A Constrained Approach to Manual Total Recall

Jeremy Pickens, Tom Gricks, Bayu Hardi, Mark Noel
Catalyst Repository Systems
1860 Blake Street, 7th Floor
Denver, CO 80202
jpickens,tgricks,bhardi,mnoel@catalystsecure.com

ABSTRACT

The Catalyst participation in the manual at home Total Recall Track was both limited and quick, a TREC submission of practicality over research. The group's decision to participate in TREC was made three weeks, and data was not loaded until six days, before the final submission deadline. As a result and to reasonably simulate an expedited document review process, a number of shortcuts were taken in order to accomplish the runs in limited time. Choices about implementation details were due primarily to time constraint and necessity, rather than out of designed scientific hypothesis. We detail these shortcuts, as well as provide a few additional post hoc, non-official runs in which we remove some of the shortcuts and constraints. We also explore the effect of different manual seeding approaches on the recall outcome.

1. PRIMARY METHOD

Given the manner in which Team CATRES approached this project there was no formal hypothesis, as such. Rather, the project was primarily an evaluation of the extent to which a constrained continuous active learning [1] tool can effectively assimilate and adjust the disparate skills and knowledge of multiple independent, time-pressured reviewers tasked solely with the obligation to expeditiously locate potential seeds to commence ranking. In that sense, the working hypothesis was that a continuous active learning tool, when combined with an initial seeding associated with tight deadlines, limited knowledge and experience, and potentially inconsistent perspectives, will produce a reasonable result.

The manual seeding effort itself was intentionally limited and necessarily relatively cursory. Three users each worked for no more than one hour apiece to locate potential seed documents based on their personal conjecture as to the potential scope of the terse descriptions for each topic. Within that hour, each had to (individually and separately) carry out all three aspects of the task: (1) familiarize themselves with the topic, (2) issue queries to find relevant information, and (3) read and mark that information for relevance. One of the three users was well-versed on the search tool and its capabilities, query operators, and syntax, but the other two users were essentially brand new to the system. All three users averaged between limited to no knowledge of the topics. Further details are given in Section 1.1.

After this work was completed, the system was set to an automated, continuous learning (iterative) mode with no additional human intervention other than the official judgments. Top ranked, as yet unseen documents were continuously (at least until time ran out) fed to the TREC server in batches, truth values from the TREC server were fed back in to the core learning algorithm, and then the remaining unjudged documents in the collection were re-ranked. Normally, iterative batch sizes would be constant, but given time constraints and in order to expedite the process, batch sizes were increased over time. Batch sizes started small to enhance continuous active learning (100 docs per iteration) and then were gradually increased (250, 500, 1000, 2000, and 5000) as the time deadline neared. Final batches were submitted just hours before the deadline.

Continuing in the theme of constraint, the algorithm underlying the continuous learning protocol was also constrained. These constraints included naive reduction of the feature space, extensive use of pseudo-negatives in training, and lack of explicit diversification. We will explore these and more constraints in greater detail in Section 1.2. We also present results of some non-official, post hoc runs in which we relax some of these constraints. Finally, we investigate the effect that different searchers (and the union thereof) have on the recall-oriented outcome.

1.1 User Involvement

The initial, manual stage was conducted in the following manner: All three people worked for one hour each, yielding a total of three person-hours per topic. All three team members worked independently of each other and at different times in different geographic locations and no information was communicated about the topics or topic-related searches between the three team members. No restrictions on the strategies or methods that each adopted were stipulated other than a request to set a filter on one's searches to remove documents that had already been marked by another user. However, due either to time or lack of clarity in communication, this request was only followed by reviewer 1 and not by reviewers 2 and 3. However, as reviewer 1 completed his work before reviewers 2 and 3 began, this has the net effect of zero effort deduplication, i.e complete independence and the possibility of duplication of effort. Thus, each reviewer was free to work as he wished, with activities including (1) researching the topic, (2) searching for relevant information, and (3) reading and marking documents for relevance. Some team members spent more time researching, others spent more time marking, but each person had only a single hour, per topic, to do all activities.

