Examining Overfitting in Relevance Feedback: Sabir Research at TREC 2007

Chris Buckley
Sabir Research, Inc.
cabuckley@sabir.com

1 Introduction

Sabir Research participated in TREC-2007 in the Million Query and Legal tracks. This writeup focuses on the Legal track, and in particular on the relevance feedback and interactive tasks within the Legal track.

The information retrieval software used was the research version of SMART 16.0. SMART was originally developed in the early 1960's by Gerard Salton and since then has continued to be a leading research information retrieval tool. It continues to use a statistical vector space model, with stemming, stop words, weighting, inner product similarity function, and ranked retrieval.

2 Million Query Track

Sabir submitted one very simple run to the Million Query Track. Documents and queries were weighted with SMART ltu.Lnu weights [3]. No web structure was used, and there was no expansion of the original very short topic statements. The Terabyte collection was indexed and compressed as described in TREC-2005 [2]. The weighted index was 4.1 Gbytes.

The official evaluation results for the run sabmq07a1 are

Million Query Results, Run sabmq07a1	
Measure	Value
valid_topics	1153
statMAP_on_valid_topics	0.1519
statMRP_on_valid_topics	0.1737
statMPC_30 _on_valid_topics	0.1342
statMNDCG_on_valid_topics	0.2268
ExpectedMAP	0.0490

This was a very mediocre run, using weighting and matching algorithms that are at least 12 years old - nothing exciting here at all!

3 Legal Track

I participated in the Ad Hoc, Relevance Feedback, and Interactive tasks of the TREC-2007 Legal track. The particular options of SMART used in the track were very "vanilla" retrieval options. Except for the optimizer described below, all algorithms date from TREC 4 or before. Unless otherwise stated, retrieval was simple single term ltu.Lnu weighting, inner product similarity function, simple Rocchio feedback weighting (when done), and simple Rocchio expansion by highest weighted terms within the collection. The OCR nature of the collection discouraged non-directed experimentation, especially expansion. On other collections, for example Terabyte, the algorithms used here perform about 25% to 33% less well than the current top published algorithms - MAP on TREC-2006 Terabyte was .245 compared to MAP of around .37 for top groups who did not use link information.

The basic collection was indexed using a hybrid controlled-dictionary/hash-pool as the concept dictionary in order to reduce the impact of the OCR errors. If an ordinary naive dictionary was used, it would contain about 188 million distinct entries (mostly OCR errors), and the text of the unique words would take over 2 Gigabytes (uncompressed). Instead, a base list of known good words was formed from words that

- occur in on-line dictionaries (eg wordnet, moby, collins)
- occur more than once in TREC volumes 1-5
- occur more than 5 times in a 5% sample of the terabyte collection, with some manual editing afterwards.
- are a SMART stem form of above (about 80,000 new "words")

Any word that appears in those 443,542 words was assigned a concept number <= 700,000 (very common stop words like "a" or "the" were in the list, but were ignored and not assigned a concept). Any word that does not occur in this controlled list of words was assigned a concept number between 700,000 and 20,000,000. The concept number was a simple hash in that range. There were many (often 20-30) very different words assigned to each concept number – there was no distinction made between those words at all.

3.1 Legal Ad Hoc Task

I submitted 4 Ad Hoc runs. All runs were automatic runs, using the OCR version of the collection only, indexing the metadata portion of the documents as part of the document representation, but otherwise not treating the metadata in any special way.

Legal Ad Hoc Runs and Results		
Runname	Est-RB	Description
SabL07ar1	0.0966	Query used request only. Original OCR collection
SabL07ar2	0.0979	Query used request only. OCR with rare OCR terms deleted
SabL07arbn	0.1495	Query used request and all negotiated Boolean forms.
SabL07ab1	0.1320	Query used request and all negotiated Boolean forms with Blind feedback. Added 20 terms from top 30 docs retrieved.

SabL07ar1 was the base run, using the full indexed OCR collection and only the textual request portion of the query.

SabL07ar2 used a copy of the OCR collection that had a string of '*' characters of the same length that were textually substituted for each suspected OCR error. A suspected OCR error was defined as a token that was not in the controlled dictionary, and was not used in a document's metadata, and did not occur in more than 5 documents (thanks to Stephen Tomlinson for the occurrence list of these terms). This reduced the number of unique terms by over a factor of 20, from 188 million to about 9 million. This improved ease of indexing tremendously, and the results show that there was no impact on basic retrieval effectiveness.

