RealSakaiLab at the TREC 2020 Health Misinformation Track

Sijie TAO† and Tetsuya SAKAI†

† Department of Computer Science and Engineering, Waseda University
3–4–1 Okubo, Shinjuku-ku, Tokyo 169–8555, Japan
E-mail: †tsjmailbox@ruri.waseda.jp, †tetsuyasakai@acm.org

Abstract In this paper, we describe our experiments conducted for the AdHoc Retrieval task of the TREC 2020 Health Misinformation Track. This task offers a challenges to participants to design a ranking model that promotes retrieval of both credible and correct health information. To address both relevance and credibility, we combined several techniques to re-rank a BM25 baseline ranking. The results from a language identification model, a news category classifier and a majority score calculation were used to modify the BM25 scores of the baseline ranking.

1 Introduction

Incorrect information in a search engine results page (SERP) can be detrimental. Previous research has shown that a bias towards misinformation in SERP can increase the possibility of the users making incorrect choices in a decision-making task [2].

The AdHoc Retrieval task of the TREC 2020 Health Misinformation Track aims to provide a venue for participants to design ranking models to retrieve information that is both credible and correct. Participants are required to develop approaches to rank not only relevant but also credible and correct documents over those with incorrect information. This year, the track focuses on the pandemic, and hence, the topics are all related to COVID-19.

We submitted four runs to the track. One is a BM25 ranking as the baseline, and the other three runs are re-ranked based on the BM25 ranking scores. Several techniques are combined in the re-ranking runs. First, as non-English documents are considered irrelevant in this track, we utilized a language identification model to filter out documents that are not in English. Further, to boost the rankings of the documents that may contain health information, we trained a news category classifier to detect documents that have relevant contents. Next, the credibility of a document is modeled by calculating a majority score, which is based on its similarities with other documents. The results from the language identification model, the news category classifier, and the majority score calculation are used to modify the BM25 scores of the baseline ranking.

2 Submitted Runs

In this section, details of the submitted runs are discussed.

2.1 Language Identification

Although the corpus used in this task contains documents in different languages, non-English documents are considered irrelevant in this track. Therefore, it is necessary to detect the language in which a document is prepared and filter the non-English documents from the ranking. A pre-trained language identification model using fastText [3] [4][1] is used to identify the languages in which a document is prepared.

2.2 News Category Classification

To rank more documents that contain health information, we trained a classifier to detect the categories of the documents' content. As all documents are news articles, a news category classifier was trained on an external dataset named News Category Dataset[2]. This dataset contains 200,000 news headlines and each headline is assigned to a corresponding category. There are 41 categories, and we chose four categories, politics, science & tech, wellness, and world news as "relevant categories", that could contain health-related information about COVID-19. If a document is classified under one of the relevant categories, then the result from the classifier will be used to boost the ranking of the document.

2.3 Majority Score

As the task requires participants to retrieve credible information, we address the credibility of a document by calculating a "Majority Score". This idea is based on a hypothesis that "the more similar a document is to others, the more likely the document is credible". For each topic, the baseline BM25 ranking returns 1,000 documents. The Majority Score of a retrieved document is calculated as the average of the similarity values with the other 999 documents. For calculating document similarity, the documents are first represented

[1]: https://dl.fbaipublicfiles.com/fasttext/supervised-models/lid.176.bin

[2]: https://www.kaggle.com/rmisra/news-category-dataset

Runs	cam_map_three	nDCG	Compatibility (harmful only)	Compatibility (helpful only)
RSL_BM25	0.139	0.318	0.048	0.217
RSL_BM25LC	0.139	0.331	0.067	0.239
RSL_BM25LM	0.138	0.319	0.045	0.257
RSL_BM25LMC	0.148	0.338	0.070	0.268
Median	0.139	0.331	0.075	0.334

Table 1 Evaluation results

as tf-idf vectors, and the similarity value of two documents is defined as the cosine similarity of the document representations.

2.4 Runs

We submitted four runs to the track. The first, RSL_BM25 is a baseline BM25 ranking, automatically generated by Anserini [6][3] with default parameter settings.

The second run is named RSL_BM25LC, where L stands for Language Identification and C stands for Category Classification. The BM25 scores are set to 0 if the detected language is not English. For English documents, the BM25 scores are modified as the following equation if they are classified under a relevant category:

$$RSL\_BM25LC(d) = BM25(d) * (1 + P(d)) \quad (1)$$

where $P(d)$ is the output of the news category classifier for a given document d, which is the probability of the document being classified under a relevant category (See Section 2.2).

The third run RSL_BM25LM (Language Identification and Majority Score) re-ranks documents by boosting the BM25 scores using the majority scores:

$$RSL\_BM25LM(d) = BM25(d)*(1+MajorityScore(d)) (2)$$

The last run RSL_BM25LMC is a run combining RSL_BM25LC and RSL_BM25LM, where the scores of the documents are calculated as the sum of the scores from the two runs.

3 Result and Discussion

The results were assessed to judge whether the retrieved documents are useful, credible, and correct. Table 1 shows the evaluation results of our runs and the medians of the scores from all the participant teams. The evaluation measure cam_map_three is an extended version of the convex aggregating measure (CAM) [5]. The organizers designed this measure based on the original one considering all the three aspects (usefulness, credibility, and correctness). Other measures such as nDCG and compatibility [1] were also used to evaluate the runs.

From the evaluation results, it is difficult to say that our method can retrieve more credible and correct documents. Although the combined run RSL_BM25LMC slightly outperforms the other runs, the run RSL_BM25LM does not produce better results than the BM25 baseline run or the RSL_BM25LC in terms of cam_map_three and nDCG. Therefore, from the evaluation results, it is not clear whether reranking using majority score can promote retrieval of more credible and correct documents. Re-ranking using majority score does not work as expected because using tf-idf may not be effective in capturing document similarity.

4 Conclusions

In this paper, we described our participation in the TREC 2020 Health Misinformation Track. We used several techniques to build our runs: a language identification model, a news category classifier, and a majority score calculation to modify the BM25 baseline ranking. The evaluation results show that our method cannot significantly impact the retrieval of correct and credible documents. In future work, we hope to find alternative ways for better identification of correct and credible documents to help users avoid making erroneous decisions.

References

[1] Clarke, C., Vtyurina, A. and Smucker, M.: Assessing top-k preferences, *ArXiv*, Vol. abs/2007.11682 (2020).

[2] Ghenai, A., Smucker, M. and Clarke, C.: A Think-Aloud Study to Understand Factors Affecting Online Health Search, *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval* (2020).

[3] Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H. and Mikolov, T.: FastText.zip: Compressing text classification models, *arXiv preprint arXiv:1612.03651* (2016).

[4] Joulin, A., Grave, E., Bojanowski, P. and Mikolov, T.: Bag of Tricks for Efficient Text Classification, *arXiv preprint arXiv:1607.01759* (2016).

[5] Lioma, C., Simonsen, J. and Larsen, B.: Evaluation Measures for Relevance and Credibility in Ranked Lists, *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval* (2017).

[6] Yang, P., Fang, H. and Lin, J.: Anserini: Reproducible Ranking Baselines Using Lucene, *ACM J. Data Inf. Qual.*, Vol. 10, pp. 16:1–16:20 (2018).

[3] https://github.com/castorini/anserini