Technology-Assisted Review in Empirical Medicine: Waterloo Participation in CLEF eHealth 2018

Gordon V. Cormack and Maura R. Grossman

University of Waterloo

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 2018 task of screening MEDLINE abstracts to identify articles reporting studies to be considered for inclusion. We employed exactly the same approach for Sub-Task 1 and Sub-Task 2, which was in turn exactly the same approach employed for the CLEF 2017 eHealth Lab. The only difference was that for Sub-Task 1, the entire Pubmed/MedLine database was searched; whereas for Sub-Task 2, the only records searched were those identified by CLEF using Boolean queries.

1 Introduction

The University of Waterloo participated in Task 2, *Technologically Assisted Reviews in Empirical Medicine* [11], of the *CLEF 2018 eHealth Evaluation Lab* [13]. Task 2 was divided into Sub-Task 1 and Sub-Task 2, which simulate, respectively, the first and second phases, and only the second phase, 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. Our CLEF experiments investigate the following hypotheses:

1. Can continuous active learning ("CAL") can substantially improve the human efficiency of screening, without substantially compromising its effectiveness.
2. Does CAL obviate the use of keywords to select a universe of documents for screening?

2 Apparatus

Task 2 is essentially the Technology-Assisted Review ("TAR") task addressed by the TREC 2015 and TREC 2016 Total Recall Tracks [12, 9]. For our participation in CLEF 2018, we reprised our Total Recall efforts, and also our efforts from CLEF 2017 [7] 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 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.

Sub-Task 2, which was the only task for CLEF 2017, 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. Sub-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. Sub-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]).

Sub-Task 1, new for CLEF 2018, resembles the TREC Total Recall Track, in that no topic-specific culling of the document set is done; each search applies to the entire 30M-document Pubmed/MEDLINE collection.

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

3 Training and Configuration

3.1 Document Corpora

The corpus for each topic consisted of abstracts from MEDLINE/Pubmed[2] identified by PMID. On April 1, 2018, we fetched the entire MEDLINE dataset consisting of 28,256,688 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.

For Sub-Task 1, we applied BMI to the entire corpus of 28,256,688 files, thus combining the search and screening phases. 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. Nonetheless, for Sub-Task 1, we treated any document for which no qrel was available to be "not relevant" for the purpose of feedback.

In a separate manual run, the authors used their own judgement to assess the relevance of abstracts returned by BMI, in order to provide relevance feedback.

For Sub-Task 2, we used a common corpus consisting of all documents that were assessed for any of the 30 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 29 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; any unassessed document was considered "not relevant."

3.2 Relevance Feedback

For Sub-Task 2, we used two modes of relevance feedback:

1. Relevance feedback based on the screening-phase assessments (Method UWA.Task2 for official testing);
2. Relevance feedback based on a hybrid of screening-phase and selection-phase assessments (Method UWB.Task2 for official testing).

The first method is 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"). The second method operates in two

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

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.

For Sub-Task 3, we used three modes of relevance feedback:

1. Relevance feedback based on the screening-phase assessments (Method UWA .Task1 for official testing);
2. Manual feedback based on the authors' relevance assessments (Method UWG.Task1 for official testing);
3. Manual feedback based on the authors' relevance positive assessments, followed by relevance feedback based on the screening-phase assessments (Method UWX.Task1 for official testing).

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.

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.

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 assess-
       ment 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].
```

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 and Discussion

Table 1 shows the average number of documents reviewed and recall achieved when the stopping criterion is met. We note that the CAL approach achieves similar high recall for each subtask, at the expense of more assessment effort. This additional assessment effort must be balanced against the effort to construct a precise yet inclusive Boolean query, as well as the risk of missing relevant documents that are not matched by the query.

There is some question as to how realistic the simulated assessments were for Subtask 1, as unassessed documents were deemed to be not relevant. Had all documents actually been assessed, the simulated effort may have been lower.

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

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

Table 1. Average Recall vs. Number of Documents Reviewed at Threshold

Subtask	Run	#Docs	Recall
1	UWA	3559	0.951
1	UWG	3612	0.962
1	UWX	3613	0.951
2	UWA	2926	0.990
2	UWB	1764	0.927

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 CLEF 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 [10] 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 [8])."

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. G. V. Cormack and M. R. Grossman. Technology-assisted review in empirical medicine: Waterloo participation in clef ehealth 2017. *Working Notes of CLEF*, pages 11–14, 2017.
8. 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.
9. 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.
10. J. P. Higgins and S. Green. *Cochrane handbook for systematic reviews of interventions*, volume 4. John Wiley & Sons, 2011.
11. E. Kanoulas, R. Spijker, D. Li, and L. Azzopardi. CLEF technologically assisted reviews in empirical medicine overview. In *CLEF 2018 Evaluation Labs and Workshop: Online Working Notes*, CEUR Workshop Proceedings. CEUR-WS.org, 2018.
12. 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.
13. H. Suominen, L. Kelly, L. Goeuriot, E. Kanoulas, L. Azzopardi, D. Li, A. Névéol, L. Ramadier, A. Robert, J. R. M. Palotti, and G. Zuccon. Overview of the CLEF ehealth evaluation lab 2018. In *CLEF 2018 - 8th Conference and Labs of the Evaluation Forum*, Lecture Notes in Computer Science. Springer, 2018.