In TREC-2006, there was an improvement when groups used terms occurring in the Boolean query as well as just the natural language information request. SabL07arbn used the same retrieval function as the base case, but the query indexed both the natural language request and the negotiated Boolean queries, treating the Boolean queries as pure text (the Boolean operatives and stemming indications were ignored). This had a major impact due to the vocabulary of the indexed query, and a minor impact due to the weights of the indexed terms (there is a term frequency contribution to the query weights, so repeated important terms got a slightly higher weight). Retrieval performance was much better with these added terms, improving by over 50%.

The fourth official run, SabL07ab1, used the SabL07arbn run as an initial run, and then expanded the query using Rocchio blind feedback. The top 30 documents of the initial run were declared "relevant" (without looking at them) and the top 20 Rocchio-weighted terms from those documents were added to the query. In general, the blind feedback helped moderately on some topics (top two topic improvements were increases of .14 and .08 in est-RB). However, on topic 63 there was a strong decrease of .52 and on topic 60 there was a decrease of .22. These were the two topics with fewest number of relevant documents in the topic set with 10 and 11 relevant documents respectively, and clearly the top 30 documents did not describe the notion of relevance well in these cases. Overall, except for those two topics the positive and negative results of blind feedback were about the same. Thus, there is no reason to use blind feedback of the sort used here - other expansion techniques need to be examined.

3.1.1 Comments on Ad Hoc Evaluation

I have a couple of rather hand-waving side comments on the evaluation measures used in this track and comparisons against the Boolean run (a major goal of this year's track), and other ranked runs. In my view, the comparison between Boolean and other runs still had two problems this year, though better than 2006. The first is in the definition of relevance that an assessor would use. The assessors were given both the original natural language request and the Boolean query statements as well as the complaint. This means their internal notion of relevance was partly formed by the Boolean statement. In addition to affecting the conceptual notion of relevance of the assessor, as has been shown above the vocabulary used in the Boolean statements was often different than in the request. Documents that might be ambiguously relevant given the natural language statement, may have their ambiguity resolved by the Boolean statements. Since the Boolean statements are defining relevance in these cases, the Boolean search will be favored. This can be easily fixed next year by just giving the assessor the original statement of information, possibly expanded from the minimal statement given this year, and not giving the assessor the Boolean statements.

The second problem is inherent in the nature of the Est-RB measure and I don't know of a good fix. When comparing two methods, it is unfair to allow one method to fix a parameter that the other method will be evaluated over. In this case, it is the parameter B, the number of documents the Boolean query retrieved. This is obviously the ideal point for the Boolean query to be evaluated at; all other points are non-optimal and have technical problems. However, there is no known theory for ranked retrieval of how to optimize at some random point B. If B were associated with some intrinsic property of the topic or collection, for example, number of relevant documents, then techniques might exist now, but that is not the case with B. For a ranked system, averaging the results at B over multiple topics is averaging results at different random points in the effectiveness curve of the system for each topic. It will be a non-optimal measurement.

Est-RB is even worse for comparing two ranked systems in that there are two mismatches of points on the operating curves of the systems. There is much more variability in such a comparison than in using measure such as MAP or even R-precision based measures which attempt to measure the two systems over either the entire curve, or the same point of the cure, respectively. Given enough topics, the effects will average out, but it's unclear whether there are enough topics in the Legal collection. For comparison of two ranked systems, I would suggest that MAPJudged is probably the best single measure, though it has its own problems.

3.2 Legal Relevance Feedback Task

The Relevance Feedback task was an initial pilot task exploring how relevance information can be used to improve retrieval performance. The TREC-2007 task was run using TREC-2006 Legal track topics and relevance judgments - groups could use any or all of the information contained in the documents judged in the 2006 track to modify their query for submission in 2007. The modified queries were run by the groups, submitted, documents

judged in 2006 were removed from the runs by the track coordinators, and the runs were evaluated using new judgments. Groups generally submitted all the 2006 topics, but because of assessing resources, only 10 topics were chosen for the evaluation stage.

