Experiments with the Negotiated Boolean Queries of the TREC 2009 Legal Track

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

February 2, 2010

Abstract

For our participation in the Batch Task of the TREC 2009 Legal Track, we produced several retrieval sets to compare experimental Boolean, vector, fusion and relevance feedback techniques for e-Discovery requests. In this paper, we have reported not just the mean scores of the experimental approaches but also the largest per-topic impacts of the techniques for several measures. The experimental automatic relevance feedback technique was found to attain a statistically significant gain over the reference Boolean result in both the mean Precision@B and F_1@K measures.

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 [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 Batch task of the TREC 2009 Legal Track.

2 Batch Task

The Batch task of the TREC 2009 Legal Track investigated the effectiveness of ad hoc and relevance feedback search techniques for e-Discovery.

The Batch task evolved from the Ad Hoc and Relevance Feedback tasks of past Legal Tracks. We have participated in the 4 years of the Legal Track to date. (We also have helped with coordinating the Legal Track for the past three years as described in [18], [10] and [6].)

As in the previous three years, the document collection to be searched was the IIT Complex Document Information Processing (IIT CDIP) test collection [8]. 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 character reader. Uncompressed, the collection was 61,251,357,065 bytes (57.0 GB). The average record size

[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.

(including metadata markup and the ocr document) was 8864 bytes, though the records varied considerably in length.

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 Batch task of the TREC Legal Track, the organizers re-used 10 of the production requests from past years, herein called topics, numbered from 7 to 145. 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 of some of these appear below.) The final negotiated Boolean query is sometimes referred to as the "reference Boolean" query.

Furthermore, the relevance judgments from previous years for the topics were available for use as an input to the systems. On average, 2000 document assessments were available per topic (ranging from 499 for topic 145 to 6500 for topic 103). These input relevance judgments are sometimes referred to as "training judgments", "training examples" or "training qrels". This year's scoring was done with entirely new judgments, however, and if the sampling happened to select a document judged in past years, it was not guaranteed that this year's assessors would agree with the past assessment.

[19] and [6] have more details on the track and task, and [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].

2.1 Indexing

Our index was the same as the past few years. 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 3 submitted Batch runs of August 2009 and 3 other diagnostic runs are described below. The relevance ranking approach was the same as 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, 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 3 submitted Batch task runs (and 3 other diagnostic runs) were as follows:

otL09fb (final Boolean run): (This run was not submitted). The otL09fb 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 (which was not actually one of this year's topics), 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 would be

```
SELECT RELEVANCE('2:3') AS REL, DOCNO
FROM LEGAL09FULL
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;
```

otL09fv (final vector run): (This run was not submitted). The otL09fv run was the same as otL09fb 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 would be

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

otL09rvl (request text vector run): The submitted otL09rvl run was the same as otL09fv 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 would be

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

otL09frw (baseline fusion run): (This run was not submitted). The otL09frw run was a weighted fusion of the final Boolean, request text vector and final vector runs: weight 3 on otL09fb, weight 2 on otL09rvl, and weight 1 on otL09fv. Each input run was retrieved to depth 1,500,000 (or however many documents it matched if it matched fewer than 1,500,000). This year we used the "Reciprocal Rank Fusion" (RRF) algorithm to combine the input runs. [4]

otL09F (pure feedback run): The submitted otL09F run did not make any use of the topic fields. Instead, the run used a feedback technique based on the set of documents that were previously judged relevant (the "feedback set"). Documents of 10000 bytes or more (in the xml formatting of the collection) were excluded from the feedback set in hopes of reducing the percentage of input text that was not relevant. In some cases the feedback set was further restricted to a random sample to cap the number of input documents at approximately 200. The documents of the final feedback set 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.

otL09frwF (feedback fusion run): The submitted otL09frwF run was a weighted fusion of the pure feedback, final Boolean, request text vector and final vector runs: weight 3 on otL09F, weight 3 on otL09fb, weight 2 on otL09rvl, and weight 1 on otL09fv (same fusion approach as otL09frw except for additionally including the feedback run as an input).

For each run, only 1,500,000 rows were allowed to be submitted for each query.