The reviewers self-reported the following rough breakdown in time spent: Searchers 1 and 2 spent no time researching the topic, other than 15 and 45 minutes (respectively) consulting external resources on topic 109. Otherwise, half of

	Reviewer 1		Reviewer 2		Reviewer 3		Sum		Union		Diff	
Topic	Rel	Nonrel	Rel	Nonrel	Rel	Nonrel	Rel	Nonrel	Rel	Nonrel	Rel	Nonrel
athome100	30	55	27	109	5	2	62	166	62	166	0	0
athome101	75	29	55	10	26	13	156	52	156	52	0	0
athome102	91	1	51	13	20	6	162	20	149	20	13	0
athome103	199	13	199	13	16	2	414	28	214	15	200	13
athome104	67	69	27	17	27	12	121	98	95	88	26	10
athome105	150	1	107	0	34	0	291	1	291	1	0	0
athome106	–	–	100	0	26	16	126	16	126	16	0	0
athome107	129	3	49	0	33	2	211	5	211	5	0	0
athome108	168	2	70	0	23	1	261	3	261	3	0	0
athome109	208	11	76	2	13	6	406	19	297	18	109	1
average	124.2	20.4	76.1	16.4	22.3	6.0	221.0	40.8	186.2	40.4	34.8	1.6

Figure 1: Reviewer Document Judgment Counts

their time was spent searching and half of the time reviewing and marking documents, with the searching and reviewing activities done in an interleaved manner. Each of these two reviewers averaged 3-5 primary queries per topic, with 3-5 refinements of the primary queries (addition or deletion of a word or query operator) for a total of approximately 6-10 searches per topic.

Reviewer 1 explicitly attempted to diversify his efforts, never spending too much time pursuing any one topic, going wide and broad with his query selection rather than narrow and deep. This reviewer judged documents solely to determine whether they served the cause of diversity. Reviewer 2 took a more standard approach, where the purpose of each query was to find relevant documents, not diverse documents. Reviewer 3, on the other hand, spent the first 15 minutes researching the topic, the next 30 minutes composing a "synthetic document" with multiple passages of information relevant to the topic (which passages were gleaned from web searches), and the last 15 minutes reviewing and marking results from having used this synthetic document as an initial "pseudo-positive" seed document. No additional queries were issued by Reviewer 3 while in this last stage.

During the interactive querying and relevance marking (manual) session for each topic, reviewers 1 and 2 chose to not avail themselves of the official TREC relevance judgments. Instead, they decided upon a subjective assessment of relevance and use that subjective assessment to guide query formulation and development. Thus, no official judgments influenced either of the first two reviewers during their interactivity. Reviewer 3 did look up the official relevance judgment as he was examining every document, but because document examination start only at the very end, after research and the single pseudo-positive seed document "query", this also has the effect that no official judgments influenced reviewer 3 during any query formulation stage.

Once all three reviewers were finished working on a topic, every document that any reviewer had examined but that had not yet received an official TREC relevance judgment (i.e. had not yet been submitted to the Total Recall server to be recorded as official effort for the run) were submitted to the server, and then the continuous learning stage was kicked off with no further human intervention.

One fact mentioned earlier requires further elaboration. No controls were made for ensuring that the three reviewers didn't duplicate their effort. Thus, during the course of each of their one hour of work, reviewers may have unintentionally judged the same document twice, completely unaware of the judgment that a previous reviewer had given to that document. In our official TREC submission, we submitted the union of all documents that all three reviewers manually found, every relevant as well as every non-relevant document. As per task guidelines, however, it may have been more correct to submit the same documents twice, i.e. to have done the sum of all documents rather than the union. To this end, we offer the following statistics in Figure 1. This chart shows the number of documents found, relevant or non-relevant, per reviewer per topic. The values are the official TREC judgments. It also shows counts for the sum, union, and (sum - union) difference of all three reviewers.

The fact that the sum equals the union in 6 of the 10 topics means that our official run is completely accurate in terms of representing the full amount of manual effort done. Where the sum is greater than the union, it means that reviewer effort was (unintentionally) duplicated. However, as every uniquely reviewed document, both positive and negative, was submitted to the TREC server, this has no effect on the basic shape of the gain curve. Instead, it should just shift that curve a few documents to the right, anywhere from a shift of 0 for most topics, a shift of 13 documents for Topic 102, and a shift of 213 documents for Topic 103.

Of note is the fact that duplication of relevant effort (different reviewers finding the same relevant documents) was relatively much more common than duplication of non-relevant effort, as well as the slight tendency for this to happen more on sparse topics (e.g. 104 and 109). More could be written along these lines, but we save that for future work.