The overall goal of any relevance feedback task is to try and capture the essence of known relevant documents without describing those documents so well that only the known relevant documents are retrieved. The basic tool used for the Sabir runs was smart_retro, a tool that attempts to find optimal weights for a term set, ignoring over-fitting problems. Over-fitting was addressed by limiting the choice of terms over which to optimize weights. The smart_retro algorithm and properties, including upper bound experiments, are described below, and then the adaptation to the Legal track is described.

3.2.1 Smart_retro

Given a weighted doc collection D, a subset of D of known relevant documents R, and a set Q of terms, then smart_retro finds weights for Q so as to maximize MAP when ranking D with an inner product similarity function over Q. Note this is a retrospective weighting, the same documents are being used for both determining weights and for evaluation of those weights.

The algorithm is an adaptation of Dynamic Feedback Optimization presented in 1995[1]. It's a very simple hill-climbing algorithm that makes changes in weights of terms in a round-robin fashion, keeping a change in the weight of a term if it improves MAP. The challenge of the algorithm is to make it robust and fast enough.

The algorithm used here is a five pass algorithm. The initial query weights are normalized to sum to 1.0, and an inner product document similarity is calculated for all documents in D. In the first pass, conceptually 0.3 (in practice it is 0.3 / (0.7 - currentweight) due to later length normalization) is subtracted from the weight of the first term and a new similarity for all documents in D can be calculated by just going through the documents in the current term's inverted list. If MAP using the new similarities is higher than it was previously, then the new weight is kept, and all weights and document similarities are renormalized so that query weights again sum to 1.0. Then the same process is tried adding 0.3 to the term instead of subtracting. Keep on repeating this process for each term in Q in turn (starting over with the first term if needed) until the entire set of terms in Q has been gone through without increasing MAP. Then the first pass is over. Each successive pass takes as its increment, the previous increment multiplied by 0.1. Thus pass 5 considers changes of 0.00003 in weights.

3.2.2 Smart_retro Upper-bound Performance

As part of a much larger query weight investigation, not yet completed, smart_retro has been extensively run on the TREC volumes 4 and 5 collection. In that investigation, the 125 odd topics were used. The retrospective performance varying the number of terms used in the query was measured on three different versions of the indexed collection.

1. the normal tf-normalized weighted documents (Lnu in SMART nomenclature)

2. tf*idf weighted documents (Ltu)

3. tf-normalized weighting with rare terms removed (Lnu with terms occurring in 5 or more documents).

For each topic, the terms occurring in the relevant documents were ordered by decreasing Rocchio weight (average weight in the relevant documents minus average weight in the non-relevant documents for that particular collection).

TREC V45 Upperbound MAP scores with Smart_retro			
Terms Added	Lnu Collection	Ltu Collection	Lnu >= 5 Collection
50	0.7696	NotDone	0.7662
100	0.8474	0.8450	0.8396
200	0.9102	0.9098	0.8983
400	0.9534	0.9546	0.9466
800	0.9789	0.9813	0.9760

The upper-bound performance of smart_retro is quite good, and there is very little difference between the collection variations. For these cases, it's clear that effectiveness does not depend on the document weighting scheme, or on the presence of rare terms. Efficiency is also good - it takes 2 to 3 minutes per topic for optimizing an 800 term query on a cheap PC.

Other experiments show it is also quite robust when varying starting conditions. In one experiment with 50 term queries, 24 different starting conditions were examined, some using random term ordering of Q and random assignments of (un-normalized) weights going from 1.0 to 50.0. The majority of topics had a maximum difference in MAP scores of less than 0.04. Only 5 of the 125 topics had a maximum MAP difference of more than 0.2; on those topics the relative orders of pairs of terms make a difference.

Thus in practice the optimization performed by smart_retro appears to be robust, efficient, and, as far as we know, near optimum. It distinguishes the relevant documents from the non-relevant documents very well.

3.2.3 Smart_retro in TREC-2005 Robust Track

Smart_retro was used in the TREC-2005 Robust track [2]. 50 term topics were constructed using relevance information on the TREC V45 collection, and then run and evaluated on the new AQUAINT collection. The overall result was only average; there were lots of highs and lows. Examining the low performing topics, several were due to non-obvious effects of overfitting. There were topics where it was easy to get near perfect retrieval performance using only 20 terms. But since all topics were expanded with 50 terms, those easy topics had tremendous degrees of freedom in assigning weights. There were near-infinite numbers of perfect retrieval solutions, most of which ranked the non-relevant documents very differently. These queries did not, in general, perform well on the new collection which had new non-relevant documents.