Run	Avg. K	R@K	P@K	F_1@K	Gray@K	Avg. Num. Judged@K
otL09F	198939	0.167	**0.398**	**0.196**	0.005	472 (329r, 142n, 1g)
otL09frwF	270389	**0.182**	0.377	0.189	**0.025**	610 (359r, 248n, 3g)
otL09rvl	192605	0.165	0.267	0.162	0.006	632 (344r, 284n, 5g)
otL09frw	79427	0.098	0.339	0.125	0.007	570 (320r, 246n, 4g)
otL09fv	79427	0.075	0.286	0.098	0.011	510 (292r, 214n, 4g)
otL09fb	27462	0.037	0.391	0.063	0.016	277 (182r, 92n, 3g)
(high rel)	Avg. K_h	R@K_h	P@K_h	F_1@K_h	Gray@K_h	Avg. Num. Judged@K_h
otL09F	26582	0.167	**0.266**	**0.132**	0.000	260 (115r, 145n, 0g)
otL09frwF	26582	0.168	0.215	0.113	0.016	384 (134r, 248n, 1g)
otL09rvl	61608	**0.239**	0.119	0.105	0.007	465 (141r, 321n, 3g)
otL09frw	20756	0.226	0.180	0.114	0.004	412 (132r, 278n, 3g)
otL09fv	20756	0.160	0.165	0.105	0.007	353 (115r, 236n, 3g)
otL09fb	27462	0.143	0.149	0.063	**0.016**	277 (84r, 190n, 3g)

Table 1: Mean Set-based Scores of Experimental Batch Task Runs

Run	Avg. Ret	P@B	R@B	F_1@R	R@ret	indAP	GS10J	S1J
otL09F	1500000	**0.558**	**0.055**	**0.251**	0.582	**0.653**	0.957	6/10
otL09frwF	1500000	0.531	0.054	0.238	**0.609**	0.618	0.978	8/10
otL09rvl	1500000	0.476	0.041	0.194	0.535	0.591	0.909	9/10
otL09frw	1500000	0.460	0.046	0.204	0.542	0.583	0.993	9/10
otL09fv	1482237	0.447	0.037	0.194	0.575	0.550	0.940	7/10
otL09fb	27462	0.391	0.037	0.037	0.037	0.275	**1.000**	**10/10**
(high rel)	Avg. Ret	P@B	R@B	F_1@R_h	R@ret	indAP	GS10J	S1J
otL09F	1500000	**0.256**	0.228	**0.213**	**0.757**	**0.409**	0.815	2/10
otL09frwF	1500000	0.221	**0.230**	0.193	0.719	0.366	**0.868**	**5/10**
otL09rvl	1500000	0.239	0.174	0.164	0.688	0.338	0.779	**5/10**
otL09frw	1500000	0.206	0.224	0.153	0.686	0.331	0.826	**5/10**
otL09fv	1482237	0.168	0.151	0.143	0.725	0.312	0.763	4/10
otL09fb	27462	0.149	0.143	0.069	0.143	0.167	0.828	**5/10**

Table 2: Mean Rank-based Scores of Experimental Batch Task Runs

2.3 Thresholding

Like last year, the systems were required 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.

(Also, the systems were required to specify a cutoff value "K_h" for each topic for when just counting "highly relevant" documents as relevant.)

For our runs this year, we thresholded the retrieval sets as follows:

For the 2 vector runs (otL09fv and otL09rvl), 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.) This thresholding approach was the same as used for the

Table 3: Impact of Experimental Techniques on F_1@K, F_1@K$_h$, F_1@R and F_1@R$_h$

