Ranking Abstracts to Identify Relevant Evidence for Systematic Reviews: The University of Sheffield's Approach to CLEF eHealth 2017 Task 2

Working Notes for CLEF 2017

Amal Alharbi and Mark Stevenson

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

Abstract This paper describes Sheffield University's submission to CLEF 2017 eHealth Task 2: Technologically Assisted Reviews in Empirical Medicine. This task focusses on the identification of relevant evidence for systematic reviews in the medical domain. Participants are provided with systematic review topics (including title, Boolean query and set of PubMed abstracts returned) and asked to identify the abstracts that provide evidence relevant to the review topic. Sheffield University participated in the simple evaluation. Our approach was to rank the set of PubMed abstracts returned by the query by making use of information in the topic including title and Boolean query. Ranking was based on a simple TF.IDF weighted cosine similarity measure. This paper reports results obtained from six runs: four submitted to the official evaluation, an additional run and a baseline approach.

1 Introduction

Systematic reviews attempt to identify, synthesise and summarise evidence available to answer a research question. They form the backbone of evidence-based approaches to medicine where they are used to answer complex questions such as "How effective are statins for heart attack survivors?" [1].

The process of creating a systematic review is time-consuming with a single review often requiring 6 to 12 months of effort from expert reviewers [2,3]. Text mining techniques have been shown to be a useful way to reduce this effort [4,5,6,7]. CLEF eHealth Task 2 "Technologically Assisted Reviews in Empirical Medicine" focusses on the application of text mining to the process of developing systematic reviews with the aim to reduce the effort required.

This paper is organised as follows: Section 2 introduces CLEF eHealth Task 2. Section 3 describes our approach to this task. Section 4 discusses the results obtained from applying this approach to both the development and test datasets. Finally, Section 5 presents the conclusions and potential future work.

2 Task Description

The process of identify relevant evidence for a systematic review usually consists of multiple stages [8]:

1. **Boolean Search:** Experts construct a boolean query designed to identify all evidence relevant to the review question. This query is run against a medical database such as PubMed and set of titles and abstracts returned.
2. **Title and Abstract Screening:** Experts screen the titles and abstracts retrieved to identify those that are potentially relevant for inclusion in the review.
3. **Document Screening:** The full document content is then retrieved for any title and abstract that has been identified as being relevant in the previous stage. These are then examined in a second round of expert screening to form a final decision about their relevance to the review.

In CLEF eHealth 2017 [9], Task 2 [8] focuses on the second stage of systematic review (Title and Abstract Screening). Participants are required to develop methods to rank a list of PubMed abstracts returned by a boolean query (stage 1) so that relevant documents appear as early as possible.

3 Method

3.1 Datasets

Participants are provided with two datasets: a development set and a test set. The development dataset contains 20 topics and the test dataset contains 30 topics. All reviews focus on Diagnostic Test Accuracy (DTA). The queries were manually constructed by expert reviewers from the Cochrane collaboration[1]. For each topic, participants are provided with topic id, review title, boolean query and a list of PubMed documents identifiers retrieved by the query. The collection contains a total of 266,967 abstracts.

Figure 1 shows examples of two topics from the development dataset. Two different formulations were used for the Boolean queries: OVID and PubMed. The queries are generally complex and contain multiple operators. Table 1 shows operators commonly used in both types of query [3].

Participants also provided with files that indicate which of the titles and abstracts returned by the Boolean query were indicated as being relevant after the *Title and Abstract Screening* and *Document Screening* stages (see Section 2), referred to as the abstract qrels and content qrels respectively.

3.2 University of Sheffield's Approach

The University of Sheffield's submission to Task 2 ranked the list of PubMed abstracts retrieved for each topic with the intention of returning relevant ones as early as possible. The approach is completely automatic since queries are processed algorithmically and without manual intervention[2]. In addition, relevance feedback is not used.