3.2.4 Smart_retro in TREC-2007 Legal Relevance Feedback

The approach used in this year's Relevance Feedback task was to use smart_retro and try to explicitly avoid overfitting by reducing the number of terms in Q when overfitting was suspected. This was done by splitting the Legal collection in half into even and odd numbered documents. Smart_retro was run on the odd collection to form an optimized query for the relevant documents in the odd collection. Then the optimized query was run on the even collection, and results were (automatically) examined for overfitting.

More precisely, given a topic, for N varying from 5 to 50, optimize adding N terms to the original query on the odd collection, getting a retrospective MAP_{N-odd} score. The optimized query was run on the even collection getting a non-retrospective MAP_{N-even} score. N_{topic} was chosen for each topic so that $MAP_{N-odd} + MAP_{N-even}$ was maximized on that topic. Once N_{topic} is chosen, smart_retro is run again, on the full Legal collection with full 2006 judgments in order to construct the final N_{topic}-term weighted relevance feedback query.

Sabir submitted three official relevance feedback runs, all based on smart_retro relevance feedback.

Legal Relevance Feedback Runs and Results. 10 topics				
Runname	Est-RB	MAPJudged	Est-rel-ret	Description
sab07legrf1	.2778	.3807	1660	Base case. Added 50 terms for all topics and optimized weights. Accidentally used optimization over odd collection instead of full collection.
sab07legrf2	.3327	.4303	3175	Base case. Added 50 terms for all topics and optimized weights
sab07legrf3	.3212	.4296	3617	Added query dependent N_{topic} terms to each topic and optimized

After submitting the initial run, sab07legrf1, it was discovered that it had been based on optimizing over the odd half of the collection rather than the entire collection. The second run, sab07legrf2, corrected this. It was gratifying that sab07legrf2 was better than sab07legrf1. However, the third run, sab07legrf3, was the only run where the worries about overfitting were considered. There basically was no difference between that run and the base case second run.

Looking at the results in more detail, only 10 topics were chosen for evaluation. All of those topics were well behaved topics with a reasonable number of relevant documents in the TREC-2006 task; one suspects they were chosen by that criteria in order to ensure there would be enough unjudged relevant documents to evaluate in TREC-2007. In the TREC-2005 runs, the problem topics were those with fewer relevant documents, which basically was not tested here. In this track, N_{topic} was less than 42 (out of a possible 50) on only 3 of the 10 topics. Of those 3 topics, one had a minor gain, one was even, and one had a minor loss using N_{topic}. There is not enough information to know whether the overfitting avoidance methodology of sab07legrf3 is worthwhile in general.

Comparisons against other non-Sabir runs are difficult in the absence of enough topics for reliable averages, and in the absence of reliable evaluation measures. Run sab07legrf2 was 10% better than any non-Sabir run using MAPJudged, but was only average if a single topic (26) was dropped. It was 14% worse than the best non-Sabir run using Est-RB, but beat that run if topic 51 was dropped.

Looking at individual topics, topic 30 had the lowest performance compared to the median of all runs with Est-RB, and topic 13 had the second lowest performance. Those were also the two topics with lowest number of new relevant documents that any group found in TREC-2007, with 7 and 19 new relevant documents respectively. One possible explanation for the poor performance is that TREC-2006 found most of the relevant documents for those topics. Perhaps the few relevant documents left undiscovered were the difficult to find relevant documents, i.e., those that looked different from the other relevant documents.

Further evidence of this comes from looking at the Est-rel-ret measure. Using Est-RB, there was little difference on average between the aggressive relevance feedback runs of CMU and Sabir and the original run continuations of Hummingbird and the original Boolean query. However, there were large differences in Est-rel-ret, for example the Boolean run had score 1413, and the sab07legrf3 run had score 3616. For those topics with a lot of new relevant documents to be found, the aggressive feedback runs performed much better than the Boolean or continuation runs. For those topics with few new relevant documents, continuing on with the original search, whether vector or Boolean, performed better.

