were indexed (we just didn't bother to discard them). Entities (e.g. "&") were converted to the character they represented (e.g. "&").

We did not use a stopword list, and we also indexed most punctuation as 1-character words (exceptions were the hyphen and apostrophe, which were treated as 1-character stopwords). The contents of the "<dd>" section of the metadata were additionally indexed in a separate DOCDATE column, though this column was not used by the queries 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 Searching

The techniques used for our 8 submitted Ad Hoc runs of August 2008 are described below. The relevance ranking approach was the same for all runs. The relevance function dampened the term frequency and adjusted for document length in a manner similar to Okapi [9] 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, these calculations were based on the stems of the terms. For terms in phrases or proximity contraints of Boolean queries, only occurrences of the term satisfying the phrase or proximity counted towards term frequency.

The 8 submitted Ad Hoc task runs were as follows:

otL08fb (final Boolean run): The submitted otL08fb run used the final negotiated query, respecting the Boolean operators such as AND, phrase, proximity, NOT, etc. Full wildcard matching was supported. Relevance-ranking was still used to order the matching rows. For example, for topic 118, for which the final negotiated query was "`(malfunction! OR breakdown! OR failure! OR fault! OR incident!) w/25 ((manufactur! OR assembl! OR fabricat! OR produc!) OR (test! OR trial! OR exam! OR validat! OR evaluat!))`", the corresponding SearchSQL statement was

```
SELECT RELEVANCE('2:3') AS REL, DOCNO
FROM LEGAL08FULL
WHERE (FT_TEXT CONTAINS 'malfunction%', 'breakdown%', 'failure%', 'fault%', 'incident%'
        within 25 words of  'manufactur%', 'assembl%', 'fabricat%', 'produc%', 'test%',
        'trial%', 'exam%', 'validat%', 'evaluat%')
```

```
ORDER BY REL DESC;
```

otL08db (defendant Boolean run): The submitted otL08db run was the same as otL08fb except that the defendant query was used instead of the final negotiated Boolean query. For example, for topic 118, for which the initial query proposed by the defendant was "`(malfunction w/10 machinery) AND (manufacturing OR testing)`", the WHERE clause of the corresponding SearchSQL statement was

```
WHERE ((FT_TEXT CONTAINS 'malfunction' within 10 words of 'machinery')
       AND (FT_TEXT CONTAINS 'manufacturing', 'testing'))
```

otL08fv (final vector run): The submitted otL08fv run was the same as otL08fb except that the Boolean operators such as AND, phrases and proximities were dropped (all operators became an OR), and punctuation was dropped. Full wildcarding was still respected. For example, for topic 118, the WHERE clause of the corresponding SearchSQL statement was

```
WHERE (FT_TEXT CONTAINS 'malfunction%'|'breakdown%'|'failure%'|'fault%'|
                        'incident%'|'manufactur%'|'assembl%'|'fabricat%'|
                        'produc%'|'test%'|'trial%'|'exam%'|'validat%'|'evaluat%')
```

otL08rvl (request text vector run): The submitted otL08rvl run was the same as otL08fv except that (1) the terms were taken from the request text field instead of the final negotiated query field, (2) linguistic expansion from English inflectional stemming was applied, and (3) common instruction words (e.g. "please", "produce", "documents") were manually removed. For example, for topic 118, for which the request text was "Please produce all reports, written memoranda, correspondence, and other documents related to past incidents involving the malfunction of machinery in connection with manufacturing or testing activities, or which occurred within manufacturing or testing facilities.", the WHERE clause of the corresponding SearchSQL statement was

```
WHERE FT_TEXT CONTAINS 'past'|'incidents'|'involving'|'malfunction'|'machinery'|
         'connection'|'manufacturing'|'testing'|'activities'|'occurred'|
         'manufacturing'|'testing'|'facilities'
```

otL08rvlq (request text vector run, keeping instruction words): The submitted otL08rvlq run was the same as otL08rvl except that common instruction words (e.g. "please", "produce", "documents") were not removed. For example, for topic 118, the WHERE clause of the corresponding SearchSQL statement was

```
WHERE FT_TEXT CONTAINS 'Please'|'produce'|'all'|'reports'|'written'|'memoranda'|
         'correspondence'|'and'|'other'|'documents'|'related'|'to'|'past'|
         'incidents'|'involving'|'the'|'malfunction'|'of'|'machinery'|'in'|
         'connection'|'with'|'manufacturing'|'or'|'testing'|'activities'|'or'|
         'which'|'occurred'|'within'|'manufacturing'|'or'|'testing'|'facilities'
```

