Retrieving and Ranking Studies for Systematic Reviews: University of Sheffield's Approach to CLEF eHealth 2018 Task 2

Working Notes for CLEF 2018

Amal Alharbi, William Briggs and Mark Stevenson

Department of Computer Science, University of Sheffield, UK
{ahalharbi1,wbriggs2,mark.stevenson}@sheffield.ac.uk

Abstract This paper describes the University of Sheffield's approach to CLEF 2018 eHealth Task 2: Technologically Assisted Reviews in Empirical Medicine. This task focuses on identifying relevant studies for systematic reviews. The University of Sheffield participated in both subtasks. Our approach to subtask 1 was to extract keywords from search protocols and form them into queries designed to retrieve relevant documents. Our approach to subtask 2 was to enrich queries with terms designed to identify diagnostic test accuracy studies and also by making use of relevance feedback. A total of six official runs were submitted.

1 Introduction

Systematic reviews aim to identify and summarise all available evidence to answer a specific question such as 'for deep vein thrombosis is D-dimer testing or ultrasound more accurate for diagnosis?'[1]. The process of conducting a systematic review is time-consuming and a single review may require up to 12 months of expert effort [2,3]. A significant amount of this effort is spent on manually screening studies to identify those which should be included in the review. This effort can be significantly reduced by applying text mining techniques to identify relevant studies (semi-)automatically [4,5,6,7].

There are three main stages in the process of identifying relevant studies for a systematic review [8]:

- **Boolean Search:** A Boolean query is created and applied to a medical database, such as MEDLINE, to retrieve a set of candidate citations.
- **Title and Abstract Screening:** The title and abstract of all candidate citations return from the Boolean query are screened to decide which ones should be considered for inclusion in the review.
- **Content Screening:** The full text of the remaining citations are examined to determine the final set that will be included in the review.

CLEF eHealth 2018 [9] Task 2 [10] consisted of two subtasks: Subtask 1 related to the first stage ('*Boolean Search*') while Subtask 2 focussed on the second stage ('*Title and Abstract Screening*'). Both subtasks focus on Diagnostic Test Accuracy (DTA) reviews.

This paper is structured as follows: Sections 2 and 3 describe the approach to and results obtained for subtasks 1 and 2 (respectively). Conclusions are presented in Section 4.

2 Subtask 1: No Boolean Search

Before constructing a Boolean Query, reviewers design and write a search protocol that defines in detail what constitutes a relevant study for their review.

The goal of subtask 1 is to create a search strategy based on the protocol without developing a Boolean query. Participants were expected to interpret the protocol and use information from it to identify relevant studies directly from PubMed. The task is a complex problem that can be viewed as involving both Information Extraction and search.

2.1 Datasets

Participants in Subtask 1 were provided with 40 example reviews for training and an additional 30 for testing. The data provided included full protocols as well as protocol summaries. The summaries were variable in length and typically contain three main headings: topic, title and objective. See Figure 1 for an example protocol summary.

> Topic: CD008122
>
> Title: Rapid diagnostic tests for diagnosing uncomplicated P. falciparum malaria in endemic countries
>
> Objective: To assess the diagnostic accuracy of RDTs for detecting clinical P. falciparum malaria (symptoms suggestive of malaria plus P. falciparum parasitaemia detectable by microscopy) in persons living in malaria endemic areas who present to ambulatory healthcare facilities with symptoms of malaria, and to identify which types and brands of commercial test best detect clinical P. falciparum malaria.

Figure 1. Example protocol summary [11].

2.2 University of Sheffield's Approach for Subtask 1

Sheffield's approach to subtask 1 used the RAKE [12] keyword extraction algorithm to interpret protocols and Apache Lucene[1] as the IR engine.

The PubMed database was retrieved from the PubMed FTP site[2] as a set of XML files and indexed using Apache Lucene. The abstract and title of each citation is parsed

[1] `https://lucene.apache.org/`

[2] `ftp://ftp.ncbi.nlm.nih.gov/`

out of the XML files and concatenated. Each citation is pre-processed by carrying out tokenisation, stop word removal and stemming.

Information was extracted from the protocol summaries, rather than the full protocols, since our analysis suggested that this contained the key information that was useful for creating a search.

Protocol summaries were pre-processed by removing references, single characters, headers, titles and markup tags. RAKE is then applied to the remaining content with a minimum keyword frequency set to 1 (i.e. return all terms). The extracted terms are concatenated together to form a query which is used to retrieve citations from the Lucene index (see Figure 2 for example of the query generated from protocol summary shown in Figure 1).

> endemic countries objective ambulatory healthcare facilities rapid diagnostic tests falciparum parasitaemia detectable malaria endemic areas diagnostic accuracy falciparum malaria

Figure 2. Example of keywords extracted from protocol summary by RAKE.