We should also briefly note that all three reviewers had high levels of experience running search-based recall-oriented projects in the past, so all felt comfortable in the TREC task. However, reviewer 2 had only used the specific tool a small handful of times, and reviewer 1 had never used the tool before. As such, halfway into running his first topic (106), reviewer 1 felt that he had made some mistakes based on his unfamiliarity with the tool that he felt disqualified his effort from the task. Fifteen minutes into running that topic he stopped and removed all traces of his effort (marked documents). He communicated nothing of what he did (or did not) learn to the other team members. Then reviewers 2 and 3 each worked for 1.5 hours apiece rather than their usual 1 hour, stretching out their individual strategies to proportionally fill the time.

No team member had more than limited prior knowledge

Topic	Reviewer 1	Reviewer 2	Reviewer 3
athome100	limited	no	limited
athome101	limited	no	limited
athome102	limited	no	limited
athome103	limited	no	no
athome104	limited	no	no
athome105	limited	limited	limited
athome106	–	limited	limited
athome107	limited	no	limited
athome108	limited	no	no
athome109	no	no	limited

Figure 2: Reviewer Prior Topic Knowledge

(watching or reading the news, conversations with friends) about any topic. Reviewer 1 had limited, generic knowledge of a number of the topics, but no specific knowledge, and no knowledge whatsoever on the Scarlet Letter Law, Topic 109. For Topic 109, reviewer 1 spent 5-10 minutes of his hour reading Wikipedia on the Topic. The only topics that reviewer 2 had some knowledge on were topics 106 (Terri Schiavo) and 105 (Affirmative Actions). He used Google and Wikipedia to research topic 109 (Scarlet Letter Law) for 45 minutes. He did not have a lot of knowledge on the remaining topics, but felt that he got a general idea from looking at the topic names. Again, however, we note that he viewed no official assessments during the course of the hour. Reviewer 3 had some previous knowledge of many of the topics through exposure to news articles and the like. As mentioned previously, for all ten topics he started by assembling a synthetic seed document by using Google for about 30 minutes per topic to find wikipedia and news articles with sections of text that looked relevant.

The exact breakdown of prior by topic is found in Figure 2.

1.2 Algorithmic Constraints

In addition to the time constraints placed on the human reviewers, our official run was implemented with number of algorithmic constraints as well. Some of these constraints were intentionally done so as to speed execution time, some constraints unintentionally occurred due to mistakes associated with last minute haste.

The intentional constraints were two-fold: Features and ranking. On the feature side, only unigram text was extracted. Not extracted were dates, email communication information, named entities, and so on. Moreover, naive feature reduction was done by ignoring unigrams that were either too frequent or not frequent enough. Unigrams with a document frequency under about a dozen, and over about ten thousand were simply ignored. This naive approach sped iterative simulation – important given the time constraints – but might have come at a bit of an effectiveness cost, one that we anecdotally explore in Section 3.2.

The second intentional constraint that was done to speed limited iteration time was to pre-compute a "pseudo-negative" score across the entire collection and then only train on positive documents. In a tradeoff between speed an effectiveness, experiments on other datasets have shown that speed can be increased often without losing too much effectiveness. This bore out for some topics, but turns other topics into relative failures, and will be investigated further in Section 2.

There were a number of unintentional constraints as well. The first was that the the algorithms were not run fully continuously. This was simply due to running out of time; iteration necessarily stopped with the TREC deadline was hit. As that deadline was approached, topics were actively monitored as time ran out, and batch size was manually increased, subjectively and non-uniformly, based on iterative richness.

The second unintended constraint was that no algorithmic diversification was performed. Typically we have found various forms of active diversification useful in recall-oriented tasks, but unfortunately a bug in the hastily written code to interface between the algorithm and the TREC server failed to submit any diversification results. Finally, the third unintended algorithmic limitation is that another server interface code bug ended up submitting one hundred documents to Topic 101 that had actually been selected via Topic 100, i.e. the wrong documents had been submitted. This added slightly to the total cost for topic 101, but post hoc resimulation showed that overall effect was negligible.

2. ADDITIONAL RUNS

The previous section outlines in large brush strokes the primary methodology used for our official manual total recall run. After the official run was submitted we did a number of post hoc, non-official runs to remove the primary run's constraints and to test the effect of various starting points. The first non-official run removed the unintentional constraints of (1) having submitting the wrong documents to the wrong topic, and (2) with more time in which to operate we increased iteration continuousness until greater than 90% of the relevant documents were found for each topic. On the intentional constraint side, we used true negatives rather than pseudo-negatives. The second non-official run examines the effect of relaxing some of the feature frequency and feature length constraints. Frequency-wise, a wider range of terms with a document frequency between 3 and 50,000 were allowed. And term length was expanded beyond unigrams to bigrams and trigrams. Section 3.2 discusses one topic for which these expanded features were critical. We call these three runs (one official plus two non-official) "HC", "MC", and "LC", for highest constraints, medium constraints, and lowest constraints, respectively.

