Automatic and Semi-Automatic Document Selection for Technology-Assisted Review

Maura R. Grossman
University of Waterloo

Gordon V. Cormack
University of Waterloo

Adam Roegiest
Kira Systems Inc.

ABSTRACT

In the TREC Total Recall Track (2015-2016), participating teams could employ either fully automatic or human-assisted ("semi-automatic") methods to select documents for relevance assessment by a simulated human reviewer. According to the TREC 2016 evaluation, the fully automatic baseline method achieved a recall-precision breakeven ("R-precision") score of 0.71, while the two semi-automatic efforts achieved scores of 0.67 and 0.51. In this work, we investigate the extent to which the observed effectiveness of the different methods may be confounded by chance, by inconsistent adherence to the Track guidelines, by selection bias in the evaluation method, or by discordant relevance assessments. We find no evidence that any of these factors could yield relative effectiveness scores inconsistent with the official TREC 2016 ranking.

ACM Reference format:
Maura R. Grossman, Gordon V. Cormack, and Adam Roegiest. 2017. Automatic and Semi-Automatic Document Selection for Technology-Assisted Review. In *Proceedings of SIGIR '17, Shinjuku, Tokyo, Japan, August 07-11, 2017,* 4 pages.
http://dx.doi.org/10.1145/3077136.3080675

1 INTRODUCTION

The purpose of the TREC Total Recall Track [1, 5] was to evaluate, through controlled simulation, technology-assisted review ("TAR") methods to achieve very high recall with a human-in-the-loop. Towards this end, the Track provided a web-based server that simulated a human-in-the-loop, by providing (pre-coded) relevance assessments, on a document-by-document basis, in response to requests from participating teams during their completion of the task. The objective of each team was to request assessments for as many relevant documents as possible, while requesting assessments for as few non-relevant documents as possible.

Participants' methods were evaluated using the traditional set-based measures of recall and precision, as well as gain curves and a novel family of rank-based measures denoted by the Track coordinators as "recall@aR+b." For brevity, we limit our consideration to the special case of "recall@R," which is equivalent to the recall-precision breakeven point, or R-precision. The recall-precision breakeven point is the proportion of documents for which relevant assessments are returned among the first R requests, where R is the number of relevant documents in the collection.

At the outset, participants retrieved the document collection from the assessment server, as well as a short topic description. Each run was declared by the participant to be either "Automatic" or "Manual." Automatic runs were permitted no manual intervention whatsoever; the only information available to an Automatic run was the collection, the topic statement, and the results of any assessments requested from the server. Manual runs were permitted unrestricted manual intervention, including, but not limited to, independent research, search of the collection, and human review of documents in the collection. As stated in the Track guidelines, "[i]f documents are manually reviewed, the same documents must also be submitted to the assessment server, at the time they are reviewed."[1] One of the two Manual runs submitted to the TREC 2016 Total Recall Track conformed to this requirement; the other did not, in effect availing itself of pre-training not available to the other runs.

The pre-coded assessments that were used to simulate human feedback, as well as to evaluate the participating runs, were derived using a process similar to a Manual run, but with real human assessors-in-the-loop. Interactive search and judging [6], as well as two machine-learning methods, were used to identify potentially relevant documents, which were labeled as "relevant" or "non-relevant," by a team of six assessors supervised by the National Institute of Standards and Technology ("NIST"). In total, the assessors reviewed 61,985 documents for relevance to 34 different topics, labeling 36,021 as "relevant" and 25,964 as "non-relevant." For the purposes of feedback and evaluation, the 36,021 documents were deemed "relevant"; all others, whether reviewed or not, were deemed "non-relevant."

To provide an evaluation standard independent of the primary assessments, a non-uniform statistical sample [2] of 50 documents was drawn from the entire collection for each of the 34 topics; each sample was reviewed by three separate assessors from the same NIST team. The Track coordinators reported separate gain curves using each of these three alternate assessments, as well the majority vote among the three assessments, as a gold standard [1].

Figures 1 and 2, reproduce the gain curves from the TREC 2016 proceedings [1] for the three runs of interest, evaluated using the primary assessments and the majority vote of the three alternate assessments, respectively. The runs of interest for this study are BMI-Desc (the Automatic baseline that achieved recall@R of 0.71), eDiscoveryTeam (the Manual

[1] http://cormack.uwaterloo.ca/total-recall/2016/guidelines.html.

SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan

ACM ISBN 978-1-4503-5022-8/17/08.
http://dx.doi.org/10.1145/3077136.3080675

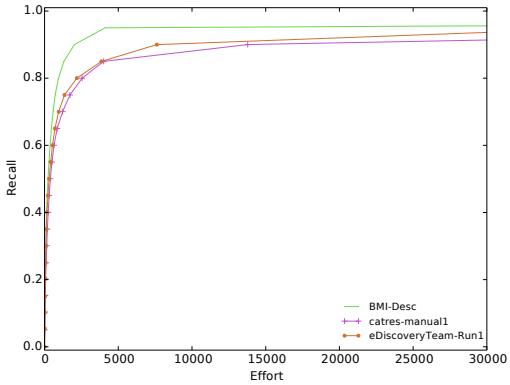

Figure 1: Gain Curves Showing Recall (Averaged Over 34 Topics) as a Function of the Number of Documents Submitted, for the Athome4 (Jeb Bush) Test Collection.

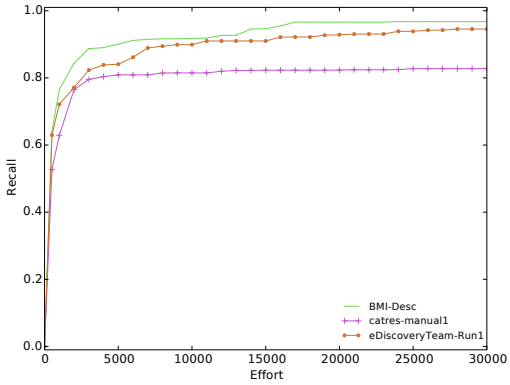

Figure 2: Gain Curves Showing Recall According to the Majority Vote of the Three Secondary Assessors (Averaged Over 34 Topics) as a Function of the Number of Documents Submitted, for the Athome4 (Jeb Bush) Test Collection.

run that achieved recall@R of 0.67), and catres (the Manual run that achieved recall@R of 0.51). A paired t-test indicates that the 95% confidence interval of the difference between the BMI-Desc score and the e-DiscoveryTeam score is between -0.012 and +0.095; in other words, the difference is neither large nor significant. In contrast, the differences between these two scores and the catres score are both significant ($p \approx 0.0001$).

Taken at face value, these results might suggest that there is little to choose between the the Automatic baseline method and one Manual method, and that both are superior to the second Manual method. Such a conclusion, we argue, would be premature without first considering the confounding factors that are investigated in this work. One such factor has already been noted – the eDiscoveryTeam run was primed with the result of assessing more than one hundred documents from

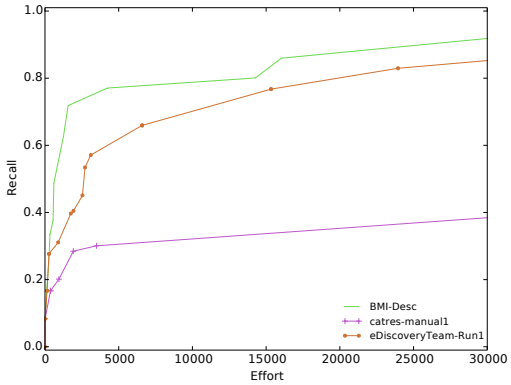

Figure 3: Gain Curves Showing Recall of Documents That Were Unjudged in the Primary Assessment, According to the Majority Vote of the Three Secondary Assessors (Averaged Over 34 Topics) as a Function of the Number of Documents Submitted, for the Athome4 (Jeb Bush) Test Collection.

the test collection, on average, per topic . For many topics, their "recall@R" results were derived from assessments of many more than R documents [3]. Another factor, suggested by the catres team [4], was that the catres run returned disproportionately fewer documents that had been reviewed by the NIST assessors, raising the question of whether or not these documents might have been coded as "relevant" had they been reviewed, thereby raising catres' score. Finally, discordant relevance assessments may have enured to the benefit of one run, at the expense of others.

2 EQUALIZING PRIOR REVIEW

We know of no way to exclude or to account for the effect of the prior review for the eDiscoveryTeam runs. However, we were able to devise a way to afford the other runs an advantage equal in magnitude to the prior review conducted by eDiscoveryTeam. Consider, for example Topic 434. eDiscoveryTeam reported having reviewed 83 documents for this topic, and requested labels for only the 38 documents they believed to be relevant, of which 33 were deemed relevant by the simulated human assessor. Coincidentally, the official value of R was also 38. Accordingly, this run scored recall@R = $\frac{33}{38} = 0.87$. According to the Track guidelines, we posit that the score should more properly be interpreted as recall@83=0.87. In contrast, if we consider the first 83 documents in the catres run, we find that recall@83=$\frac{37}{38} = 0.97$. The Automatic baseline method ("BMI-Desc") achieved an identical recall@83=0.97.

More generally, let us denote by H the number of documents reviewed prior to a Manual run. Some smaller number $h \leq H$ of these documents will be deemed "relevant" and, we assume, submitted to the assessment server. In general, it will be the case that $h \leq R$, and the next $R - h$ documents will be considered in computing recall@R, when in fact, R+H-h

Run	Recall	Std. Dev.	p (vs. next)
BMI-Desc	0.79	0.12	0.0004
eDiscoveryTeam	0.67	0.21	0.05
catres	0.61	0.22	

Table 1: Prior-Review Adjusted Recall@R+H-h, Averaged Over 34 Topics.