2.3 Runs

Three runs were submitted using a range of retrieval strategies from Lucene.

- The **sheffield-Boolean** run uses terms that occur most frequently in the document and query as a basis for ranking. Documents that contain more query terms will feature higher in the overall rankings. The Apache Lucene Boolean similarity class[3] was used for the implementation.
- The **sheffield-tfidf** run uses a cosine similarity measure to compare the similarity between the query and the PubMed article. Documents and queries are represented as tf.idf weighted vectors. The Apache Lucene tf.idf similarity class[4] was used.
- The **sheffield-bm25** run uses the BM25 similarity measure [13] implemented using the Apache Lucene BM25 similarity class[5].

2.4 Results and Discussion

Results were computed over the training and test datasets and shown in Table 1 and Table 2.

[3] `https://lucene.apache.org/core/7_0_1/core/org/apache/lucene/search/similarities/BooleanSimilarity.html`

[4] `https://lucene.apache.org/core/7_0_1/core/org/apache/lucene/search/similarities/TFIDFSimilarity.html`

[5] `https://lucene.apache.org/core/7_0_1/core/org/apache/lucene/search/similarities/BM25Similarity.html`

Training Dataset Results for the training dataset are shown in Table 1. The Boolean search method achieves 0.341 recall for the 5000 documents returned. This approach is limited by the fact there is no weighting of term importance and the information used for ranking is based only on the number of query terms contained in documents. Using tf.idf leads to a slight improvement in performance (0.375 recall), presumably due to the availability of term weighting information. The best recall score for the training data (0.587) is obtained using BM25. The overall pattern of results is as expected for the training data, particularly the relatively strong performance of BM25.

Table 1. Results of No-Boolean search for training dataset (5000 documents returned).

RUN-ID	recall@50	recall@500	recall@5000	norm_area	ap
sheffield-boolean	0.036	0.146	0.341	0.247	0.007
sheffield-tfidf	0.026	0.126	0.375	0.263	0.007
sheffield-bm25	0.078	0.273	0.587	0.431	0.034

Test Dataset Results for the test dataset are shown in Table 2. Performance using the tfidf ranking method was surprisingly poor on this dataset. This may be due to the use of RAKE to extract keywords which reduces the impact of the idf element of the tfidf similarity measure. As with the training data, BM25 achieves the best performance, although generally lower than for the training data.

Table 2. Results of No-Boolean search for test dataset (5000 documents returned).

RUN-ID	recall@50	recall@500	recall@5000	norm_area	ap
sheffield-boolean	0.022	0.124	0.299	0.244	0.018
sheffield-tfidf	0.005	0.057	0.266	0.184	0.005
sheffield-bm25	0.045	0.169	0.426	0.372	0.039

Overall RAKE worked well with Apache Lucene as an approach for retrieving relevant documents using protocol summaries as search information. In future, we could make use of RAKE's ability to extract phrases, rather than just terms.

3 Subtask 2: Abstract and Title Screening

Subtask 2 focuses on the second stage of conducting systematic reviews (*Title and abstract screening*). Participants are asked to rank the list of PubMed Document Identifiers (PMIDs) returned from the Boolean query with the goal that all relevant citations appear as early as possible.

3.1 Datasets

The dataset for this subtask consists of 72 DTA reviews. This dataset divided into 42 reviews for training dataset and 30 reviews for test dataset. For each review, participants are provided with the topic ID, title of the review (written by Cochrane experts), a Boolean query using either OVID or PubMed syntax (manually constructed by Cochrane experts), set of PMIDs returned by running the query in MEDLINE database and relevance judgement at both abstract and content levels. Figure 3 shows an example topic from the training dataset.

Topic: CD010705

Title: The diagnostic accuracy of the GenoType® MTBDRsl assay for the detection of resistance to second-line anti-tuberculosis drugs.

Query:

```
MTBDR*.ti,ab.
Genotype MTBDR*.ti,ab
or/1-2
exp Tuberculosis, Pulmonary/
exp Tuberculosis, Multidrug-Resistant/
MDR-TB.ti,ab
XDR-TB.ti,ab
Mycobacterium tuberculosis/
TB.ti,ab
tuberculosis.ti,ab
or/4-10
3 and 11
```

Pids:
24429319 24197880 24172155 24098523 24056651 24046537 24039735 24029194 23895665 23883707 23808160 23782980 23689727 23658272 23633684 23467605 23392466 23383320 ....

Figure 3. Example topic from Cochrane reviews used in training dataset [14].

3.2 University of Sheffield's Approach to Subtask 2

The University of Sheffield's submission for subtask 2 extended the approach used for CLEF 2017 [15] by augmenting queries with terms designed to identify DTA studies and by making use of relevance feedback.

