Technology-Assisted Review in Empirical Medicine: Waterloo Participation in CLEF eHealth 2017

Gordon V. Cormack and Maura R. Grossman

Cheriton School of Computer Science
University of Waterloo
Waterloo ON N2L 3G1, Canada

Abstract. Screening articles for studies to include in systematic reviews is an application of technology-assisted review ("TAR"). In this work, we applied the Baseline Model Implementation ("BMI") from the TREC Total Recall Track (2015-2016) to the CLEF eHealth 2017 task of screening MEDLINE abstracts to identify articles reporting studies to be considered for inclusion. According to rank-based evaluation measures, this approach identified every article describing a study that should have been included in each of 30 systematic reviews, by examining 461 abstracts, on average, per review—12.6% of the 3,655 abstracts that would have had to be examined, on average, if instead, a manual approach had been used. While this result indicates TAR's promise to substantially reduce the time and cost of abstract screening, this promise can be realized only if it can be known with reasonable certainty for each review how many abstracts must be examined before all, or substantially all, articles that should be included have been identified. To this end, we applied our "knee-method" stopping criterion to BMI to determine how many abstracts should be examined for each topic. According to threshold-based evaluation, the knee method identified every article that should have been included (100% recall), while examining 2,659 abstracts, on average, per topic—72.8% of the 3,655 abstracts, that would have required examination, on average, had a manual approach been used instead. While our results suggest that TAR can substantially improve the efficiency of abstract screening without compromising recall, there remains room for improvement both in ranking and stopping criterion, as well as important factors that were not addressed in the CLEF eHealth 2017 framework: the completeness of the universe of abstracts gathered using keyword search, and the accuracy of the human assessments of the collected abstracts.

1 Introduction

The University of Waterloo participated in Task 2, *Technologically Assisted Reviews in Empirical Medicine* [10], of the *CLEF 2017 eHealth Evaluation Lab* [12]. Task 2 simulates the second phase—*screening*—in a prototypical three-phase workflow to identify studies for inclusion in a systematic review:

1. *Search:* First, Boolean queries are used to identify as many articles as possible that may describe studies that should be included;
2. *Screening:* Second, titles and abstracts of the articles identified in the search phase are examined to eliminate those which could not possibly describe studies that should be included; and
3. *Selection:* Finally, articles that survived the screening phase are read in full to determine whether or not they meet the systematic review inclusion criteria.

The overall objective of our research is to improve the human efficiency, as well as the effectiveness, of workflows to identify studies for inclusion in systematic reviews. The results of our CLEF experiments support the hypothesis that continuous active learning ("CAL") can substantially improve the human efficiency of screening, without substantially compromising its effectiveness. The results also are consistent with the further hypothesis that CAL actually improves effectiveness by identifying articles missed in the search phase, or articles mistakenly eliminated during the screening phase. While this hypothesis cannot be tested immediately within the framework of Task 2, we have identified a set of articles that, were it determined that they describe one or more studies that should have been included in the review, would demonstrate CAL's superior effectiveness.

Run Name	Method	Rank/Threshold	Simple/Cost Sensitive
A-rank-cost	A	Rank	Cost Sensitive
A-rank-normal	A	Rank	Simple
A-thresh-cost	A	Threshold	Cost Sensitive
A-thresh-normal	A	Threshold	Simple
B-rank-cost	B	Rank	Cost Sensitive
B-rank-normal	B	Rank	Simple
B-thresh-cost	B	Threshold	Cost Sensitive
B-thresh-normal	B	Threshold	Simple

Table 1. Official Waterloo CLEF Task 2 Submissions.

2 Apparatus

Task 2 is essentially the Technology-Assisted Review ("TAR") task addressed by the TREC 2015 and TREC 2016 Total Recall Tracks [11, 8]. For our participation in CLEF, we reprised our Total Recall efforts using the same apparatus.

At TREC, the systems under test were given, at the outset, a corpus of documents and a set of topics. For each topic, a system under test repeatedly submitted documents from the corpus to a server, and in return, was given a simulated human assessment of "relevant" or "not relevant" for each document.

The objective was to identify as many relevant documents as possible, while submitting as few non-relevant documents as possible. The tension between these

Algorithm 1 The AutoTAR Continuous Active Learning ("CAL") Method, as Implemented by the TREC Baseline Model Implementation ("BMI") and deployed by Waterloo for the CLEF Technologically Assisted Review Task.