The third non-official run examines reviewer effect. Using the newer, less-constrained algorithm we then compared the effect of various seeding options. Specifically, we compared the effect of the initialization of the runs using the contributions (manual search effort) of each of the three reviewers, separately and individually, against the union thereof. Some reviewers found a higher ratio of relevant-to-nonrelevant documents, some found a lower ratio. Some reviewers found more total documents, some found fewer.

3. RESULTS

The first thing that we should note is that due to the last minute nature of our effort, and software that was still being written even as the experiments began, we failed to capture the exact order in which documents were judged. It is our understanding that this information will be made available by TREC at some point, but as we did not have access to that information at the time of this writing, we decided to do a re-simulation of the primary run. We knew which documents the reviewers manually examined during their sessions, so using those same documents as starting points,

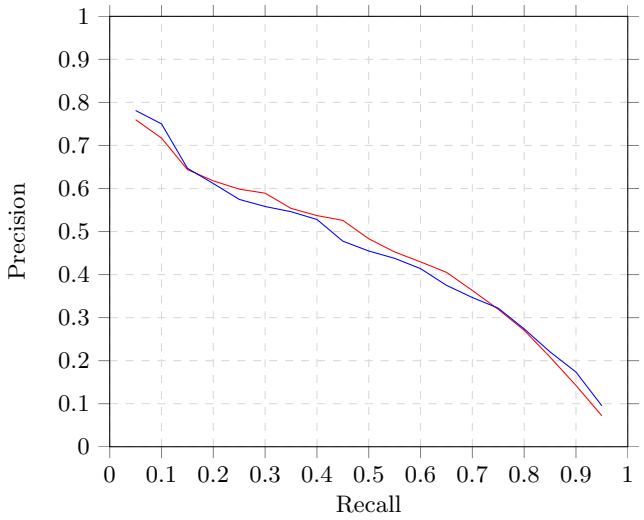

Figure 3: Precision-Recall curves averaged across all ten athome1 topics. Official run (red line) and re-simulated run ("HC") with similar parameters (blue line).

we set the same (fully constrained) algorithmic parameters and ran the simulated review in the same constrained manner (i.e. learning iteration ceased after a certain number of documents, approximately at the same point as was done in the official run). Figure 3 shows a Precision-Recall curve of an average of all ten athome1 topics, with the red line as the official TREC result and the blue line the re-simulation. The results were close enough that we felt comfortable using the re-simulation's seen document ordering as an honest substitute to the official run.

Figure 5 shows the results of all ten athome1 topics in the form of gain curves, with percentage of collection judged on the x-axis and recall on the y-axis. The two black lines show the gain curves of an expected linear review and a theoretical perfect run (not a single non-relevant document viewed). The red solid line is the (re-simulated) official run with full (highest) constraints ("HC"), the blue dashed line is the secondary run in which the main constraints of pseudo-negativity and non-continuousness were removed ("MC"), and the green line the tertiary run in which document frequency constraints were removed and word bigrams and trigrams were added ("LC").

The results of the manual individual versus the joint seed set are found in the gain curves of Figure 6. The joint seeding (i.e. the union of the manual effort of all three reviewers) is in black, and the various individual reviewers are in red lines of various patterns. The x-axes of these curves are not shown; the purpose of the visualization is to compare the relative rise between the various starts. As such, the data is scaled so that the range in which recall rises from 0% to approximately 90% is clearly shown. This range is, of course, different for topics of different richness levels and system performance. But again, the purpose of Figure 6 is to visualize the effect of various starts, not of absolute performance levels.

3.1 Discussion

3.1.1 *Main Experiment*

From Figure 5 we see that the primary, constrained approach was, with the exception of Topics 105 and 108, fairly reasonable. On Topics 101, 103, 104, and 109 the highest constrained HC approach (red line) is essentially on par with the medium constrained MC secondary approach (dotted blue line), with HC even slightly outperforming MC at mid-range recall levels. And up to about 50% recall there is little difference between the two approaches for almost all of the topics.

Of course, the goal of this task is Total Recall and it is at higher recall levels that the constraints become a hindrance. A cursory post hoc analysis found that, where the MC diverged from HC was at a point at which continuous active learning was still operating for both approaches. Thus, while non-stop continuousness undoubtedly helped the secondary approach, the more important constraint seems to be the use of pseudo-negative weighting versus true negative weighting.