otL08rv (uninflected request text vector run): The submitted otL08rv run was the same as otL08rvl except that linguistic expansion from English inflectional stemming was not applied. The WHERE clauses of the SearchSQL statements were the same as for otL08rvl; the difference was just that this run preceded the searches with "`SET TERM_GENERATOR ''`" instead of "`SET TERM_GENERATOR 'word!ftelp/inflect'`".

otL08frw (fusion run): The submitted otL08frw run was a weighted fusion of the final Boolean, request text vector and final vector runs: weight 3 on otL08fb, weight 2 on otL08rvl, weight 1 on otL08fv. Each input run was retrieved to depth 100,000 (or B in the case of the final Boolean run). This year we changed our fusion algorithm to be rank-based rather than rsv-based. Last year we just added together the weighted relevance scores (also known as retrieval status values or rsv scores), which we found last year [14] caused low-scoring Boolean documents to often be omitted despite the 3x weight on them. This year we replaced the rsv score with (7000000 - rank)/7000000 (where rank ranged from 1 to 100,000) before multiplying by the weights and adding the resulting scores. Hence documents in all 3 input runs (fb, rvl, fv (weight sum

Table 1: Mean Scores of Submitted Ad Hoc Task Runs

Run	Avg. K	(P@K, R@K)	F_1@K	F_1@R	GS10J, S1J, P5	R@B, R@ret
otL08fbe	75228	(0.241, **0.409**)	**0.215**	**0.246**	0.842, 17/26, 0.654	0.272, 0.451
otL08frw	64232	(0.239, 0.380)	0.207	0.220	**0.938**, **21/26**, **0.769**	0.269, **0.461**
otL08fv	46369	(0.243, 0.345)	0.190	0.186	0.856, 13/26, 0.485	0.254, 0.447
otL08rv	52081	(0.242, 0.311)	0.185	0.216	0.864, 15/26, 0.592	0.268, 0.422
otL08rvl	81826	(0.190, 0.402)	0.179	0.215	0.856, 16/26, 0.585	**0.278**, 0.443
otL08fb	40402	(0.280, 0.240)	0.161	0.165	0.905, 17/26, 0.577	0.240, 0.240
otL08rvlq	50865	(0.186, 0.253)	0.126	0.160	0.802, 14/26, 0.415	0.217, 0.352
otL08db	3180	(**0.407**, 0.035)	0.050	0.037	0.717, 16/26, 0.469	0.034, 0.035
(high rel)	Avg. K_h	(P@K_h, R@K_h)	F_1@K_h	F_1@R_h	GS10J, S1J, P5	R@B, R@ret
otL08frw	21799	(0.100, 0.316)	**0.095**	0.164	0.656, 8/24, **0.333**	0.385, 0.643
otL08fb	39930	(0.078, 0.335)	0.091	0.104	0.539, 3/24, 0.192	0.335, 0.335
otL08fbe	14124	(0.105, 0.267)	0.086	0.117	0.489, 4/24, 0.267	0.349, 0.531
otL08rv	30039	(0.100, 0.372)	0.079	**0.165**	0.643, **10/24**, 0.300	**0.500**, **0.696**
otL08rvl	43531	(0.059, **0.441**)	0.065	0.150	**0.667**, 7/24, 0.325	0.464, 0.645
otL08fv	15389	(0.076, 0.316)	0.064	0.086	0.435, 3/24, 0.108	0.354, 0.519
otL08db	3445	(**0.143**, 0.062)	0.063	0.057	0.444, 7/24, 0.192	0.062, 0.062
otL08rvlq	8891	(0.063, 0.242)	0.033	0.108	0.495, 6/24, 0.133	0.299, 0.541

6)) would be at the top of the result, followed by those in (fb, rvl (weight sum 5)), followed by those in (fb, fv (weight sum 4)), followed by those in (either fb or both rvl, fv (weight sum 3)), etc.

otL08fbe (blind feedback run): The submitted otL08fbe run was a blind feedback run based 50% on otL08fb and 25% each on expansion queries from the first 2 rows of otL08fb. Unlike last year, we used the new rank-based fusion algorithm for combining the expansion queries with the Boolean run.