1. The initial training set consists of a synthetic document containing only the topic title, labeled as "relevant."
2. Set the initial batch size B to 1.
3. Temporarily augment the training set by adding 100 random documents from the collection, provisionally labeled as "not relevant."
4. Apply logistic regression to the training set.
5. Remove the random documents added in step 3.
6. Select the highest-scoring B documents that have not yet yet been screened.
7. Label each of the B documents as "relevant" or "not relevant" by consulting:
   (a) Previous "abstract" assessments supplied by CLEF [Method A]; or,
   (b) Previous "document" assessments, once the first "relevant" document assessment is encountered [Method B].
8. Add the labeled documents to the training set.
9. Increase B by $\lceil \frac{B}{10} \rceil$.
10. Repeat steps 3 through 10 until either:
    (a) All documents have been screened [for ranked evaluation]; or,
    (b) The "knee-method" stopping criterion is met [for threshold evaluation].

Measure	Method A	Method B
num_rels	607	607
rels_found	607	575
r	1.000	0.979
num_docs	109,560	109,560
num_shown	79,765	52,934
%shown	72.8%	48.3%
loss_r	0.0	0.008
loss_e	0.657	0.526
loss_er	0.657	0.534

Table 2. Thresholding Results for Waterloo Method A and Method B, Measured Against Full-Document Selection. All measures were calculated using the CLEF evaluation tool, except %shown=num_shown÷num_docs. %shown is num_shown expressed as a fraction of the total number of documents screened.

Measure	Method A	Method B
wss_95	0.814	0.824
wss_100	0.823	0.830
num_docs	109,560	109,560
last_rel	461	469
num_shown*	13,830	14,070
%shown*	12.6%	12.8%
NCG@10	0.699	0.727
NCG@20	0.944	0.956
NCG@30	0.985	0.987
NCG@40	0.997	0.997
NCG@50	0.997	0.998
NCG@60	0.998	1.000
NCG@70	1.000	1.000
NCG@80	1.000	1.000
NCG@90	1.000	1.000
NCG@100	1.000	1.000
norm_area	0.948	0.955
ap	0.189	0.231

Table 3. Ranking Results for Waterloo Method A and Method B over 30 Topics, Measured Against Full-Document Selection. All measures were calculated using the CLEF evaluation tool, except num_shown*= 30·last_rel, and %shown*=num_shown*÷num_docs. num_shown* indicates the number of documents shown during screening at the point when all relevant articles have been identified; %shown* expresses num_shown* as a fraction of the total number of documents screened.

two criteria was evaluated using rank-based measures (*e.g.*, recall as a function of the number of documents submitted), as well as set-based measures (*e.g.*, recall at a point when a certain number of documents, specified contemporaneously by the system, had been submitted).

Prior to TREC, we made available a Baseline Model Implementation ("BMI"),[1] to illustrate the client-server protocol, as well as to provide baseline results for comparison. BMI, which encapsulates our AutoTAR Continuous Active Learning ("CAL") method [1], yielded rank-based results that compared favorably will all systems under test. During the course of our participation in TREC, we developed and tested the "knee method" stopping procedure [3, 2, 5], with the purpose of achieving high recall with high probability.

Task 2 differed operationally from the TREC Total Recall Track in that a list of document identifiers, rather than a corpus, was supplied at the outset, and a complete set of relevance assessments, rather than an assessment server were used to simulate human assessments. Task 2 also differed substantively from the Total Recall Track in that the corpus for each topic was narrowed by a search phase specific to that topic, and therefore yielded a much smaller set that was richer in relevant documents. Task 2 differed further in that two sets of relevance assessments were available: the assessments from a previously conducted screening phase, and the assessments from a previously conducted selection phase, raising the question of which assessments (or combination of assessments) should be used to simulate relevance feedback, and which should be used to evaluate the results (*cf.* [6]).

Task 2 provides no method equivalent to TREC's "call your shot" for a system under test to specify a stopping criterion (for threshold-based evaluation), while at the same time continuing until every document in the corpus has been submitted for assessment (for rank-based evaluation).

Task 2, however, unlike TREC, afforded participants the opportunity to conduct task-specific tuning and configuration, by supplying 20 training topics (with corresponding corpora and assessments) in advance of the exercise, followed by 30 test topics, which were used for evaluation.