Expt	ΔF_1@K	95% Conf	vs.	3 Extreme Diffs (Topic)
F-frwF	0.008	(−0.019, 0.034)	6-4-0	0.08 (105), 0.05 (80), −0.07 (138)
frwF-frw	0.064	(−0.027, 0.155)	7-3-0	0.35 (103), 0.22 (102), −0.16 (7)
frw-rvl	−0.037	(−0.088, 0.013)	4-6-0	−0.19 (103), −0.12 (102), 0.06 (89)
rvl-fv	0.064	(0.017, 0.111)	8-2-0	0.21 (103), 0.13 (102), −0.04 (80)
F-frw	0.072	(−0.024, 0.167)	7-3-0	0.32 (103), 0.25 (102), −0.17 (7)
F-rvl	0.034	(−0.032, 0.101)	6-4-0	0.18 (105), 0.13 (103), −0.16 (7)
F-fv	0.098	(0.009, 0.188)	8-2-0	0.34 (103), 0.28 (105), −0.08 (7)
F-fb	0.134	(0.041, 0.226)	8-2-0	0.39 (103), 0.29 (80), −0.04 (138)
frwF-fb	0.126	(0.042, 0.210)	9-1-0	0.42 (103), 0.24 (80), −0.03 (51)
frw-fb	0.062	(−0.008, 0.132)	7-3-0	0.24 (80), 0.18 (7), −0.14 (102)
rvl-fb	0.099	(0.031, 0.167)	8-2-0	0.26 (103), 0.23 (145), −0.02 (51)
fv-fb	0.035	(−0.033, 0.104)	6-4-0	0.24 (80), 0.15 (145), −0.14 (102)
	ΔF_1@K$_h$			(high relevance)
F-frwF	0.019	(−0.009, 0.048)	7-3-0	0.10 (103), 0.08 (145), −0.03 (138)
frwF-frw	−0.001	(−0.038, 0.035)	4-6-0	0.09 (102), 0.09 (103), −0.08 (80)
frw-rvl	0.009	(−0.068, 0.085)	6-4-0	0.30 (89), −0.07 (102), −0.17 (103)
rvl-fv	0.001	(−0.089, 0.091)	7-3-0	−0.38 (89), 0.10 (105), 0.14 (103)
F-frw	0.018	(−0.034, 0.070)	5-5-0	0.19 (103), 0.09 (102), −0.07 (138)
F-rvl	0.027	(−0.037, 0.091)	6-4-0	0.27 (89), 0.10 (145), −0.09 (105)
F-fv	0.027	(−0.022, 0.077)	6-4-0	0.16 (103), 0.11 (145), −0.11 (89)
F-fb	0.069	(0.004, 0.134)	9-1-0	0.25 (89), 0.17 (145), −0.10 (138)
frwF-fb	0.050	(−0.004, 0.103)	8-2-0	0.24 (89), 0.10 (80), −0.07 (138)
frw-fb	0.051	(−0.017, 0.119)	5-5-0	0.28 (89), 0.18 (80), −0.06 (102)
rvl-fb	0.042	(−0.006, 0.091)	7-3-0	0.17 (80), 0.13 (103), −0.07 (138)
fv-fb	0.042	(−0.039, 0.122)	6-4-0	0.36 (89), 0.12 (80), −0.10 (138)
	ΔF_1@R			3 Extreme Diffs (Topic)
F-frwF	0.013	(−0.024, 0.049)	3-7-0	0.13 (105), 0.09 (89), −0.04 (7)
frwF-frw	0.035	(0.006, 0.064)	8-2-0	0.11 (105), 0.11 (80), −0.02 (104)
frw-rvl	0.010	(−0.001, 0.021)	8-2-0	0.04 (80), 0.02 (103), −0.02 (105)
rvl-fv	0.000	(−0.022, 0.023)	5-5-0	−0.06 (80), 0.03 (145), 0.05 (105)
fv-fb	0.157	(0.074, 0.240)	10-0-0	0.42 (103), 0.27 (80), 0.00 (51)
F-frw	0.048	(−0.009, 0.104)	6-4-0	0.24 (105), 0.16 (89), −0.04 (104)
F-rvl	0.058	(0.004, 0.112)	8-2-0	0.22 (105), 0.16 (89), −0.04 (104)
F-fv	0.058	(0.000, 0.116)	7-3-0	0.28 (105), 0.14 (89), −0.02 (104)
	ΔF_1@R$_h$			(high relevance)
F-frwF	0.019	(−0.009, 0.047)	7-3-0	0.08 (103), 0.07 (89), −0.05 (138)
frwF-frw	0.040	(0.017, 0.064)	9-1-0	0.09 (145), 0.09 (105), −0.00 (102)
frw-rvl	−0.011	(−0.041, 0.019)	4-5-1	−0.09 (89), −0.09 (105), 0.05 (102)
rvl-fv	0.021	(−0.004, 0.045)	7-2-1	0.10 (105), 0.05 (89), −0.04 (102)
fv-fb	0.074	(−0.020, 0.167)	7-2-1	0.44 (89), 0.13 (103), −0.11 (138)
F-frw	0.060	(0.015, 0.104)	7-3-0	0.15 (89), 0.13 (105), −0.05 (138)
F-rvl	0.049	(0.019, 0.078)	8-2-0	0.09 (103), 0.09 (80), −0.03 (104)
F-fv	0.070	(0.031, 0.108)	8-2-0	0.15 (105), 0.13 (80), −0.01 (102)