For each run, only 100,000 rows were allowed to be submitted for each query.

2.3 Thresholding

New this year was the requirement to specify a cutoff value "K" for each topic at which the track's main measure, F_1@K, would be computed. F_1 is (2*Precision*Recall)/(Precision+Recall) and hence requires both high precision and high recall to produce a high score. If K is chosen too small, F_1 will be low from low recall. If K is chosen too large, F_1 will be low from low precision. Ideally, the retrieval set will have all of the relevant documents at the top, and K would be set to the number of relevant documents.

We did some diagnostic experiments with last year's runs using the SearchServer relevance() function score for thresholding (which is a score between 0 and 1000 for each document). We found that a fairly wide range of thresholds produced similar mean F_1 scores.

For our official submission, we thresholded the retrieval sets as follows:

For the 4 vector runs (otL08fv, otL08rvl, otL08rvlq, otL08rv), we set K so that just documents of relevance() score of 200 or higher were included. (And for highly relevant documents, we set K_h so that just documents of relevance() score of 225 or higher were included. Note that there was no training data for the highly relevant category.)

For the 2 Boolean runs (otL08fb and otL08db), we set K (and K_h) to the number retrieved. (Hence relevance-ranking did not matter for our K values for the Boolean runs.)

For the fusion run (otL08frw), K was set to include documents of at least weight sum 3 (i.e. any Boolean document in the result and any document from both the vector results (rvl and fv) would be in the top-K). K_h was set to include documents of at least weight sum 4 (i.e. documents that were both Boolean matches and in at least one of the vector runs).

Table 2: Impact of Experimental Techniques on F_1@K and F_1@K$_h$ (Ad Hoc task)

Expt	ΔF_1@K	95% Conf	vs.	3 Extreme Diffs (Topic)
frw-fb	0.046	(0.011, 0.081)	18-7-1	0.28 (121), 0.25 (147), −0.13 (129)
fbe-fb	0.055	(−0.008, 0.117)	15-11-0	0.50 (109), 0.38 (142), −0.22 (129)