The least constrained LC approach (green line) is even better still for a majority of topics. On the "problematic" topics (e.g. 105) there is a massive improvement over not only HC but also over MC. On most of the remainder of the topics, LC continues to outperform the other two approaches. We note, however, that on a few topics such as 104 and 108, the LC approach, while outperforming next best alternative up to about 95% recall, drops off in effectiveness past that point and does not seem to recover. Perhaps the additional higher order features (bigrams and trigrams) increase the chance of overfitting.

3.1.2 *Individual Reviewer Experiment*

The individual versus joint seeding experiments shown in Figure 6 were all done under the MC constraints. The results here are an interesting first step. On the one hand, the overall magnitude of the differences are quite small. At 50% recall, the average difference across all topics between the lowest and highest performing reviewers is 671 documents (std dev 495), and at 90% recall the average difference is 851 documents (std dev 797). In terms of practical impact, the difference is very low. Furthermore, no matter how the process is seeded (even if there is larger or smaller variation along the way) all starts converge to relatively high recall at approximately the same point. This is not too much of a surprise, as a number of researchers [2, 3, 4] have tested the "single seed hypothesis", i.e. the notion that high recall can be achieved with a single relevant seed and continuous iteration. Thus, no matter if one seed, a dozen seeds, or a few hundred seeds are used, high recall can be achieved.

On the other hand, when there is a difference between starting points, that difference seems to come about by one of the "bolder" approaches. Recall from Section 1.1 that Reviewer 1 and Reviewer 3 each took "extreme" (or as extreme as one can get with a single hour's worth of work) approaches. Whereas Reviewer 2 ran straightforward relevance-seeking queries, Reviewer 1 explicitly tried to make each query as different as possible from the next query, and marked as many documents as possible, taking "scattered buckshot" approach to the task, quickly marking only those documents that appeared to be diverse from previously marked documents. Evidence of this is found in Figure 1, in that Reviewer 1 marked an average of 144.6 documents per topic

in his allotted hour. Reviewer 2, on the other hand, only marked 92.5 documents on average in each hour.

At the other extreme, Reviewer 3 spent the majority of each hour (45 minutes) reading and learning about the topic itself and constructing a pseudo-relevant seed document filled with information gleaned from external sources. It was only in the last 15 minutes of each hour that Reviewer 3 even started engaging with the collection, and then not even to run any additional queries beyond the single, heavily researched initial query that had been created during the first 45 minutes of research. This approach is also reflected in the documents counts: Reviewer 3 marked on average 28.3 documents per topic, over five times fewer than Reviewer 1.

Now, against this backdrop, examine Topics 104 and 109. These were the two sparsest topics in the athome1 set. And Reviewer 1's approach of "buckshot" trying to hit as many targets as possible yielded an approach that not only found more rare documents initially, but kept the lead ahead of the other two approaches up past 70% or 80% recall. On the other hand, Topics 101 and 103 are two of the richer topics, and also happen to be ones in which Reviewer 3's single seed, deeper investigatory approach, while slow in its start, caught up and surpassed the other approaches. Topic 102 is the only one for which Reviewer 2's approach maintained a lead. Reviewer 2's approach can be characterized as somewhere between Reviewer 1 and Reviewer 3 in that the quantity and diversity of initial seeds was greater than Reviewer 3, but less than Reviewer 2. As it so happens, Topic 102 also lies between the other topics in terms of its richness. For the other half of the topics (100, 105, 106, 107, 108) there was little difference between the three approaches. This relationship between topic richness versus the reviewer whose seeding approach fared the best can be found in Figure 4 below:

Topic	Richness (#rel)	Best Reviewer Seeds
106	17135	–
101	5836	Reviewer 3
103	5725	Reviewer 3
100	4542	–
105	3635	–
107	2375	–
108	2375	–
102	1624	Reviewer 2
109	506	Reviewer 1
104	227	Reviewer 1

Figure 4: Best Reviewer vs Topic Richness

We believe this suggests that seeding approaches that are more diversified, scattered, higher diversity may tend to work better on lower richness topics, while seeding approaches that are deeper and more investigatory may tend to work better on higher richness topics. However, these are post hoc explanations, and the number of examples that fit this pattern is so small that it could be due to random chance. Nevertheless, it may be worth further research into when the high diversity, high volume buckshot approach (Reviewer 1) versus the deep investigatory low volume approach (Reviewer 3) versus the middle of the road (medium diversity, medium volume) approach (Reviewer 2) yields the best results.