corresponding vector runs last year.

For the Boolean run (otL09fb), we set K (and K$_h$) to the number retrieved. (Hence relevance-ranking

Table 4: Impact of Experimental Techniques on P@B

Expt	ΔP@B	95% Conf	vs.	3 Extreme Diffs (Topic)
F-frwF	0.027	(−0.056, 0.110)	7-3-0	−0.26 (138), 0.15 (105), 0.21 (104)
frwF-frw	0.071	(0.013, 0.129)	7-3-0	0.25 (7), 0.17 (105), −0.04 (102)
frw-rvl	−0.016	(−0.073, 0.040)	5-5-0	−0.16 (7), −0.14 (80), 0.12 (102)
rvl-fv	0.029	(−0.067, 0.125)	6-4-0	0.27 (80), 0.21 (138), −0.22 (89)
F-frw	0.098	(−0.005, 0.201)	7-3-0	0.32 (105), 0.26 (7), −0.18 (138)
F-rvl	0.082	(0.000, 0.163)	9-1-0	0.26 (105), 0.21 (103), −0.16 (138)
F-fv	0.111	(−0.005, 0.227)	7-3-0	0.39 (105), 0.28 (80), −0.21 (89)
F-fb	0.167	(0.042, 0.292)	7-3-0	0.37 (105), 0.36 (7), −0.20 (138)
frwF-fb	0.140	(0.062, 0.217)	9-1-0	0.35 (7), 0.32 (89), −0.00 (51)
frw-fb	0.069	(0.007, 0.130)	8-2-0	0.31 (89), 0.10 (7), −0.03 (138)
rvl-fb	0.085	(−0.008, 0.178)	6-4-0	0.34 (89), 0.26 (7), −0.07 (51)
fv-fb	0.056	(−0.078, 0.190)	5-5-0	0.56 (89), 0.17 (7), −0.25 (138)
	ΔP@B			(high relevance)
F-frwF	0.035	(0.004, 0.066)	7-3-0	0.10 (103), 0.09 (7), −0.04 (138)
frwF-frw	0.015	(−0.015, 0.046)	5-5-0	0.12 (7), 0.06 (145), −0.04 (104)
frw-rvl	−0.033	(−0.071, 0.004)	4-5-1	−0.14 (80), −0.10 (7), 0.06 (138)
rvl-fv	0.072	(−0.050, 0.193)	8-2-0	0.57 (80), 0.13 (145), −0.18 (89)
F-frw	0.050	(−0.005, 0.105)	6-3-1	0.20 (7), 0.14 (103), −0.07 (138)
F-rvl	0.017	(−0.018, 0.052)	6-4-0	0.11 (7), 0.08 (105), −0.06 (80)
F-fv	0.089	(−0.023, 0.201)	8-2-0	0.51 (80), 0.16 (105), −0.17 (89)
F-fb	0.107	(0.033, 0.180)	7-3-0	0.30 (7), 0.22 (89), −0.07 (138)
frwF-fb	0.072	(0.016, 0.128)	7-3-0	0.21 (7), 0.20 (89), −0.04 (138)
frw-fb	0.057	(0.014, 0.099)	8-2-0	0.22 (89), 0.10 (7), −0.00 (138)
rvl-fb	0.090	(0.030, 0.150)	8-2-0	0.21 (89), 0.19 (7), −0.06 (138)
fv-fb	0.018	(−0.104, 0.140)	6-4-0	0.39 (89), 0.16 (7), −0.38 (80)

did not matter for our K values for the Boolean runs.)