fv-fb	0.029	(−0.033, 0.092)	16-10-0	0.39 (109), 0.27 (128), −0.28 (138)
rvl-fb	0.018	(−0.034, 0.070)	14-12-0	0.38 (121), 0.29 (147), −0.23 (129)
rv-fb	0.024	(−0.039, 0.087)	12-14-0	0.38 (121), 0.31 (124), −0.23 (138)
rvlq-fb	−0.035	(−0.091, 0.020)	10-16-0	0.31 (121), −0.28 (113), −0.30 (129)
db-fb	−0.111	(−0.168,−0.054)	3-23-0	−0.37 (128), −0.31 (110), 0.17 (119)
rvl-rv	−0.006	(−0.035, 0.024)	13-13-0	−0.20 (124), −0.18 (118), 0.16 (133)
rvl-rvlq	0.054	(0.021, 0.086)	20-6-0	0.27 (145), 0.18 (113), −0.10 (118)
frw-rvl	0.028	(−0.006, 0.061)	16-10-0	0.21 (132), 0.16 (139), −0.13 (124)
frw-fv	0.017	(−0.042, 0.075)	12-14-0	0.33 (139), −0.30 (128), −0.30 (109)
rvl-fv	−0.011	(−0.072, 0.050)	13-13-0	−0.37 (109), −0.36 (128), 0.25 (121)
fbe-rvl	0.036	(−0.015, 0.088)	19-7-0	0.48 (109), 0.30 (142), −0.18 (121)
fbe-fv	0.025	(−0.028, 0.079)	14-12-0	0.32 (142), 0.24 (139), −0.30 (128)
fbe-frw	0.008	(−0.037, 0.054)	11-15-0	0.41 (109), 0.23 (142), −0.19 (132)
	ΔF_1@K$_h$			(high relevance)
frw-fb	0.003	(−0.013, 0.020)	16-3-5	−0.13 (145), −0.06 (126), 0.08 (124)
fbe-fb	−0.006	(−0.037, 0.025)	14-6-4	−0.24 (126), −0.19 (138), 0.11 (124)
fv-fb	−0.028	(−0.069, 0.013)	11-13-0	−0.26 (126), −0.22 (138), 0.17 (110)
rvl-fb	−0.027	(−0.065, 0.012)	10-14-0	−0.21 (145), −0.21 (126), 0.18 (124)
rv-fb	−0.013	(−0.059, 0.033)	11-12-1	0.30 (124), −0.19 (138), −0.23 (126)
rvlq-fb	−0.058	(−0.096,−0.020)	6-17-1	−0.30 (145), −0.27 (126), 0.08 (121)
db-fb	−0.029	(−0.087, 0.029)	4-17-3	0.41 (148), 0.26 (119), −0.26 (145)
rvl-rv	−0.014	(−0.032, 0.004)	10-13-1	−0.12 (124), −0.10 (110), 0.07 (129)
rvl-rvlq	0.031	(0.006, 0.057)	16-8-0	0.26 (124), 0.09 (145), −0.06 (139)
frw-rvl	0.030	(−0.005, 0.065)	13-11-0	0.22 (138), 0.20 (139), −0.13 (121)
frw-fv	0.031	(−0.014, 0.077)	13-11-0	0.25 (138), 0.24 (139), −0.19 (145)
rvl-fv	0.001	(−0.041, 0.044)	12-11-1	−0.27 (145), 0.21 (121), 0.26 (124)
fbe-rvl	0.021	(−0.009, 0.051)	10-14-0	0.26 (145), 0.15 (139), −0.07 (124)
fbe-fv	0.022	(−0.009, 0.053)	11-12-1	0.18 (139), 0.18 (124), −0.10 (110)
fbe-frw	−0.009	(−0.041, 0.022)	12-8-4	−0.22 (138), −0.18 (126), 0.18 (145)