Our approach used the topic title and terms from the Boolean query. A simple parser was used to extract terms from the Boolean query automatically. The topic title and terms extracted from the query were pre-processed by tokenisation, stemming and removal of stop words[6]. The same pre-processing steps were applied to the title and abstract for each PMID returned for the Boolean query.

3.3 Runs

Three runs were submitted to the subtask 2 official evaluation: sheffield-query_terms, sheffield-general_terms and sheffield-feedback.

– The **sheffield-query_terms** run ranks abstracts by comparing each citation against topic title and terms extracted from the query. We used tf.idf weighted vectors to represent information obtained from the topic and citations then calculate the similarity between them using the cosine metric[7]. Abstracts are ranked based on this similarity score. (N.B. This run is the same as our best performing run from CLEF 2017, i.e. Sheffield-run-4.)
– The **sheffield-general_terms** run extends the previous approach (sheffield-query_terms) by enriching queries with terms designed specifically to identify DTA studies. Terms from standard filters developed to identify DTA studies [16] were added to the queries (see Figure 4).

```
'sensitivity', 'specificity', 'diagnos', 'diagnosis',
             'predictive', 'accuracy'
```

Figure 4. Sample of search filters for diagnosis - MEDLINE.

– The **sheffield-feedback** run used relevance feedback to re-rank abstracts based on relevance judgements. 10% of the abstracts (up to a maximum of 1,000) were used to update the query vector using Rocchio's algorithm (equation 1) and the abstracts re-ranked [17].

$$\vec{q}_m = \alpha\vec{q} + \frac{\beta}{N_r}\sum_{\forall\vec{d_j}\in D_r}\vec{d}_j - \frac{\gamma}{N_n}\sum_{\forall\vec{d_j}\in D_n}\vec{d}_j \qquad (1)$$

where $\vec{q}$ is the original query vector, $\vec{d}_j$ is a weighted term vector associated with abstract j. D_r is the set of relevant abstracts among the abstracts retrieved and N_r is the number of abstracts in D_r. D_n is the set of non-relevant abstracts among the abstracts retrieved and N_n is the number of abstracts in D_n. A range of values for the weighting parameters (α, β and γ) were explored using the training data and it was found that the best results were achieved by setting α = β = 1 and γ = 1.5.

[6] NLTK's `tokenize` and `LancasterStemmer` packages were used for tokenisation and stemming. PubMed's stop word list was used: `https://www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T.stopwords/`

[7] Scikit-learn's `TfidfVectorizer` and `linear_kernel` packages were used for this steps

3.4 Results and Discussion

Training Dataset Table 3 shows the results[8] of applying our approach on the training dataset. As expected, all submitted runs outperform the baseline where the list of PubMed abstracts is randomly ordered.

Performance improves when general terms are added to the queries (comparing sheffield-query_terms and sheffield-general_terms). These results demonstrate the usefulness of including general terms which provide information about the types of citations that are likely to be relevant for DTA reviews, independently of their specific topic.

The best performance was achieved using relevance feedback (sheffield-feedback). The average precision (ap) increased by 0.577 when compared with the baseline. It also produced the best results for work saved oversampling (wss@100 and wss@95), and the norm area was also improved by 0.359, 0.522 and 0.423 respectively. An improvement in performance is to be expected when relevance feedback is used, given that additional information available in the relevance judgements. The scale of the improvement demonstrates just how useful this information is for the task.

Table 3. Results of runs evaluated against training dataset using abstract qrels.

RUN-ID	ap	last_rel	wss@100	wss@95	norm_area
sheffield-baseline	0.043	4417	0.076	0.085	0.503
sheffield-query_terms	0.199	2741	0.370	0.505	0.850
sheffield-general_terms	0.220	2731	0.376	0.529	0.868
sheffield-feedback	**0.620**	**2410**	**0.435**	**0.607**	**0.926**

Test Dataset Table 4 shows the results on the test dataset. The pattern of performance is similar to that observed for the training dataset. The best performance was achieved by using relevance feedback (sheffield-feedback). The ap improved by 0.556 when compared with baseline. The wss@100, wss@95 and the norm area were also improved by 0.421, 0.606 and 0.418 respectively.

Results from both training and test datasets demonstrate that retrieval performance for technology-assisted reviews can be improved by adding additional terms indicating the type of citation likely to be on interest for the review are added to the queries and by applying relevance feedback.

[8] Using the script provided by task organisers: `https://github.com/leifos/tar`.

Table 4. Results of runs evaluated against test dataset using abstract qrels.

RUN-ID	ap	last_rel	wss@100	wss@95	norm_area
sheffield-baseline	0.051	7221	0.023	0.029	0.500
sheffield-query_terms	0.224	5737	0.377	0.506	0.849
sheffield-general_terms	0.258	5519	0.431	0.552	0.871
sheffield-feedback	**0.607**	**5171**	**0.444**	**0.635**	**0.918**

4 Conclusions

This paper presented the University of Sheffield's approach to CLEF2018 task 2. For subtask 1, we established a method for retrieving document using search protocol summaries. We showed RAKE was an effective way of identifying keywords in the protocol from which queries could be created. For subtask 2, results demonstrated that augmenting queries with terms designed to identify DTA studies and applying relevance feedback improve retrieval performance.