For the baseline fusion run (otL09frw), we just used the same K (and K_h) values as otL09fv this year.

For the feedback runs (otL09F and otL09frwF), the training qrels were a factor in choosing K (and K_h). One input was the "retrospective optimal K value" from using the retroK option of the l07_eval scoring utility to determine what value K would have produced the highest F_1 for the run when using the training qrels. Another input was the estimated number of relevant documents for the topic based on the training qrels (which was listed in the provided estRelL09.append file). The submitted K value was the greater of the retrospective optimal K value and estRelL09.append K value, plus 10 percent (in case the deeper runs this year led to greater numbers of relevant documents). (For K_h, we just used the K_h values from estRelL09.append, which for the 5 topics with past highly relevant judgments was the estimated number of highly relevant documents, and for the other 5 topics just used 14% of the estimated number of relevant documents.)

2.4 Results

Tables 1 and 2 list several mean scores for the 6 experimental runs. The retrieval measures are defined in Section 3.1 of the Glossary at the end of the paper. The highest mean scores of each measure are in bold; however, see Tables 3-6 for which mean differences are statistically significant. (The columns of Tables 3-6 are explained in Section 3.2 of the Glossary.)

We see that the feedback runs outperformed the plain vector runs when comparing to the Boolean run.

Table 5: Impact of Experimental Techniques on R@B

Expt	ΔR@B	95% Conf	vs.	3 Extreme Diffs (Topic)
F-frwF	0.000	(−0.016, 0.017)	5-5-0	0.06 (105), −0.02 (138), −0.03 (102)
frwF-frw	0.008	(−0.008, 0.024)	6-4-0	0.08 (105), 0.01 (145), −0.01 (102)
frw-rvl	0.005	(−0.001, 0.011)	7-3-0	0.02 (102), 0.02 (138), −0.01 (103)
rvl-fv	0.004	(−0.005, 0.014)	6-4-0	0.03 (105), 0.02 (89), −0.01 (102)
F-frw	0.009	(−0.023, 0.040)	7-3-0	0.14 (105), −0.02 (138), −0.05 (102)
F-rvl	0.014	(−0.016, 0.043)	8-2-0	0.14 (105), 0.02 (145), −0.03 (102)
F-fv	0.018	(−0.018, 0.054)	8-2-0	0.17 (105), 0.02 (89), −0.04 (102)
F-fb	0.018	(−0.018, 0.054)	7-3-0	0.17 (105), 0.04 (89), −0.03 (102)
frwF-fb	0.018	(−0.004, 0.040)	9-1-0	0.10 (105), 0.05 (89), −0.01 (138)
frw-fb	0.009	(−0.001, 0.020)	8-2-0	0.04 (89), 0.03 (105), −0.00 (145)
rvl-fb	0.004	(−0.008, 0.017)	5-4-1	0.04 (89), 0.03 (105), −0.02 (138)
fv-fb	−0.000	(−0.010, 0.010)	7-3-0	−0.04 (138), 0.01 (102), 0.02 (89)
	ΔR@B			(high relevance)
F-frwF	−0.002	(−0.059, 0.054)	3-6-1	0.19 (105), −0.06 (89), −0.17 (51)
frwF-frw	0.006	(−0.015, 0.027)	7-2-1	0.07 (105), 0.02 (145), −0.06 (138)
frw-rvl	0.050	(−0.036, 0.135)	4-5-1	0.42 (51), 0.08 (138), −0.05 (103)
rvl-fv	0.022	(−0.036, 0.081)	8-2-0	0.20 (89), 0.10 (105), −0.17 (51)
F-frw	0.004	(−0.066, 0.074)	5-5-0	0.26 (105), −0.08 (138), −0.17 (51)
F-rvl	0.054	(−0.010, 0.118)	9-1-0	0.25 (51), 0.24 (105), −0.00 (104)
F-fv	0.076	(0.007, 0.146)	9-1-0	0.33 (105), 0.21 (89), −0.00 (104)
F-fb	0.085	(−0.061, 0.231)	7-3-0	0.61 (89), 0.35 (105), −0.20 (51)
frwF-fb	0.087	(−0.048, 0.223)	7-3-0	0.67 (89), 0.16 (105), −0.06 (138)
frw-fb	0.081	(−0.050, 0.212)	9-1-0	0.66 (89), 0.09 (105), −0.03 (51)
rvl-fb	0.031	(−0.129, 0.191)	8-2-0	0.60 (89), 0.11 (105), −0.45 (51)
fv-fb	0.009	(−0.096, 0.113)	6-4-0	0.40 (89), −0.10 (138), −0.27 (51)