For the blind feedback run (otL08fbe), K was set to include documents of at least weight sum 2 (i.e. any Boolean document in the result and any document from both the expansion query results). K$_h$ was set to include documents of at least weight sum 3 (i.e. documents that were Boolean matches and also in at least one of the expansion query results).

2.4 Results

Table 1 lists several mean scores for the 8 submitted runs. The retrieval measures are defined in Section 4.1 of the Glossary at the end of the paper. The highest mean scores of each measure are in bold; however, see Tables 2 and 3 for which mean differences are statistically significant. (The columns of Tables 2 and 3 are explained in Section 4.2 of the Glossary.)

The updated fusion technique (rank-based merging the results of the vector runs with the Boolean run) produced a statistically significant increase in mean F_1@K (and also in last year's Recall@B measure) compared to the Boolean run (as per the "frw-fb" entries of Tables 2 and 3).

Table 3: Impact of Experimental Techniques on Recall@B (Ad Hoc task)

Expt	ΔR@B	95% Conf	vs.	3 Extreme Diffs (Topic)
frw-fb	0.029	(0.006, 0.053)	18-6-2	0.18 (137), 0.18 (121), −0.03 (128)
fbe-fb	0.032	(−0.021, 0.085)	15-8-3	0.53 (136), 0.41 (121), −0.08 (126)
fv-fb	0.014	(−0.043, 0.071)	14-12-0	0.53 (136), 0.18 (137), −0.21 (138)
rvl-fb	0.038	(−0.064, 0.140)	12-14-0	0.84 (149), 0.74 (121), −0.48 (137)
rv-fb	0.028	(−0.087, 0.143)	10-16-0	0.84 (149), 0.61 (124), −0.48 (137)
rvlq-fb	−0.023	(−0.119, 0.073)	9-16-1	0.79 (149), −0.48 (137), −0.57 (128)
db-fb	−0.206	(−0.303,−0.109)	1-25-0	−0.81 (137), −0.79 (128), 0.12 (119)
rvl-rv	0.010	(−0.038, 0.059)	12-12-2	0.40 (136), −0.19 (118), −0.33 (124)
rvl-rvlq	0.061	(0.008, 0.114)	19-6-1	0.45 (121), 0.42 (128), −0.16 (118)
frw-rvl	−0.009	(−0.108, 0.090)	15-11-0	−0.78 (149), −0.56 (121), 0.66 (137)
frw-fv	0.015	(−0.038, 0.068)	15-10-1	−0.53 (136), 0.17 (129), 0.21 (138)
rvl-fv	0.024	(−0.092, 0.139)	16-10-0	0.81 (149), 0.71 (121), −0.66 (137)
fbe-rvl	−0.006	(−0.104, 0.092)	14-12-0	−0.84 (149), 0.48 (137), 0.53 (136)
fbe-fv	0.018	(−0.026, 0.062)	13-12-1	0.38 (121), 0.21 (138), −0.18 (137)
fbe-frw	0.003	(−0.047, 0.053)	8-18-0	0.53 (136), 0.23 (121), −0.18 (137)
	ΔR@B			(high relevance)
frw-fb	0.051	(−0.003, 0.104)	17-2-5	0.50 (149), 0.40 (148), −0.06 (105)
fbe-fb	0.015	(−0.029, 0.058)	10-3-11	0.40 (148), 0.23 (121), −0.15 (139)
fv-fb	0.019	(−0.111, 0.149)	10-11-3	1.00 (109), 0.50 (149), −0.84 (138)
rvl-fb	0.129	(−0.015, 0.273)	13-8-3	1.00 (150), 1.00 (131), −0.32 (138)
rv-fb	0.165	(0.004, 0.327)	13-8-3	1.00 (150), 1.00 (131), −0.35 (138)
rvlq-fb	−0.036	(−0.203, 0.130)	12-11-1	1.00 (109), −0.90 (118), −0.94 (128)
db-fb	−0.273	(−0.418,−0.127)	1-20-3	−1.00 (128), −1.00 (137), 0.25 (119)
rvl-rv	−0.037	(−0.102, 0.028)	6-10-8	−0.56 (110), −0.41 (124), 0.21 (132)
rvl-rvlq	0.165	(−0.006, 0.336)	16-4-4	−1.00 (109), 0.94 (128), 1.00 (131)
frw-rvl	−0.078	(−0.214, 0.058)	12-7-5	−1.00 (150), −1.00 (131), 0.32 (138)
frw-fv	0.031	(−0.087, 0.149)	13-4-7	−1.00 (109), 0.28 (139), 0.84 (138)
rvl-fv	0.109	(−0.055, 0.274)	13-7-4	1.00 (150), 1.00 (131), −1.00 (109)
fbe-rvl	−0.114	(−0.248, 0.020)	9-12-3	−1.00 (150), −1.00 (131), 0.32 (138)
fbe-fv	−0.005	(−0.128, 0.119)	13-7-4	−1.00 (109), −0.50 (149), 0.84 (138)
fbe-frw	−0.036	(−0.082, 0.010)	4-13-7	−0.50 (149), −0.16 (139), 0.10 (121)

Removing instruction words from the request text also produced a statistically significant increase in both mean F_1@K and Recall@B (as per the "rvl-rvlq" entries of Tables 2 and 3).

As future work, we would like to walk through some of the individual topic differences to better understand the reasons for these impacts.

3 Relevance Feedback Task

The Relevance Feedback task re-used 40 topics from past years (though only 12 turned out to be assessed this year). Residual evaluation was used, i.e. documents judged in past years were dropped from the results before evaluating.