Bibliography

1. M. Di Nisio, A. Squizzato, A. W. S. Rutjes, H. R. Büller, A. H. Zwinderman, and P. M. M. Bossuyt, "Diagnostic accuracy of D-dimer test for exclusion of venous thromboembolism: a systematic review," *Journal of Thrombosis and Haemostasis*, vol. 5, no. 2, pp. 296–304, 2007.
2. A. M. Cohen, K. Ambert, and M. McDonagh, "A prospective evaluation of an automated classification system to support evidence-based medicine and systematic review," *AMIA Annual Symposium Proceedings*, vol. 2010, pp. 121–125, 2010.
3. S. Karimi, S. Pohl, F. Scholer, L. Cavedon, and J. Zobel, "Boolean versus ranked querying for biomedical systematic reviews," *BMC medical informatics and decision making*, vol. 10, no. 1, pp. 1–20, 2010.
4. A. O'Mara-Eves, J. Thomas, J. McNaught, M. Miwa, and S. Ananiadou, "Using text mining for study identification in systematic reviews: a systematic review of current approaches," *Systematic Reviews*, vol. 4, no. 1, pp. 1–5, 2015.
5. S. Paisley, J. Sevra, M. Stevenson, R. Archer, L. Preston, and J. Chilcott, "Identifying Potential Early Biomarkers of Acute Myocaridal Infarction in the Biomedical Literature: A Comparison of Text Mining and Manual Sifting Techniques," in *Proceedings of the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) 19th Annual European Congress*, (Vienna, Austria), 2016.
6. M. Miwa, J. Thomas, A. O'Mara-Eves, and S. Ananiadou, "Reducing systematic review workload through certainty-based screening," *Journal of Biomedical Informatics*, vol. 51, pp. 242–253, 2014.
7. Shemilt, Khan, Park, and Thomas, "Use of Cost-effectiveness Analysis to Compare the Efficiency of Study Identification Methods in Systematic Reviews," *Systematic reviews*, 2016.
8. L. Goeuriot, L. Kelly, H. Suominen, A. Névéol, A. Robert, E. Kanoulas, R. Spijker, J. Palotti, and G. Zuccon, "CLEF 2017 eHealth Evaluation Lab Overview ," *CLEF 2017 - 8th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS), Springer, September*, 2017.

9. H. Suominen, L. Kelly, L. Goeuriot, E. Kanoulas, L. Azzopardi, R. Spijker, D. Li, A. Névéol, L. Ramadier, A. Robert, G. Zuccon, and J. Palotti, "Overview of the CLEF eHealth Evaluation Lab 2018," in *CLEF 2018 - 8th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS)*, (France), Springer, September 2018.
10. E. Kanoulas, R. Spijker, D. Li, and L. Azzopardi, "CLEF 2018 Technology Assisted Reviews in Empirical Medicine Overview," in *CLEF 2018Evaluation Labs and Workshop: Online Working Notes*, (France), CEUR-WS, September 2018.
11. K. Abba, J. Deeks, P. Olliaro, C. Naing, S. Jackson, Y. Takwoingi, S. Donegan, and P. Garner, "Rapid diagnostic tests for diagnosing uncomplicated *P. falciparum* malaria in endemic countries," *The Cochrane Database of Systematic Reviews*, 2011. CD008122.
12. R. Stuart, E. Dave, C. Nick, and C. Wendy, *Automatic Keyword Extraction from Individual Documents*, ch. 1, pp. 1–20. Wiley-Blackwell, 2010.
13. S. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford, "Okapi at TREC–3," in *Overview of the Third Text REtrieval Conference (TREC–3)*, pp. 109–126, Gaithersburg, MD: NIST, 1995.
14. "The diagnostic accuracy of the GenoType(®) MTBDRsl assay for the detection of resistance to second-line anti-tuberculosis drugs," *The Cochrane database of systematic reviews*, 2014. CD010705.
15. A. Alharbi and M. Stevenson, "Ranking abstracts to identify relevant evidence for systematic reviews: The University of Sheffield's approach to CLEF eHealth 2017 Task 2 ," in *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings, (Dublin, Ireland), CEUR-WS.org, September 11-14 2017.
16. "Health Information Research Unit. Search Filters for MEDLINE in Ovid Syntax and the PubMed translation [Internet]. 2016 [updated 2016 February 09 ;cited 2018 January 15]."
17. R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval: The Concepts and Technology Behind Search*. USA: Addison-Wesley Publishing Company, 2nd ed., 2011.