For example, in Table 4, the "fv-fb" entries for Precision@B (where B is the depth to which the Boolean run retrieved) show that the final vector run outscored the final Boolean run on just half of the queries, whereas the "F-fb" entries indicate a statistically significant advantage for the pure feedback run over the final Boolean run in Precision@B using either All Relevant or just Highly Relevant documents.

The tables flag topic 105 as one for which feedback was particularly helpful. This topic's request text was "Please produce all reports, written memoranda, correspondence, and other documents related to building design compliance or noncompliance with structural standards, and compliance or noncompliance with structural regulations.", and its final Boolean query was "`(build! OR structure!) AND (design! OR plan OR scheme OR blueprint) AND ((compliance OR comply OR complies OR obey! OR correspond! OR meet! OR adhere! OR conform) w/5 (regulat! OR code! OR law! OR ordinanc! OR rule! OR statut!))`". In the feedback query, we see helpful looking terms that were not in the request text or final Boolean query such as "ASHRAE", "CONTAMINANTS", "HVAC", "IAQ", "OSHA", "SBSC" and "VENTILATION".

Last year, our pure feedback run (otRF08F) did not do well compared to the Boolean run. We suspect the decision this year to just use relevant documents of less than 10000 bytes improved the quality of the feedback run substantially this year. We have verified that Precision@B was higher with otL09F than otRF08F on the two topics in common (topic 80, 0.989 vs. 0.902; topic 89, 0.460 vs. 0.365). We suspect further improvements to the feedback run are possible by using a formula that favors terms that appear in multiple relevant documents (instead of appending all of the relevant documents together before picking the terms).