3.3 Legal Interactive Task

The goal of the Interactive task was to find new relevant documents manually for a small number of topics (a subset of those used in the Relevance Feedback task). The evaluation measure was to score 1 point for each relevant document submitted, and deduct 0.5 points for each non-relevant document submitted.

For my participation in the task, I decided to find as many relevant documents as possible in 15 minutes per topic, by doing iterative relevant feedback, judging 10 new documents per iteration, and starting with relevant feedback on all TREC-2006 judged documents.

I used basic Rocchio relevance feedback, with a hacked together user interface that presented the current query in an editable window, and 10 large snippets of documents in a scrollable window. Each document had a pointer to the full document, in case that was needed (and it often was). Each snippet/document was judged Relevant, Non-relevant, or Undetermined. The only human actions performed were judging documents, and adding or deleting text from the natural language query. No new iterations were started after the 15 minutes from the time the initial topic was automatically submitted.

Overall, it was incredibly frustrating to judge documents with this interface. The problem was not that the task was uninteresting or the interface was that poor, but that there were major limitations of the relevance feedback algorithm used. For half of the topics done (4 of 8), there were large numbers of duplicate documents and form letters, mostly non-relevant, that needed to be judged. Interactively, I was eventually able to escape the duplicate documents by either going through them all or altering the textual topic, but

for those topics, most of the effort was still spent dealing with duplicates and not on any intellectual effort judging. This has implications for both automatic relevance feedback algorithms and interactive relevance feedback interfaces that will be discussed later.

Legal Interactive Topics and Results						
Qid	Topic	Num Seen	Num Sub	Num Rel	Num NonRel	Score
7	G movie placement	50	15	3	12	-3.0
45	pigeon deaths	70	47	11	36	-7.0
51	memory loss	50	1	0	1	-0.5
8	live theater	150	1	0	1	-0.5
13	candy cigarette	172	2	1	1	0.5
26	San Diego prices	60	21	13	8	9.0
27	calif prop placement	80	51	50	1	49.5
30	Cartwright act	40	7	1	6	-2.0
Totals		672	145	79	66	

Over the 8 topics done, I examined 672 documents, and submitted 145 of those documents as relevant. Of those 145, the assessor judged 79 as being relevant and 66 as being non-relevant. My overall scores per topic were low in comparison with the other two groups, though I was the only group submitting more than 3 topics. The other groups had stronger agreements with the assessor and/or were more active in weeding out documents with possible disagreements.

3.3.1 Interactive Disagreements

Historically, given 2 assessors per topic and a large pool of judgments over many topics, the expected overlap of their relevant document sets is only about 40–60%.[6, 5, 4, 7]. About half the topics have little disagreement, but some have massive disagreements. Despite these massive disagreements, all investigations have concluded the disagreements make no difference when comparing systems over all topics.

The disagreements come (mostly) from blunders and scope:

- Blunder: a differing assessor would agree they made a mistake
- Scope: One assessor more lenient than the other on some aspect

My personal estimate on TREC newswire is 5% "blunder" rate per assessor on the union of the two assessor's relevant judgments. Most of the rest of the disagreements come from scope disagreements.

For the interactive task, I re-examined all documents that I submitted, but the assessor judged non-relevant. I rejudged each document as Relevant, if I thought the assessor would agree, as Maybe, if I considered it relevant, but I could see an argument against it, or Non-relevant if I though I had blundered in my original judgment.

Legal Interactive Disagreements						
		Assessor		Rejudged		
Qid	Topic	Num Rel	Num NonRel	Rel	Maybe	NonRel
7	G movie placement	3	12	0	12	0
45	pigeon deaths	11	36	14	21	1
51	memory loss	0	1	0	0	1
8	live theater	0	1	1	0	0
13	candy cigarette	1	1	1	0	0
26	San Diego prices	13	8	1	5	2
27	calif prop placement	50	1	0	0	1
30	Cartwright act	1	6	4	2	0
Totals		79	66	21	40	5

The documents rejudged as Relevant are those I believe can be called assessor blunders.