Table 4: Mean Scores of Submitted Relevance Feedback Task Runs

Run	K_r	(P@K_r, R@K_r)	F_1@K_r	F_1@R_r	GS10J, S1J, P5	R@B_r, R@ret_r
otRF08fb	3488	(**0.367**, 0.228)	**0.142**	0.151	**0.970**, 9/12, 0.656	0.228, 0.228
otRF08rvl	87738	(0.126, **0.607**)	0.134	0.185	0.727, 6/12, 0.450	0.096, 0.640
otRF08fv	70255	(0.099, 0.495)	0.119	0.202	0.811, 8/12, 0.533	0.103, 0.596
otRF08frw	39726	(0.140, 0.549)	0.116	**0.307**	0.953, **10/12**, **0.800**	0.236, **0.659**
otRF08fbF	36620	(0.118, 0.392)	0.084	0.261	0.879, 8/12, 0.633	**0.261**, 0.472
otRF08fbFR	8776	(0.266, 0.261)	0.082	0.261	0.879, 8/12, 0.633	**0.261**, 0.472
otRF08F	36716	(0.109, 0.327)	0.075	0.156	0.763, 7/12, 0.600	0.118, 0.446
otRF08FR	8826	(0.203, 0.129)	0.047	0.156	0.763, 7/12, 0.600	0.118, 0.446
(high rel)	K_{hr}	(P@K_{hr}, R@K_{hr})	F_1@K_{hr}	F_1@R_{hr}	GS10J, S1J, P5	R@B_r, R@ret_r
otRF08frw	2392	(**0.124**, 0.336)	**0.129**	0.145	0.671, 2/9, **0.289**	0.347, 0.647
otRF08fb	3870	(0.112, 0.336)	0.118	0.148	**0.750**, 2/9, 0.259	0.336, 0.336
otRF08fbFR	10150	(0.096, 0.400)	0.050	**0.202**	0.696, **3/9**, 0.267	**0.367**, 0.661
otRF08fbF	15305	(0.032, **0.516**)	0.044	**0.202**	0.696, **3/9**, 0.267	**0.367**, 0.661
otRF08rvl	39342	(0.018, 0.418)	0.029	0.036	0.497, **3/9**, 0.200	0.133, **0.718**
otRF08FR	10213	(0.016, 0.311)	0.028	0.089	0.538, 2/9, 0.244	0.237, 0.658
otRF08fv	22248	(0.018, 0.371)	0.025	0.085	0.583, 1/9, 0.156	0.185, 0.686
otRF08F	15390	(0.009, 0.444)	0.015	0.089	0.538, 2/9, 0.244	0.237, 0.658

3.1 Submitted Runs

We submitted 8 Relevance Feedback task runs. 4 were baseline runs (using the same techniques as 4 of our Ad Hoc runs). The other 4 runs used relevance feedback.

The 4 baseline runs were as follows:

otRF08fb (final Boolean run): Same approach as described above for the Ad Hoc otL08fb run.

otRF08fv (final vector run): Same approach as described above for the Ad Hoc otL08fv run.

otRF08rvl (request text vector run): Same approach as described above for the Ad Hoc otL08rvl run.

otRF08frw (fusion run): Same approach as described above for the Ad Hoc otL08frw run.

The 4 feedback runs were as follows:

otRF08F (pure feedback run): The known relevant documents (excluding those larger than 1 million bytes) were input to the SearchServer IS_ABOUT predicate which created a vector query from the highest weighted terms (based on a tf.idf calculation). English inflections were enabled, and stems in more than 5% of the collection documents were omitted. The same thresholding was used as for the other vector runs, i.e. K was set so that just documents of relevance() score of 200 or higher were included, and K_h was set so that just documents of relevance() score of 225 or higher were included.

otRF08fbF (fusion of Boolean and feedback): This run used ranked-based fusion of the Boolean run (otRF08fb) and the pure feedback (otRF08F) runs, with equal weights. K (and K_h) were set to the the same as for otRF08F. (Note that using the same K value for 2 runs does not imply that the (residual) K_r will be the same for the 2 runs because there may be a different number of previously judged documents in the first K retrieved by the 2 runs.)

otRF08FR (pure feedback run with sampling-based thresholds): This run used the same retrieval list as otRF08F, but the threshold K was set based on estimates of the number of relevant documents from past year's sampling. For the first 15 topics (which were from 2006) K was set to 1000 times the P100 of the humL06tvz depth probe run (or 50 if P100 was 0). For the latter 25 topics (which were from 2007) K was set to the ":est_rel:" for the topic in 2007. K_h was set to same as K for each topic.

otRF08fbFR (fusion of Boolean and feedback with sampling-based thresholds): This run used the same retrieval list as otRF08fbF and the same K and K_h thresholds as otRF08FR.

Table 5: Impact of Experimental Techniques on F_1@K$_r$ and F_1@K$_{hr}$ (Relevance Feedback Task)