Table 6: Impact of Experimental Techniques on R@ret and indAP

Expt	ΔR@ret	95% Conf	vs.	3 Extreme Diffs (Topic)
F-frwF	−0.027	(−0.059, 0.006)	2-7-1	−0.10 (145), −0.08 (80), 0.06 (138)
frwF-frw	0.067	(−0.010, 0.144)	7-3-0	0.30 (51), 0.16 (105), −0.13 (138)
frw-rvl	0.007	(−0.033, 0.048)	6-3-1	−0.13 (89), 0.05 (80), 0.12 (138)
rvl-fv	−0.040	(−0.120, 0.041)	3-7-0	−0.30 (51), −0.16 (80), 0.13 (89)
fv-fb	0.538	(0.430, 0.645)	10-0-0	0.84 (89), 0.74 (105), 0.26 (104)
F-frw	0.040	(−0.030, 0.110)	7-3-0	0.30 (51), 0.12 (105), −0.07 (138)
F-rvl	0.047	(−0.028, 0.122)	7-3-0	0.30 (51), 0.13 (80), −0.13 (89)
F-fv	0.008	(−0.032, 0.047)	3-6-1	0.11 (138), 0.08 (105), −0.06 (103)
	ΔR@ret			(high relevance)
F-frwF	0.037	(−0.042, 0.117)	3-3-4	0.36 (138), 0.10 (80), −0.07 (102)
frwF-frw	0.033	(−0.070, 0.137)	4-2-4	0.41 (105), 0.13 (145), −0.21 (102)
frw-rvl	−0.001	(−0.024, 0.021)	3-2-5	−0.08 (80), 0.03 (7), 0.06 (103)
rvl-fv	−0.037	(−0.168, 0.094)	2-5-3	−0.40 (105), −0.26 (138), 0.29 (7)
fv-fb	0.582	(0.470, 0.694)	10-0-0	0.96 (105), 0.76 (89), 0.39 (80)
F-frw	0.071	(−0.059, 0.200)	5-2-3	0.41 (105), 0.36 (138), −0.28 (102)
F-rvl	0.069	(−0.058, 0.197)	6-2-2	0.41 (105), 0.36 (138), −0.28 (102)
F-fv	0.032	(−0.062, 0.126)	6-2-2	0.32 (7), 0.11 (138), −0.27 (104)
	ΔindAP			3 Extreme Diffs (Topic)
F-frwF	0.035	(−0.028, 0.098)	6-4-0	0.29 (105), 0.08 (7), −0.07 (51)
frwF-frw	0.035	(0.006, 0.063)	7-3-0	0.13 (105), 0.08 (7), −0.01 (51)
frw-rvl	−0.007	(−0.044, 0.029)	4-6-0	0.11 (51), −0.06 (105), −0.07 (7)
rvl-fv	0.041	(−0.007, 0.088)	7-3-0	0.19 (138), 0.11 (80), −0.06 (51)
fv-fb	0.275	(0.108, 0.442)	8-2-0	0.58 (89), 0.58 (7), −0.13 (51)
F-frw	0.070	(−0.020, 0.160)	6-4-0	0.42 (105), 0.16 (7), −0.08 (51)
F-rvl	0.063	(−0.007, 0.132)	8-2-0	0.36 (105), 0.09 (7), −0.02 (102)
F-fv	0.104	(0.017, 0.190)	8-2-0	0.44 (105), 0.17 (138), −0.03 (51)
	ΔindAP			(high relevance)
F-frwF	0.043	(0.009, 0.078)	7-3-0	0.13 (89), 0.13 (105), −0.03 (138)
frwF-frw	0.036	(0.010, 0.061)	9-1-0	0.12 (89), 0.07 (138), −0.02 (104)
frw-rvl	−0.007	(−0.032, 0.018)	4-6-0	−0.08 (138), −0.05 (80), 0.05 (145)
rvl-fv	0.025	(−0.030, 0.081)	4-6-0	0.20 (138), 0.16 (80), −0.05 (7)
fv-fb	0.146	(0.046, 0.246)	8-2-0	0.39 (80), 0.34 (89), −0.05 (138)
F-frw	0.079	(0.026, 0.132)	8-2-0	0.25 (89), 0.17 (105), −0.04 (104)
F-rvl	0.072	(0.008, 0.136)	7-3-0	0.29 (89), 0.16 (105), −0.04 (104)
F-fv	0.097	(0.041, 0.153)	9-1-0	0.25 (89), 0.18 (80), −0.03 (104)

The tables show that topic 138 was an exceptional case in which the Boolean query still outperformed both the plain vector and pure feedback approaches in various measures, including Precision@B and F_1@K. This topic's request text was "All documents describing or detailing instances of government subsidies for competitive local products.", and its final Boolean query was "`(China OR CN OR PRC OR "Hong Kong" OR HK OR Japan OR JP OR Taiwan OR TW OR ROC OR India OR Philippines OR PH OR Cambodia OR KH OR Vietnam OR VN OR "North Korea" OR "South Korea" OR KP OR Thailand OR Asia OR EMEA OR Government OR Market) AND (Subsidy OR subsidies)`". We suspect that requiring 'subsidy' or 'subsidies' to be in the result was helpful for precision, whereas in the plain vector and pure feedback approaches these terms had relatively little weight. We should investigate further to see if alternate feedback-based formulations would perform better. In general though it may be advisable to manually