Obvious “Blunder” Example 1
All documents discussing or referencing the California Cartwright Act.
Doc gdc90e00 ...In the original complaint filed July 16, 1980,plaintiff charged Philip Morris with breach of contract, unfair competition, and violation of the Cartwright Act.After eighteen months of exhaustive discovery, thep raintiff amended his complaint and deleted all causes of action against Philip Morris exce t the Cartwriqht Act claim, ...
Doc mae78c00 ... False Claims Act (CaI. Gov’t Code 9*****-12655) (id. at 24); and damages equivalent to the State’s Medi-Cal expenditures for alleged Relief Rectuested: Prohibitory injunctive relief (id. At 3-24); civil fines and penalties under the IICA and the California action for violation of the Cartwright Act (?d. 1170-74) and one ...

Obvious “Blunder” Example 2
All documents to or from employees of a tobacco company or tobacco organization referring to the marketing, placement, or sale of chocolate candies in the form of cigarettes.
Doc vqg61e00 ... ”’ -’ . - your letter of April 18, 1961 (by Mr. Shigihara) and the proposed candy packages bearing CLD *****, KENT and PIE‘APaRT label facsir*gt;:iles as -they are proposed to be amended by .- overprint or substitution of such legends as ”8ubble Oum”, ”Chocolate Cigarettes” or like **********, we approve of the proposed packages and are willing,to grant you permission to market them in Japan, subject to the following conditionsi 1. Packages containing the bubble gum or chocolate cigarettes will conform precisely to the sample packages which ********** your letter and may not be changed without our written conse ...

Non-Obvious "Blunder" Example
All documents that refer or relate to pigeon deaths during the course of animal studies.
Doc gwp94f00 ...We have successfully conducted and met the protocol requirements for periodic three-month sacrifices over a now eighteen consecutive months of tobacco smoke inhalation in pigeons. ...
There were 7 copies of this 8 page document where "sacrifices" refers to intentional pigeon killings.

Maybe Example 1
All documents that refer or relate to pigeon deaths during the course of animal studies.
Any explicit pigeon autopsy was considered relevant by the assessor.
My Maybe documents included 5 implied autopsies (for example, liver or heart examinations). 4 proposals of explicit pigeon autopsies not yet happened. 12 proposals of implied pigeon autopsies not yet happened.

Maybe Example 2
All documents discussing, referencing, or relating to company guidelines, strategies, or internal approval for placement of tobacco products in movies that are mentioned as G-rated.
Doc alp15f00 ... You agree not to place the Products in any motion picture that is made for or intended for display on television or any motion picture intended for or likely to appeal to an audience under the age of twenty-one. ...
There were 10 docs with this exact text (2 different form letters). 2 (one of each form letter) were judged relevant by the assessor and 8 were judged non-relevant.

The last example here brings up an important issue. There are numerous textually duplicate documents in the collection, since copies of the same document can come from different sources or lawsuits. There are also large numbers of form letters in the collection (or at least in these topics), where the information making a document relevant or not is contained in the form part of the document. These documents are essentially duplicate as far as relevance goes, while not being actually textually duplicate. (There is also the 51 submitted documents for topic 27 which were all the same form letter, but the non-repeated part of each document contained information essential to relevance.) Of the 6 duplicate or essentially duplicate document sets occurring in the rejudged documents, 4 sets had some documents originally judged relevant by the assessor, and some documents originally judged non-relevant by the assessor.

Legal Interactive Disagreements within Essentially Duplicate Sets						
		Assessor		Rejudged		
Qid	Topic	Num Rel	Num NonRel	Rel	Maybe	NonRel
7	G movie placement	3	12	0	12	0
	Duplicate Group 1	1	4	0	4	0
	Duplicate Group 2	2	8	0	8	0
45	pigeon deaths	11	36	14	21	1
	Duplicate Group 1	2	3	0	3	0
	Duplicate Group 2	1	2	2	0	0
	Duplicate Group 3	0	7	7	0	0
	Duplicate Group 4	0	11		11	0

Some of these documents in these sets should be counted as blunders, since the single official assessor is judging essentially duplicate documents as different, but it's not clear how many.

The overall disagreement rate is a bit higher than historically expected (especially since the rejudging here is a one-sided rejudging – there is no good way to count the documents I thought were non-relevant, but the assessor would have judged relevant if the assessor had been given a chance to judge them). However, it's still within the ballpark of previous experiments. The blunder rate, though, is noticeably higher than historically observed. Possible explanations for that include

- Poor OCR or scanned images make it more difficult to judge
- non-newswire documents
- non-professional writers
- essentially duplicate documents increase detection of borderline relevant conflicts
- volunteer assessors instead of paid, though the assessors spent more time per document than the normal TREC assessors.