3 Training and Configuration

3.1 Document Corpora

The corpus for each topic consisted of abstracts from MEDLINE/Pubmed[2] identified by PMID. On March 8, 2017, we fetched the entire MEDLINE dataset consisting of 27,348,935 XML files, each containing the titles, abstracts, and metadata for an article. We used the raw XML files as documents in the corpora that were supplied at the outset to BMI.

[1] Available under GNU General Public License at http://cormack.uwaterloo.ca/trecvm.

[2] *See* https://www.nlm.nih.gov/bsd/pmresources.html.

Our original intent had been to apply BMI to the entire corpus of 27,348,935 files, thus combining the search and screening phases. When we employed this strategy in a pilot experiment on the test topics, we found that no assessments were available for many, if not most, of the highly ranked documents returned by BMI. To our eye, these documents were indistinguishable from those for which "relevant" assessments were provided. We investigated, without success, the reasons why these documents were not retrieved by the previously conducted search phase. For example, the documents in question were neither newer nor older than those for which assessments were available, and appeared to contain relevant terms from the search query. As we were unable to reproduce the results of the CLEF search phase, we chose to ignore—for the purpose of relevance feedback and evaluation—documents for which no assessments were available. Ignoring these unjudged documents, our pilot experiment yielded what appeared to be reasonable rank-based results.

Ignoring documents for feedback and evaluation yields a substantially different result from removing them from the corpus altogether. In a second pilot experiment, we constructed a separate corpus for each topic, consisting of only those documents for which relevance assessments were available. While BMI ran much faster on these reduced corpora than on the 27M dataset, results were apparently inferior. We conjecture that this inferior result can be explained by skewed term-frequency statistics in the reduced corpora.

As a compromise between the effectiveness of searching the 27M dataset and the (computational) efficiency of searching the reduced corpora, we conducted a third pilot experiment using a common corpus consisting of all documents that were assessed for any of the 20 test topics. That is, for any given topic, the corpus consisted of all the documents assessed for that topic, as well as all the documents assessed for each of the other 19 topics. Our rationale was that including documents retrieved for all topics would introduce enough diversity to unskew sufficiently the term-frequency statics. This approach appeared to achieve the efficiency of using reduced corpora and the effectiveness of using the full dataset, and was chosen for our official tests: For the official tests, the corpus consisted of all documents assessed for any of the 30 test topics (less four documents whose PMIDs were not present in our MEDLINE database); from this corpus, we submitted and solicited feedback only for documents for which assessments were available.

3.2 Relevance Feedback

We investigated three modes of relevance feedback, of which only two were selected for official testing:

1. Relevance feedback based on the screening-phase assessments (selected as Method A for official testing);
2. Relevance feedback based on the selection-phase assessments (not selected for official testing);
3. Relevance feedback based on a hybrid of screening-phase and selection-phase assessments (selected as Method B for official testing).

The first and second methods are straightforward: When BMI identifies a document for assessment, the judgment returned to BMI is that supplied by CLEF for either the screening phase (the "abstract qrels") or the selection phase ("the content qrels"). The third method operates in two phases: At the outset, the judgment returned to BMI is that of the abstract qrels. The abstract qrels continue to be used until BMI identifies one document that is relevant not only according to the abstract qrels, but also according to the content qrels. Thereafter, the judgment returned to BMI is that of the content qrels.

In our pilot experiments, we found that the first method consistently yielded superior rank-based results, whether evaluated using the abstract qrels or the content qrels. The second method yielded consistently inferior results. The third method showed similar, but slightly inferior, results, to the first method, when evaluated using the content qrels. Based on our pilot results, we selected the first and third methods, denoted as Method A and Method B, respectively, for our official experiments.

3.3 Stopping Criterion

For threshold-based evaluation, it was necessary to implement a stopping procedure to terminate screening when the best compromise between recall and effort had been achieved, for some definition of "best." In our opinion, technology-assisted review should be considered a satisfactory alternative to manual review only if it yields comparable or superior recall, with high probability. Toward this end, we deployed our knee method with default parameters ($\rho = 156 - \min(relret, 150)$, $\beta = 100$ [3]), which interprets a sharp fall-off in the slope of the gain curve (recall vs. review effort) as evidence that substantially all relevant documents have been identified.