Run	Judged	Precision	Std. Dev.
BMI-Desc	0.88	0.80	0.17
eDiscoveryTeam	0.88	0.83	0.17
catres	0.59	0.80	0.13

Table 2: Precision of the Top R Results That Were Judged by NIST Assessors. None of the differences are significant ($p > 0.05$).

Run	Recall	Std. Dev.	p (vs. Next)
eDiscoveryTeam	0.67	0.34	0.4
BMI-Desc	0.63	0.34	0.01
catres	0.51	0.34	

Table 3: Recall@R as Evaluated by the Majority Vote of the Three Alternate Assessments, Averaged Over 34 Topics.

	Precision	Std. Dev.	p (vs. Next)
Judged	0.79	0.22	0.0001
Unjudged	0.10	0.18	

Table 4: catres' Precision Among Judged and Unjudged Documents, According to a Sample Reviewed by the Second Author.

documents were assessed. We should compare any runs that received a "head start" using the measure recall@$R + H - h$, where h is the number of relevant documents among the first H submitted. In both cases, we are replacing the threshold value of R by a somewhat greater value, affording a controlled comparison, at the expense of altering somewhat the objective function.

Table 1 shows the average recall@R+H-h over 34 topics, as well as the standard deviation, and p-values versus the next-ranked run. Recall of both BMI-Desc and catres increase substantially, from 0.51 and 0.71, to 0.61 to 0.79, respectively, when given the benefit of H documents of prior review. With this adjustment, we see that BMI-Desc achieves substantially and significantly higher recall than eDiscoveryTeam, for the same number of documents reviewed, according to the NIST primary assessments. eDiscoveryTeam also achieves significantly higher recall than catres, but the margin is substantially reduced.

3 SELECTION BIAS

In the first R documents, BMI-Desc and eDiscoveryTeam submitted a substantially higher proportion of documents that had been judged by the NIST assessors, as compared to catres. Table 2 shows the proportion of judged documents for each run, and the precision – the proportion of relevant documents – among that set. It is clear from this table that the catres run achieves similar precision on the judged documents that it returns, but returns fewer judged documents overall, and hence achieves a lower recall@R score. If a substantial number of the unjudged documents were in fact relevant, the catres result would be under-reported.

Our first avenue of investigation was to consider the secondary assessments reported by the Track coordinators. If there were a substantial number of relevant unassessed documents returned by some runs and not others, this effect should manifest itself in the results reported with respect to these alternate assessments. Figure 2 shows the gain curve achieved by the three runs, when evaluated by the majority vote of the three alternate assessors. We see very little difference compared to Figure 1, which is calculated with respect to the primary NIST assessments. It may be the case that eDiscoveryTeam's curve is closer to BMI-Desc, but the overall ordering remains intact.

To isolate the effect of unjudged documents, we obtained the raw scoring data from NIST, and calculated the recall of each run, as assessed by the majority vote of the three alternate assessors, considering only documents that were unjudged in the primary assessment. The result is shown as a gain curve in Figure 3. We had no prior hypothesis regarding what this gain curve would show. Somewhat to our surprise, we observed that the relative heights of the curves for the three runs were the same – there is no indication that the catres system was better at finding relevant, unjudged, documents than the others.

We also computed recall@R based on the majority-vote alternate assessments, the results of which are shown in Table 3. The results are generally consistent with those for the primary assessments, with larger variances, as expected. In this evaluation, eDiscoveryTeam achieves a higher score, but, as for the primary assessments, the difference is not significant. catres achieves lower recall, by a significant margin.

While the sampling and alternative assessments were independent of the primary method, the small sample size could miss small but important populations of unjudged relevant documents. For example, for Topic 434, an additional 38 relevant documents, over and above the 38 that were found by NIST, could escape sampling if they were dissimilar to the judged-relevant documents. We would expect, however, over 34 topics, that if this were a systematic issue, at least some such relevant documents would have come to light.

To avoid reliance on small samples, we sought further, direct confirmation or refutation of the hypothesis that the catres run included a substantial number of relevant unjudged documents. Towards this end, we reviewed a stratified random sample of ten documents for each topic: five judged

Method	Recall	Std. Dev.	p (vs. eDT)
BMI-Desc	0.74	0.17	0.8
BMI-eDT	0.79	0.14	0.08
eDiscoveryTeam	0.73	0.22	-
catres	0.62	0.22	0.002

Table 5: Recall@R+H-h, Averaged Over 34 Topics, Evaluated According to the "Corrected" Gold Standard. BMI-Desc is the official TREC baseline run, trained on feedback from the NIST gold standard; BMI-eDT is the same method, trained on feedback from the "corrected" gold standard.

documents that were among the first R documents returned by catres, and five unjudged documents that were among the first R documents returned by catres. The review was conducted blind by the second author, who was familiar with the subject matter.[2] The results, shown in Table 4 are consistent with the results in Table 2 with regard to precision among judged documents (0.80 vs. 0.79), and show eight times lower precision (0.10) among unjudged documents, leading us to conclude that unjudged relevant documents were not a major factor in the TREC 2016 Total Recall evaluation results.