Even with the increased blunder rate, the number of assessing disagreements due to blunders is still much less than the number of assessing disagreements due to scope of relevance. The blunder disagreements may be problematic for some systems with learning algorithms, but they should not interfere with TREC-style comparison of systems which average enough topics so that disagreements which are independent of the systems involved will not affect rankings of systems.

Lawyers, though, are less interested in averages over topics, and more interested in worst case scenarios (since the worst case might be their case). Both the blunder and the scope disagreements need to be addressed in practice. The blunder rate is caused by uncertainty in document judgment; this uncertainty can be addressed by "good practices", such as ensuring all duplicate or essentially duplicate documents are treated the same. The scope error rate is caused by uncertainty in what the topic means. In practice, this should be theoretically

addressed by the two parties in a litigation. There needs to be agreement on what should be considered relevant. Historically, this agreement in information retrieval has to involve looking at documents in the collection; humans are quite poor at coming up with all possible borderline cases of relevance without documents.

3.3.2 Legal Relevance Feedback and Interactive Task Conclusions

It would have been very nice to have done the interactive task before the relevance feedback task. I wouldn't have spent nearly as much time worrying about marginal overfitting problems for relevance feedback, and focused substantial efforts on dealing with near-duplicate documents, and whether a particular set of near-duplicate documents can be treated uniformly! If they can, substantial improvements should be obtainable for both automatic relevance feedback runs, and the manual interactive runs. This is particularly important on the Legal collection with OCR documents. Unlike "perfect text" collections like the TREC newswire collections, duplicate documents in the Legal collection can have very different similarities to a topic due to OCR differences.

Unfortunately, it will be tough to detect whether a set of documents is essentially duplicate as regards relevance. The blunder rate is high enough so that relevance disagreements among nearly duplicate documents cannot be attributed to the differences in those documents. In just the 8 topics examined in the interactive task, there were two topics with disagreements among near-duplicates caused by assessment blunders, and one topic with the disagreements caused by the differences in the near-duplicate documents.

Overall, it's clear that relevance feedback algorithms need to be improved for both automatic relevance feedback and manual interaction in order to satisfy the requirements of legal discovery. The current algorithms which tend to try to classify all relevant documents in one class are not going to be appropriate in real-life. There needs to be provisions for faceted relevance feedback: dealing with different types of relevant documents differently.

References

[1] C. Buckley and G. Salton. Optimization of relevance feedback weights. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 351–357, 1995.

[2] Chris Buckley. Looking at limits and tradeoffs: Sabir research at TREC 2005. In E.M. Voorhees and L.P. Buckland, editors, *Information Technology: The Fourteenth Text REtreival Conference (TREC 2005)*, 2006.

[3] Chris Buckley, Amit Singhal, Mandar Mitra, and Gerard Salton. New retrieval approaches using SMART:TREC-4. In D. K. Harman, editor, *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*, pages 25–48, October 1996. NIST Special Publication 500-236.

[4] G. Cormack, C. Palmer, M. Van Biesbrouck, and C. Clarke. Deriving very short queries for high precision and recall multitext experiments for TREC-7. In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, pages 121–132, August 1999. NIST Special Publication 500-242. Electronic version available at http://trec.nist.gov/pubs.html.

[5] Gordon V. Cormack, Charles L.A. Clarke, Christopher R. Palmer, and Samuel S. L. To. Passage-based refinement (MultiText experiments for TREC-6. In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the Sixth Text REtrieval Conference (TREC-6)*, pages 303–319, August 1998. NIST Special Publication 500-240. Electronic version available at http://trec.nist.gov/pubs.html.

[6] Ellen M. Voorhees and Donna Harman. Overview of the fifth Text REtrieval Conference (TREC-5). In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the Fifth Text REtrieval Conference (TREC-5)*, pages 1–28. NIST Special Publication 500-238, November 1997.

[7] Justin Zobel. How reliable are the results of large-scale information retrieval experiments? In W. Bruce Croft, Alistair Moffat, C.J. van Rijsbergen, Ross Wilkinson, and Justin Zobel, editors, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 307–314, Melbourne, Australia, August 1998. ACM Press, New York.