3.4 Runs and Evaluation

The Task 2 guidelines specify a plethora of run types and evaluation measures, which may be classified on two orthogonal dimensions:

1. Rank-based vs. threshold-based (or set-based) evaluation; and
2. Simple vs. cost-sensitive scoring.

The strategies to optimize these measures are incompatible, occasioning us to submit four versions of the output from each of our two runs, for a total of eight submissions, detailed in Table 1. The only difference between the "rank" and "thresh" runs is that the latter are truncated using the knee-method stopping procedure; the only difference between the "normal" and "cost" runs is that the "interaction field" "AF" is replaced by "AFS" where the document receives a "relevant" assessment, and by "AFN" where the document receives a "non-relevant" assessment.

4 AutoTAR

In 2015, we published the details and rationale for AutoTAR [1], which remains, to this date, the most effective TAR method of which we are aware. BMI implements AutoTAR exactly as described above, except for the substitution of Sofia-ML logistic regression in place of SVM^{light} (see [4, Section 3.1]). It has no dataset- or topic-specific tuning parameters; except for modifications to incorporate the CLEF corpora and relevance assessments, and our knee-method stopping procedure, we used BMI "out of the box."

The AutoTAR/BMI algorithm, as modified for CLEF, is detailed in Algorithm 1, which is reproduced from [1] with the following changes:

- In Step 1, AutoTAR gives the option of starting with a relevant document, or with a synthetic document. Here, we used a synthetic document consisting of the title of the topic, and nothing else.
- In Step 7, we introduced two different ways to simulate user feedback, corresponding to Method A and Method B, described above in Section 3.2.
- In Step 10, we introduced the option to terminate the process when the knee-method stopping criterion was met.

Internally, BMI constructs a normalized TF-IDF ($(1 + \log tf) \cdot \log \frac{N}{df}$) word-vector representation of each document in the corpus (which, as noted in Section 3.1, consists of raw XML files), where a word is considered to be any sequence of two or more alphanumeric characters not containing a digit, that occurs at least twice in the corpus. Scoring is effected by Sofia-ML[3] with parameters "`--learner_type logreg-pegasos --loop_type roc --lambda 0.0001 --iterations 200000`." As noted above, these parameters were fixed when BMI was created in 2015.

5 Results

We present separately the results for our threshold-based and rank-based runs, reporting only simple threshold-based and simple rank-based measures for each, computed using the content qrels. At the time of writing, cost-sensitive evaluation was not available to CLEF participants.

5.1 Threshold-Based Results

Our threshold-based results are shown in Table 2. Perhaps the most important result is shown in the first three lines: Across 30 topics, Method A identified all 607 articles referencing studies that should have been included, thus achieving 100% recall. Method B, on the other hand, identified 575 of the articles, achieving 97.9% recall. Method A, however, entailed the review of 79,765 (72.8%) of the

[3] *See* https://github.com/glycerine/sofia-ml.

109,560 abstracts identified by the search phase, while method B entailed the review of only 52,934 (48.3%) of the documents.

In other words, Method A was more effective, but Method B was more efficient. According to the combined loss measure which considers both factors, Method B was superior.

5.2 Rank-Based Results

Our rank-based results are shown in Table 3. Work saved over sampling ("WSS")—a measure commonly reported for systematic review—reflects how many fewer documents would have been needed to have been reviewed to achieve a particular level of recall, *if it were somehow known exactly when that level had been achieved*. Thus, WSS, along with all other rank-based measures, is a measure of what might have been, rather than achieved effectiveness. According to WSS, Method A is marginally inferior to Method B at 95% recall (0.815 vs. 0.824), and at 100% recall (0.823 vs. 0.830).

Conversely, Method A is marginally superior to Method B in terms of the number of documents that had to be examined per topic before 100% recall was achieved (461 vs. 469, representing 12.6% and 12.8%, respectively, of the average number of documents per topic). In other words, Method A could have achieved 100% recall with roughly on-sixth the review effort, had a stopping procedure been able to determine when 100% recall had occurred. Similarly, Method B could have achieved 100% recall with roughly four times less effort that it actually required to achieve 97.9% recall, had a stopping procedure been available.

The Normalized Cumulative Gain ("NCG") results—which report the recall achieved when a specified fraction (between 10% and 100%) of the documents have been reviewed—tell much the same story: Very high recall could have been achieved at a fraction of the review effort, had it been know when high recall had been achieved.

In our opinion, cumulative measures like norm_area and average precision yield very little insight into the actual or hypothetical effectiveness of technology-assisted review for screening purposes.