Table 7: Impact of Experimental Techniques on GS10J

Expt	ΔGS10J	95% Conf	vs.	3 Extreme Diffs (Topic)
F-frwF	−0.022	(−0.044, 0.001)	0-3-7	−0.07 (138), −0.07 (145), 0.00 (103)
frwF-frw	−0.014	(−0.051, 0.022)	1-2-7	−0.14 (104), −0.07 (51), 0.07 (80)
frw-rvl	0.083	(−0.101, 0.268)	1-1-8	0.91 (51), 0.00 (103), −0.07 (80)
rvl-fv	−0.030	(−0.237, 0.177)	3-1-6	−0.91 (51), 0.14 (89), 0.32 (138)
fv-fb	−0.060	(−0.130, 0.009)	0-3-7	−0.32 (138), −0.14 (89), 0.00 (103)
F-frw	−0.036	(−0.081, 0.009)	1-4-5	−0.14 (51), −0.14 (104), 0.07 (80)
F-rvl	0.047	(−0.116, 0.210)	1-3-6	0.77 (51), −0.07 (145), −0.14 (104)
F-fv	0.017	(−0.063, 0.098)	3-3-4	0.25 (138), 0.14 (89), −0.14 (104)
	ΔGS10J			(high relevance)
F-frwF	−0.053	(−0.211, 0.105)	2-6-2	−0.56 (104), −0.25 (7), 0.39 (51)
frwF-frw	0.042	(−0.088, 0.172)	4-3-3	0.46 (51), 0.21 (80), −0.32 (104)
frw-rvl	0.047	(−0.061, 0.155)	5-2-3	0.37 (103), 0.21 (7), −0.21 (80)
rvl-fv	0.016	(−0.089, 0.120)	3-4-3	0.32 (138), 0.21 (102), −0.21 (7)
fv-fb	−0.064	(−0.203, 0.074)	2-5-3	−0.50 (103), −0.25 (138), 0.29 (105)
F-frw	−0.011	(−0.289, 0.267)	5-5-0	−0.88 (104), −0.32 (7), 0.86 (51)
F-rvl	0.036	(−0.244, 0.316)	4-5-1	−0.88 (104), 0.30 (103), 0.86 (51)
F-fv	0.051	(−0.236, 0.339)	6-3-1	−0.88 (104), 0.43 (103), 0.85 (51)

supervise the feedback process.

We found that fusion of the baseline runs with the pure feedback run did not, on average, improve upon the pure feedback run in most measures (i.e. the F run typically had a higher mean score than the frwF run). Note that we do not blame the new RRF fusion technique for this result; an (unreported) experiment comparing the frw-style run with the RRF technique and last year's fusion technique did not find much difference. The tables often flag topic 105 as one of the topics that declined with fusion.

3 Glossary

3.1 Retrieval Measures

The retrieval measures of Tables 1 and 2 are defined as follows:

"Avg. K": The Average K value.

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

"Gray@K": Estimated percentage of the top-K that was "gray" documents (e.g. documents too long for the assessor to fully review).

"Avg. Num. Judged@K" is the actual number of judged documents in the top-K, followed in parentheses by the actual number of relevant (r), non-relevant (n) and gray (g) documents.

"Avg. Ret": The Average number of Retrieved documents per topic.

"P@B" and "R@B": Estimated Precision and Recall at Depth B (where B is the number of documents matching the final negotiated Boolean query).

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

"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).

"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.

"S1J": Success of the First Judged Document.

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

"R_h": Estimated number of Highly relevant documents.

3.2 Difference Tables

For the comparison tables (such as Table 3), 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; strictly speaking, for 10 topics, it would have been a little more accurate to have used a multiplier of 2.3 instead of 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).

4 Conclusions

For our participation in the Batch Task of the TREC 2009 Legal Track, we produced several experimental runs to compare experimental Boolean, vector, fusion and relevance feedback techniques for e-Discovery requests. This paper reported not just the mean scores of the runs but also the largest per-topic impacts of the techniques for several measures. We found that the experimental automatic relevance feedback technique (which just used shorter relevant documents for feedback) produced a statistically significant gain over the reference Boolean result in both the mean Precision@B and F_1@K measures. However, there were still cases (most notably topic 138) in which the reference Boolean query produced the higher score. Fusion of the Boolean result (and other baseline vector approaches) with the relevance feedback result further increased the scores for some individual topics, but not on average in most measures. While this paper focused on automated approaches (aside from the reference Boolean query), we suspect that for best results in practice one should manually supervise the approaches.

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, Charles L. A. Clarke and Stefan Büttcher. Reciprocal Rank Fusion outperforms Condorcet and Individual Rank Learning Methods. *SIGIR 2009*, pp. 758-759.

[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. (To appear in) 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, Douglas W. Oard, Jason R. Baron and Paul Thompson. Overview of the TREC 2007 Legal Track. Proceedings of TREC 2007.

[19] TREC 2009 Legal Track: Batch Task Guidelines. http://trec-legal.umiacs.umd.edu/batch09a.html