Perhaps the more interesting question is why the union of the three reviewer seeds did not fare well. With the exception of Topics 104 and 109, which were sparse enough that the manual phase alone (rather than the continuously iterative algorithmic phase) gave the union a recall boost, the union approach does not surpass, and sometimes even falls behind the best individual reviewer, such as in topics 101 and 103. The difference cannot be due to difference in reviewer relevance assessment, as every coding call was checked against the official TREC value before being used in any fashion. We have a few hypothesis, but that none that we're yet willing to commit to writing.

3.2 Failure Analysis

For a final bit of discussion we wish to engage in some brief failure analysis. The following observation came from Reviewer 1, who took the buckshot approach by issuing as many queries as possible and not spending too much time engaged with any one particular query. Even when a particular query yielded rich results, he moved on from that query to others, to ensure diversity.

For topic 105 (Affirmative Action), for which there were 3635 total relevant documents, two of the many queries that Reviewer 1 did were the two word *phrase* queries "affirmative action" and "one florida", the latter being the name of the Florida state government program related to Affirmative Action. Those two queries yielded a large number of relevant documents, but in the interest of "buckshotting" the process he moved on to other queries, leaving the remainder to the continuous, algorithmic process. However, had he stuck with those two queries, which would not have been unreasonable given that 150 of the 151 documents that he examined during that hour were relevant, a post hoc analysis shows that he would have achieved the following results with the following boolean queries:

1. "one florida" = 1678 relevant out of 2739 total hits (61.3% precision at 72.2% recall)

2. "affirmative action" = 799 relevant out of 971 total hits (82.3% precision at 34.4% recall)

3. ("one florida" OR "affirmative action") = 2123 relevant out of 3,337 total hits (63.6% precision at 91.4% recall)

Thus, with absolutely no additional machine learning technology, or even ranked retrieval, the simple boolean query ("one florida" OR "affirmative action") achieves a high 91.4% recall at a not unreasonable 63.6% precision, which was actually higher than our highest constrained HC approach.

Further analysis showed that the problem likely resided with feature selection. As mentioned in Section 1.2, the HC run did no higher order feature extraction, only unigrams. And in a further naive step, unigrams with a document frequency under about a dozen, and over about ten thousand were simply ignored. This seemed to pose no problem for a number of topics (e.g. 101, 103, 109), but for others such as 105, it was likely the cause of the large loss in fidelity. Case in point: Most of the words in the ("one florida" OR "affirmative action") were ignored by the model. The collection counts (document frequency) for these four terms are:

