MRG_UWaterloo Participation in the TREC 2020 Precision Medicine Track

Maura R. Grossman and Gordon V. Cormack
University of Waterloo

Ba' Pham
University of Toronto

Overview

The MRG_UWaterloo group from the University of Waterloo, in collaboration with Ba' Pham of the University of Toronto Health Economics and Assessment Collaborative, participated for the first time in the TREC 2020 Precision Medicine Track.

Our baseline run (**uwbm25**) used the BM25 relevance ranking method, as implemented by the Wumpus Search Engine,[1], with default parameters. The remaining runs examined the impact of various forms of pseudo-relevance feedback and manual relevance feedback, using relevance assessments by one author (Pham) who has extensive experience in conducting rapid systematic reviews, but not with precision medicine. To our surprise, we found that none of the feedback methods differed substantially in effectiveness from our baseline run. Based on preliminary NIST feedback, the effectiveness of all runs was near the median of all TREC submissions.

Our pseudo-relevance feedback run (**uwpr**) selected the 20 highest-ranked documents from uwbm25 as positive training examples, and 100 randomly selected documents from the corpus as negative training examples. These training examples were converted to tf-idf feature vectors and used as input to Sofia-ML[2] to compute a log-likelihood score for each document. Documents were sorted by score, and the top-scoring 1,000 documents were submitted as run uwpr.

Our positive-only manual feedback run (**uwr**) started with the same documents as uwpr, but the 20 highest-ranked documents from uwbm25 were assessed for relevance by our expert (Pham), and only those assessed to be relevant were included as positive training examples. Those documents assessed to be non-relevant were excluded from the training set.

Our postive-and-negative manual feedback run (**uwrn**) used the documents and relevance assessments from uwpr, treating documents assessed to be relevant as positive training examples, and documents assessed to be non-relevant, as well as the 100 random documents, as negative training examples.

Our Continuous Active Learning ("CAL") approach [1] (**uwman**) started with the result of uwrn, repeatedly selecting the top-20 scoring documents for assessment, adding them to the training set, creating a new model, and re-scoring the documents. Our expert assessor also consulted external resources such as Web of Science to find other relevant documents that should be included. The final ranking consisted of those documents assessed relevant during assessment, ranked by score, followed by all other documents also ranked by score. Documents assessed non-relevant were ranked no differently from documents that were never assessed.

Results

Discussion

It is apparent that our expert's assessments differed substantially from the official TREC assessments, and that these differences thwarted our attempts to leverage them for relevance feedback. It may be that our expert's assessments were too conservative. It has been observed elsewhere that more liberal assessments, even by less skilled reviewers, may yield better results for relevance feedback [2].

[1] http://stefan.buettcher.org/cs/wumpus/index.html.
[2] https://code.google.com/archive/p/sofia-ml/ with parameters --learner_type logreg-pegasos --loop_type roc --lambda 0.0001.

	infNDCG	P_10	R_prec
uwbm25	0.4539	0.5097	0.3449
uwpr	0.4269	0.4355	0.2987
uwr	0.4371	0.4677	0.3331
uwrn	0.3926	0.4516	0.2885
uwman	0.4510	0.5194	0.3497

Tab. 1: Mean results over all topics, for each MRG_UWaterloo run.

The precision medicine topics may present retrieval difficulties similar to the "intersection topics" that were problemetic in the TREC 2002 Filtering Track [3]. When relevance is defined to be the conjunction of several criteria, documents meeting some, but not all, of the criteria are considered to be non-relevant. Our working hypothesis is that when the learning algorithm is trained on a number of documents meeting some, but not all of the criteria, it may infer that each is an indicator of non-relevance, while failing to infer that, in combination, they indicate relevance.

References

[1] Gordon V Cormack and Maura R Grossman. Autonomy and reliability of continuous active learning for technology-assisted review. *arXiv:1504.06868*, 2015.

[2] Adam Roegiest and Gordon V. Cormack. Impact of review-set selection on human assessment for text classification. In *SIGIR 2016*.

[3] Ian Soboroff and Stephen Robertson. Building a filtering test collection for TREC 2002. In *SIGIR 2003*.