4 ASSESSOR DISCORD

Since the earliest days of IR evaluation, disputes over relevance have raised concerns [7]. The eDiscoveryTeam report [3] suggested that the primary NIST assessments were flawed, and that, instead, their results should be evaluated according to their own "corrected" gold standard.

Confusion matrices reported by eDiscoveryTeam suggest that the magnitude of discord between its assessments and the primary NIST assessments is well within the bounds of what would be expected for independent assessments [8]. Considering our running example of Topic 434, eDiscoveryTeam reports that, according to their "corrected" gold standard, the NIST assessments contain five false positives and five false negatives. In other words, the overlap (*i.e.*, Jaccard coefficient) between the two gold standards is $\frac{33}{43} = 0.77$ – much higher than the overlap values reported in the literature for informed expert assessors, which have not proven to affect the reliability of ad hoc system evaluation [8].

eDiscoveryTeam provided us with a copy of its "corrected" gold standard.[3] The average per-topic overlap between the eDiscoveryTeam and primary NIST assessments was calculated to be 0.75. Table 5 shows recall@R+H-r for the three methods, evaluated by the eDiscoveryTeam's "corrected" gold standard. Even when BMI-Desc is trained using the primary NIST assessments, it achieves recall comparable to the eDiscoveryTeam run. When trained and evaluated using eDiscoveryTeam's "corrected" assessments, BMI-Desc achieves exactly the same recall – 0.79 – as when trained and evaluated using the primary NIST assessments. When trained on the primary NIST assessments and evaluated with respect to the eDiscoveryTeam assessments, BMI-Desc yields slightly, but not significantly, higher recall than the eDiscoveryTeam submission: 0.74 vs. 0.73. catres yields insubstantially different results, whether evaluated with respect to the eDiscoveryTeam "corrected" assessments or the primary NIST assessments: 0.62 vs. 0.61.

The impact of assessor discord on the reliability of Total Recall system evaluation has not previously been studied; the results above, and the agreement between Figures 1 and 2, suggest that, as with ad hoc retrieval, reliable evaluation of the relative effectiveness of Total Recall systems does not hinge on precise relevance assessments, either for feedback or for evaluation.

5 CONCLUSIONS

Prior to the TREC Total Recall Track, we assumed that the best Manual runs would substantially outperform the best Automatic runs, as they did in previous TREC ad hoc tasks. We were surprised that they did not. For some topics, Manual runs achieved higher recall scores with less effort (discounting prior review), but no Manual method consistently improved on the fully Automatic TREC baseline method. In this study, we sought to determine whether the apparent superiority of the Automatic baseline method at TREC 2016 was real, or attributable to chance, failure to follow the Track guidelines, selection bias, or assessor discord. We found no evidence to suggest that the Manual runs were superior to the Automatic baseline, and we found evidence to suggest that, when review effort was properly controlled, the Automatic baseline method found more relevant documents with less effort than any Manual run in the TREC 2016 Total Recall Track.

Based on all currently available evidence, the TREC Automatic baseline method remains the method to beat.

[2]The 340 documents and their corresponding assessments are available from the Authors.

[3]Whether or not eDiscoveryTeam's relevance assessments are more "correct" than the primary NIST assessments is a subjective question that is beyond the scope of this work, and has no bearing on our results.

REFERENCES

[1] M. R. Grossman, G. V. Cormack, and A. Roegiest. TREC 2016 Total Recall Track Overview. In *TREC 2016*.
[2] D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.
[3] R. C. Losey, J. Sullivan, T. Reichenberger, L. Kuehn, and J. Grant. e-Discovery Team at TREC 2016 Total Recall Track. In *TREC 2016*.
[4] J. Pickens, T. Gricks, B. Hardi, M. Noel, and J. Tredennick. An exploration of Total Recall with multiple manual seedings. In *TREC 2016*.
[5] A. Roegiest, G. V. Cormack, M. R. Grossman, and C. L. A. Clarke. TREC 2015 Total Recall Track Overview. In *TREC 2015*.
[6] M. Sanderson and H. Joho. Forming test collections with no system pooling. In *SIGIR 2004*.
[7] T. Saracevic. Why is relevance still the basic notion in information science? In *ISI 2015*.
[8] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Informantion Processing & Management*, 36(5), 2000.