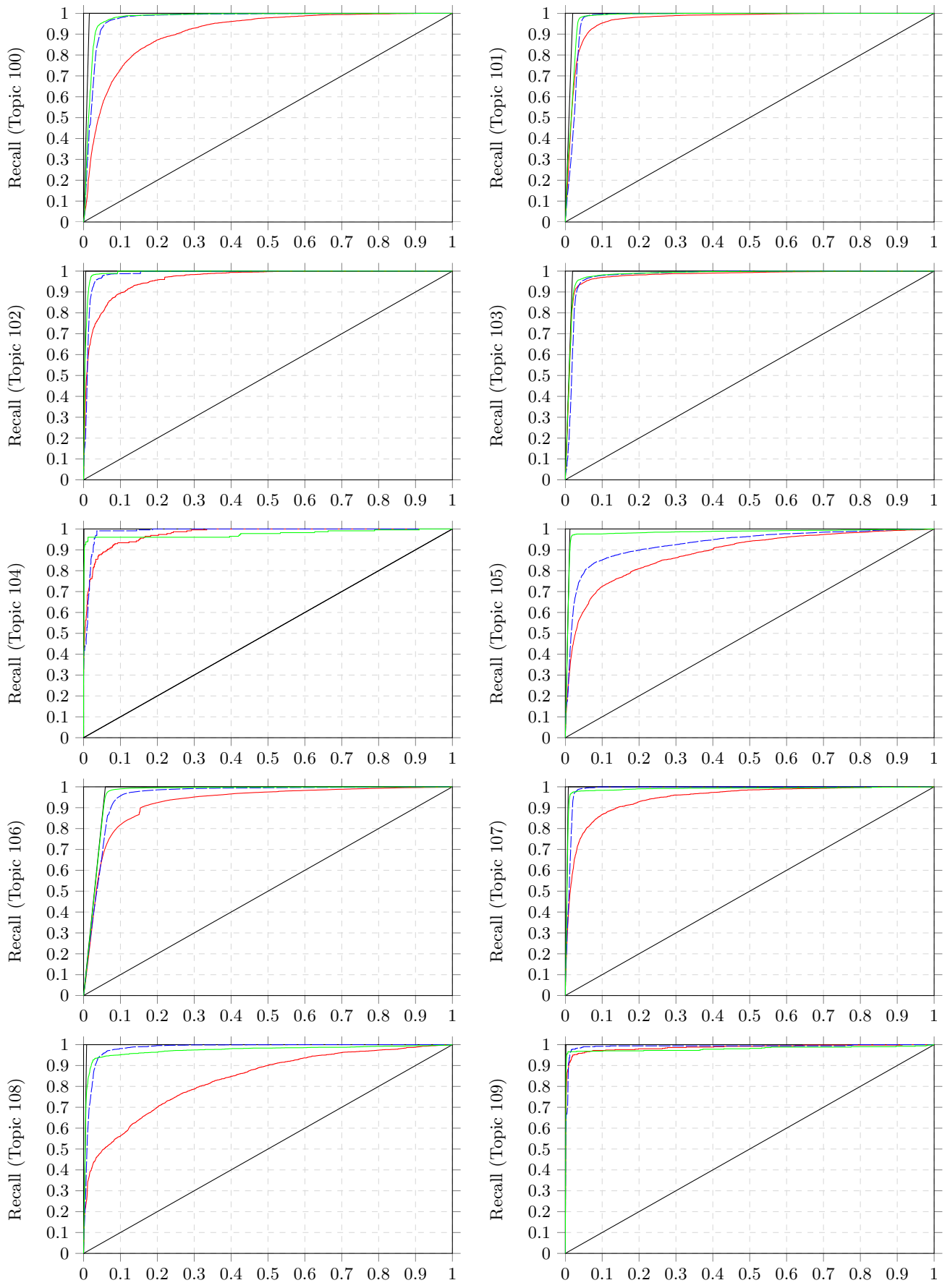

Figure 5: Gain curves for HC run in solid red, the MC run in dashed blue, and the LC run in green. Random and theoretical perfect curves given in black.