Expt	ΔF_1@K$_r$	95% Conf	vs.	3 Extreme Diffs (Topic)
frw-fb	−0.027	(−0.178, 0.125)	7-5-0	−0.59 (85), −0.34 (31), 0.27 (60)
fv-fb	−0.023	(−0.189, 0.142)	8-4-0	−0.61 (85), −0.35 (31), 0.44 (73)
rvl-fb	−0.008	(−0.183, 0.166)	7-5-0	−0.63 (85), −0.35 (31), 0.48 (73)
F-fb	−0.067	(−0.213, 0.078)	5-7-0	−0.63 (85), −0.35 (31), 0.30 (79)
FR-fb	−0.096	(−0.211, 0.020)	4-8-0	−0.60 (85), −0.27 (31), 0.12 (80)
fbF-fb	−0.058	(−0.203, 0.086)	5-7-0	−0.60 (85), −0.35 (31), 0.30 (79)
fbFR-fb	−0.061	(−0.146, 0.025)	4-8-0	−0.36 (85), −0.27 (31), 0.13 (28)
FR-F	−0.029	(−0.094, 0.037)	6-6-0	−0.27 (79), −0.18 (60), 0.09 (28)
fbFR-fbF	−0.002	(−0.094, 0.090)	5-7-0	−0.28 (79), −0.21 (60), 0.24 (85)
frw-rvl	−0.018	(−0.085, 0.048)	8-4-0	−0.35 (73), −0.08 (89), 0.11 (47)
frw-fv	−0.003	(−0.063, 0.056)	9-3-0	−0.31 (73), 0.07 (80), 0.10 (47)
rvl-fv	0.015	(−0.004, 0.034)	8-4-0	0.09 (89), 0.05 (79), −0.02 (28)
	ΔF_1@K$_{hr}$			(high relevance)
frw-fb	0.011	(0.002, 0.019)	7-0-2	0.03 (60), 0.02 (85), 0.00 (80)
fv-fb	−0.093	(−0.208, 0.022)	4-5-0	−0.50 (31), −0.21 (60), 0.04 (28)
rvl-fb	−0.089	(−0.210, 0.032)	3-6-0	−0.51 (31), −0.22 (60), 0.08 (73)
F-fb	−0.103	(−0.220, 0.014)	3-6-0	−0.52 (31), −0.24 (60), 0.02 (28)
FR-fb	−0.090	(−0.205, 0.025)	2-7-0	−0.48 (31), −0.26 (60), 0.08 (28)
fbF-fb	−0.073	(−0.188, 0.041)	3-6-0	−0.52 (31), −0.10 (79), 0.03 (28)
fbFR-fb	−0.068	(−0.182, 0.047)	2-7-0	−0.48 (31), −0.18 (60), 0.12 (28)
FR-F	0.013	(−0.002, 0.029)	6-2-1	0.05 (28), 0.04 (31), −0.01 (60)
fbFR-fbF	0.006	(−0.044, 0.055)	6-3-0	−0.17 (60), 0.06 (79), 0.08 (28)
frw-rvl	0.100	(−0.022, 0.222)	6-3-0	0.51 (31), 0.25 (60), −0.08 (73)
frw-fv	0.104	(−0.012, 0.219)	5-4-0	0.50 (31), 0.25 (60), −0.04 (89)
rvl-fv	0.004	(−0.021, 0.028)	4-4-1	0.08 (73), 0.03 (89), −0.06 (28)

3.2 Results

Table 4 lists several mean scores for the 8 submitted runs. Note that mean scores may not be so reliable over small numbers of topics (12 counting all relevant documents, 9 when just counting highly relevant documents). We can still learn from investigating individual topic differences however.

For individual topics, the investigated relevance feedback technique, when merged in with the reference Boolean results, led to some substantial increases in Recall@B$_r$ without any large decreases (as per the "fbF-fb" entries of Table 6). However, in F_1@K$_r$ we do see some large decreases (e.g. topic 85 in Table 5) which might suggest that the experimental thresholding techniques (setting of K) were not very reliable. We should investigate further.

4 Glossary

4.1 Retrieval Measures

The retrieval measures of Tables 1 and 4 are defined as follows:

"Avg. K": The Average K value.

"P@K", "R@K" and "F_1@K": Estimated Precision, Recall and F_1 at Depth K (respectively).

"F_1@R": Estimated F_1 at Depth R (where R is the estimated number of relevant documents).