6 Discussion

We believe that both sets of the CLEF assessments are incomplete with respect to the overall objective of identifying *all* studies that should be included in the review: The screening assessments are available only for documents retrieved by the search phase; the selection assessments are available only for documents retrieved by the search phase, and judged relevant during the screening phase. Therefore, from the assessments, it is impossible to determine whether an article not retrieved by the search phase, or an article eliminated during the screening phase, describes a study that should have been included in the review. The Task 2 architecture tacitly assumes that no such articles exist; in other words, that

the search and screening phases used to generate the relevance assessments were infallible, and each attained 100% recall.

Such an assumption is unrealistic, and limits the recall of any simulated TAR method to that of the manual review to which it is compared [6]. As noted in the Cochrane Handbook [9] with regard to the search phase: "[T]here comes a point where the rewards of further searching may not be worth the effort required to identify the additional references." And with regard to the screening phase: "Using at least two authors may reduce the possibility that relevant reports will be discarded (Edwards 2002 [7])."

Our hypothesis that our TAR runs found relevant articles that were missed by the search phase, or incorrectly discarded in the screening phase, is based on results from other domains [6], where TAR acting as a "second assessor" was able to identify potentially relevant documents that had been judged "non-relevant" by a human assessor. When we applied Method A to the 30 topics, it identified 9,250 potentially relevant articles for which the abstract qrel was "not relevant." Acquiring a second opinion on each of these documents would increase the cost of the TAR review by approximately 12%, and would, we believe, yield a substantial number of relevant documents, over and above the 670 identified in the abstract qrels.

References

1. G. V. Cormack and M. R. Grossman. Autonomy and reliability of continuous active learning for technology-assisted review. *arXiv preprint arXiv:1504.06868*, 2015.
2. G. V. Cormack and M. R. Grossman. Waterloo (Cormack) participation in the TREC 2015 Total Recall Track. In *Proceedings of The Twenty-Fourth Text REtrieval Conference, TREC 2015, Gaithersburg, Maryland, USA, November 17-20, 2015*, 2015.
3. G. V. Cormack and M. R. Grossman. Engineering quality and reliability in technology-assisted review. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, pages 75–84, 2016.
4. G. V. Cormack and M. R. Grossman. Scalability of continuous active learning for reliable high-recall text classification. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, pages 1039–1048, 2016.
5. G. V. Cormack and M. R. Grossman. "When to stop" Waterloo (Cormack) participation in the TREC 2016 Total Recall Track. In *Proceedings of The Twenty-Fifth Text REtrieval Conference, TREC 2016, Gaithersburg, Maryland, USA, November 15-18, 2016*, 2016.
6. G. V. Cormack and M. R. Grossman. Navigating imprecision in relevance assessments on the road to total recall: Roger and me. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2017, Tokyo, Japan, August 7-11, 2017*, 2017.
7. P. Edwards, M. Clarke, C. DiGuiseppi, S. Pratap, I. Roberts, and R. Wentz. Identification of randomized controlled trials in systematic reviews: accuracy and reliability of screening records. *Statistics in Medicine*, 21(11):1635–1640, 2002.

8. M. R. Grossman, G. V. Cormack, and A. Roegiest. TREC 2016 Total Recall Track overview. In *Proceedings of The Twenty-Fifth Text REtrieval Conference, TREC 2016, Gaithersburg, Maryland, USA, November 15-18, 2016*, 2016.
9. J. P. Higgins and S. Green. *Cochrane handbook for systematic reviews of interventions*, volume 4. John Wiley & Sons, 2011.
10. E. Kanoulas, D. Li, L. Azzopardi, and R. Spijker. Overview of the CLEF technologically assisted reviews in empirical medicine. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation forum, Dublin, Ireland, September 11-14, 2017.*, CEUR Workshop Proceedings. CEUR-WS.org, 2017.
11. A. Roegiest, G. V. Cormack, M. R. Grossman, and C. L. A. Clarke. TREC 2015 total recall track overview. In *Proceedings of The Twenty-Fifth Text REtrieval Conference, TREC 2015, Gaithersburg, Maryland, USA, November 17-20, 2015*, 2015.
12. H. Suominen, L. Kelly, L. Goeuriot, E. Kanoulas, A. Névéol, G. Zuccon, and J. R. M. Palotti. Overview of the CLEF ehealth evaluation lab 2017. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11-14, 2017, Proceedings*, Lecture Notes in Computer Science. Springer, 2017.