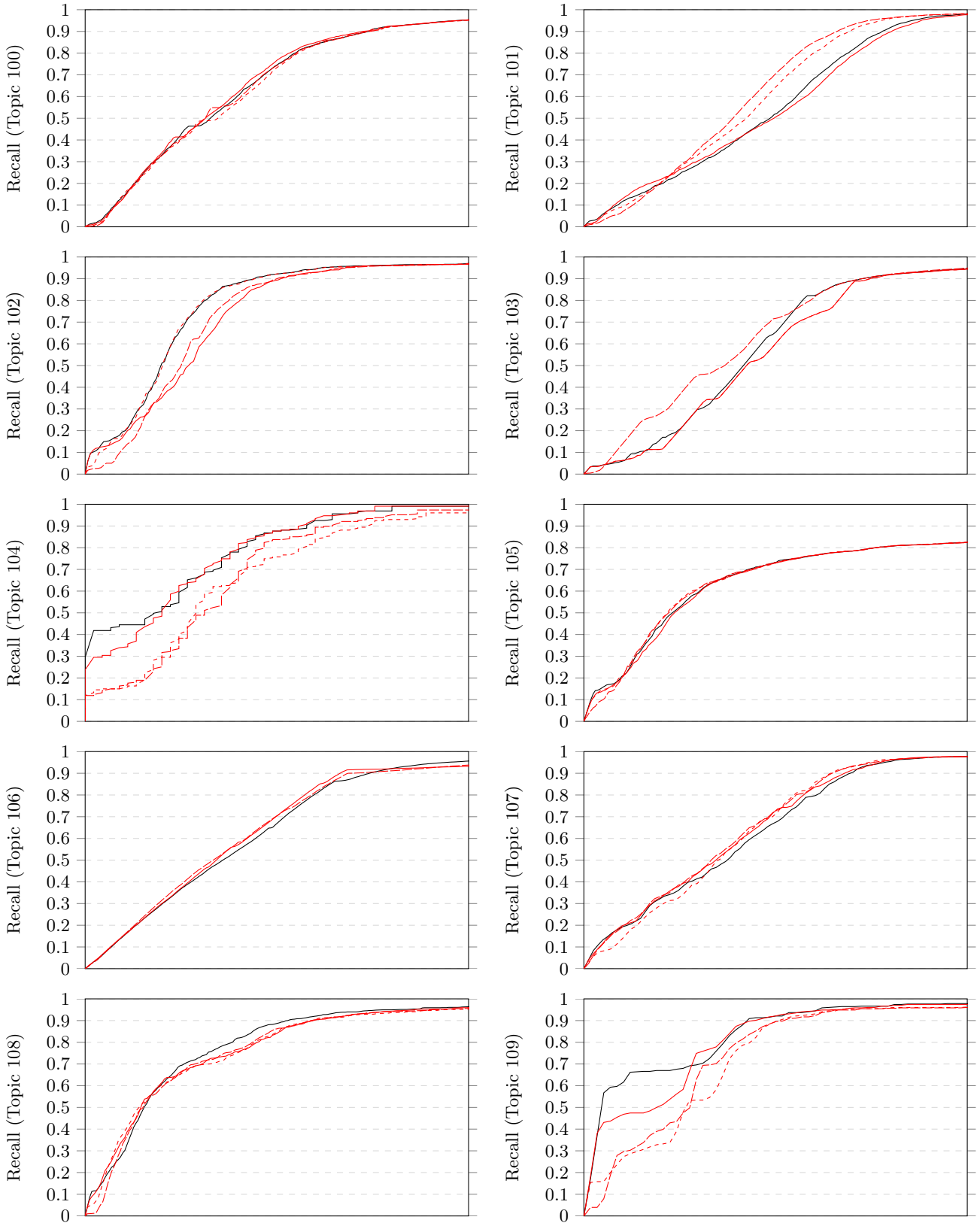

Figure 6: Gain curves for individual versus joint reviewer seeding. Reviewer 1 (solid red), Reviewer 2 (dotted red), Reviewer 3 (dashed red), Union (solid black).

1. one = 86,464
2. florida = 160,158
3. affirmative = 1,208
4. action = 16,887

Thus, all these terms other than "affirmative" were not available to the HC engine; they had been filtered out. We suspected that this is why Topic 105 underperformed, and was part of the motivation for doing additional runs with increasingly relaxed constraints. The additional experiments bore this fact out: The LC run in which the most constraints were removed – and specifically in which bigrams and trigrams were added – far outperformed the more limited approaches. To wit: For Topic 105, 95% recall under the HC run was achieved at rank 122,445, under MC at rank 97,112, and under LC at rank 4,332. Same exact initial manual seeds in all three approaches, just different levels of constraints in the automated continuous learning stage.

4. CONCLUSION AND FUTURE WORK

As mentioned at the beginning, the official run for Team CATRES had no formal hypothesis. Rather, it was an attempt to see how well a constrained continuous active learning tool could do at achieving high recall from an ad hoc, not formally organized set of reviewers tasked with manually seeding the process. Results with the highest constrained HC run were decent, though with failures on some topics (100, 105, 108) and successes on other topics (101, 103, 109). As constraints were removed and the continuous learning process re-simulated using the exact same set of manual reviewer seed documents, certain previously failing topics garnered massive improvements.

We also found that, no matter the starting condition – whether we started with a lot of documents found by Reviewer 1, over five times fewer documents found by Reviewer 3, or a union of all three sets of documents found by all three reviewers – high recall was achieved after approximately the same total document review cost. There is some slight variation in effectiveness between the various starting conditions at mid-range recall levels, not enough evidence to draw any conclusions but enough to begin to ask more questions.

5. REFERENCES

[1] G. V. Cormack and M. R. Grossman. Evaluation of machine learning protocols for technology-assisted review in electronic discovery. In *Proceedings of the ACM SIGIR Conference, Gold Coast, Australia, 6-11 July 2014*, Gold Coast, Australia, 2014.

[2] G. V. Cormack and M. R. Grossman. Autonomy and reliability of continuous active learning for technology-assisted review. *arXiv*, April 26, 2015.

[3] B. Dimm. The single seed hypothesis. `http://blog.cluster-text.com/2015/04/25/the-single-seed-hypothesis`, April 25, 2015.

[4] J. Tredennick, J. Pickens, and J. Eidelman. Predictive coding 2.0: New and better approaches to non-linear review. `http://www.legaltechshow.com/r5/cob_page.asp?category_id=72044&initial_file=cob_page-ltech_agenda.asp#ETA3`, January 31, 2012. LegalTech Presentation.