"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

Table 6: Impact of Experimental Techniques on Recall@B_r (Relevance Feedback Task)

Expt	ΔR@B_r	95% Conf	vs.	3 Extreme Diffs (Topic)
frw-fb	0.008	(−0.004, 0.020)	6-2-4	0.07 (28), 0.01 (89), −0.00 (60)
fv-fb	−0.124	(−0.311, 0.062)	5-7-0	−0.94 (85), −0.66 (14), 0.07 (31)
rvl-fb	−0.132	(−0.308, 0.045)	5-7-0	−0.94 (85), −0.59 (14), 0.03 (89)
F-fb	−0.110	(−0.279, 0.059)	3-9-0	−0.92 (85), −0.40 (14), 0.20 (28)
FR-fb	−0.110	(−0.279, 0.059)	3-9-0	−0.92 (85), −0.40 (14), 0.20 (28)
fbF-fb	0.033	(−0.007, 0.073)	7-1-4	0.19 (14), 0.17 (28), −0.00 (79)
fbFR-fb	0.033	(−0.007, 0.073)	7-1-4	0.19 (14), 0.17 (28), −0.00 (79)
FR-F	0.000	(−0.001, 0.001)	0-0-12	0.00 (73), 0.00 (28), 0.00 (89)
fbFR-fbF	0.000	(−0.001, 0.001)	0-0-12	0.00 (73), 0.00 (28), 0.00 (89)
frw-rvl	0.140	(−0.036, 0.315)	7-5-0	0.94 (85), 0.59 (14), −0.03 (36)
frw-fv	0.132	(−0.052, 0.317)	8-4-0	0.94 (85), 0.66 (14), −0.07 (31)
rvl-fv	−0.007	(−0.042, 0.028)	7-5-0	−0.14 (31), −0.11 (28), 0.07 (14)
	ΔR@B_r			(high relevance)
frw-fb	0.011	(0.000, 0.023)	6-0-3	0.04 (85), 0.03 (28), 0.00 (80)
fv-fb	−0.151	(−0.347, 0.044)	5-4-0	−0.87 (85), −0.31 (14), 0.04 (31)
rvl-fb	−0.204	(−0.375,−0.032)	2-7-0	−0.79 (85), −0.33 (31), 0.04 (89)
F-fb	−0.100	(−0.221, 0.022)	2-7-0	−0.38 (31), −0.33 (60), 0.21 (28)
FR-fb	−0.100	(−0.221, 0.022)	2-7-0	−0.38 (31), −0.33 (60), 0.21 (28)
fbF-fb	0.031	(−0.011, 0.072)	4-0-5	0.17 (14), 0.10 (28), 0.00 (73)
fbFR-fb	0.031	(−0.011, 0.072)	4-0-5	0.17 (14), 0.10 (28), 0.00 (73)
FR-F	0.000	(−0.001, 0.001)	0-0-9	0.00 (73), 0.00 (28), 0.00 (89)
fbFR-fbF	0.000	(−0.001, 0.001)	0-0-9	0.00 (73), 0.00 (28), 0.00 (89)
frw-rvl	0.215	(0.036, 0.394)	7-2-0	0.84 (85), 0.34 (31), −0.04 (89)
frw-fv	0.163	(−0.040, 0.366)	5-4-0	0.91 (85), 0.31 (14), −0.04 (31)
rvl-fv	−0.052	(−0.141, 0.037)	3-5-1	−0.37 (31), −0.14 (28), 0.07 (85)

robustness measure which exposes the downside of blind feedback techniques [11]. "Generalized Success@10" was originally introduced as "First Relevant Score" (FRS) in [12]. Intuitively, GS10J is a predictor of the percentage of topics for which a relevant document is returned in the first 10 rows.

"S1J": Success of the First Judged Document.

"P5": Estimated Precision at Depth 5.

"R@B": Estimated Recall at Depth B.

"R@ret": Estimated Recall of the full retrieval set.

"K_h": K value when just counting Highly relevant documents as relevant.

"R_h": Estimated number of Highly relevant documents.

"B": number of documents matching the final negotiated Boolean query.

"B_r" (RF task, residual evaluation): number of documents matching the final negotiated Boolean query after removing matches that were judged last year.

4.2 Difference Tables

For the comparison tables (such as Table 2), the 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). 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).

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

[6] 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.

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

[8] Douglas W. Oard, Bruce Hedin, Stephen Tomlinson and Jason R. Baron. Overview of the TREC 2008 Legal Track. (To appear in) Proceedings of TREC 2008.

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

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

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

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

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

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

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

[16] TREC 2008 Legal Track: Ad Hoc and Relevance Feedback Task Guidelines. http://trec-legal.umiacs.umd.edu/adhocRF08b.html