Our method makes use of three pieces of information from the topic: (1) the title, (2) terms extracted from the Boolean query and (3) MeSH terms extracted from the Boolean query. Information for (2) and (3) are extracted from the Boolean query using a simple parser designed to interpret both OVID and PubMed style queries. Terms

[1] http://www.cochrane.org/

[2] The approach was implemented using Python v3.6

OVID

Topic: CD009591
Title: Imaging modalities for the non-invasive diagnosis of endometriosis
Query:

```
exp magnetic resonance imaging/ or exp ultrasonography/ or exp
Imaging, Three-Dimensional/ or exp radiography/
ultraso$.tw. or magnetic resonance imaging.tw. or MRI.tw. or imag$.tw.
diagnos$.tw.
...
(animals not (humans and animals)).sh.
8 not 9
```

PubMed

Topic: CD008643
Title: Red flags to screen for vertebral fracture in patients presenting with low-back pain
Query:

```
1 Index test: clinical red flags
"Medical History Taking"[mesh] OR history[tw] OR "red flag"[tw]
OR "red flags" OR Physical examination[mesh] OR "physical examination"
[tw] OR "function test"[tw] OR "physical test"[tw]
...
1 AND 2 AND 3 NOT 4
```

Figure 1. Example topics from Cochrane reviews used in development dataset [10,11].

Table 1. OVID and PubMed common query operators

OVID	
/ or .sh.	MeSH terms
.mp.	MeSH subheading
.tw.	Text words
.ti,ab.	Title/abstract
PubMed	
[mesh] or [mh]	MeSH terms
[sh]	MeSH subheading
[tw]	Text words
[tiab]	Title/abstract

and MeSH terms modified by certain operators (e.g. `not` and `adj`) are not extracted. Figure 2 shows examples of terms extracted from the query for topic `CD008643` (see Figure 1). Some MeSH terms (e.g. `Spine`) are also standard English words that could appear as a term in an abstract. To avoid false matches all MeSH terms extracted from a query are prefixed with the string `Mesh`. In addition, MeSH terms are preprocesssed to remove whitespace and punctuation (e.g. `Lumbar vertibrae` becomes `MeshLumbarvertibrate`). Example MeSH terms extracted from the same query are shown in Figure 3.

```
'history', 'red flag', 'physical examination', 'function test',
'physical test','clinical', 'clinically','diagnosis'
```

Figure 2. Sample of terms extracted from the query of topic CD008643

```
'MeSHMedicalHistoryTaking', 'MeSHPhysicalexamination',
'MeSHra', 'MeSHri', 'MeSHWoundsandInjuries'
```

Figure 3. Sample of MeSH headings extracted from the query of topic CD008643

The abstracts returned by the Boolean query for each topic defined as the list of PMIDs (PubMed identifier) provided with the topic are downloaded from PubMed[3]. The text of the title, abstract and MeSH terms are extracted and the MeSH terms preprocessed using the same approach that was applied to the Boolean query.

Pre-processing is applied to both the PubMed abstracts and information extracted from the topics. The text is tokenised, converted to lower case, stop words/punctuation are removed and the remaining tokens stemmed[4].

The information extracted from the topic and each of the abstracts are converted into tf.idf-weighted vectors. The similarity between the topic and each of the abstracts is then generated by computing the cosine metric for the pair of vectors[5]. Abstracts are ranked based on this similarity score.

Results are output in the TREC format shown in Table 2 where:

- TOPIC-ID: topic identifier provided by CLEF 2017.
- INTERACTION: this field is assigned the value NF in all our runs to indicate that relevance feedback is not used
- PID: PubMed document identifier
- RANK: rank of the document according to the cosine similarity score
- SCORE: cosine similarity score described above
- RUN-ID: run identifier

3.3 Runs

Four runs were officially submitted for the official evaluation: Sheffield-run-1, Sheffield-run-2, Sheffield-run-3, and Sheffield-run-4. In addition, a baseline run (Sheffield-baseline) and additional approach (Sheffield-run-5) were also implemented and evaluated. A description of each run is presented below.

[3] The `Entrez` package from `biopython.org` was used.

[4] NLTK's `tokenize` and `LancasterStemmer` packages are used for tokenisation and stemming. The list of stop words provided by scikit-learn (`scikit-learn.org/stable/`) is used for most runs.

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

Table 2. Sample output for Sheffield-run-1

TOPIC-ID	INTERACTION	PID	RANK	SCORE	RUN-ID
CD010438	NF	18388501	17	0.245	Sheffield-run-1
CD010438	NF	16884987	18	0.239	Sheffield-run-1
CD010438	NF	22164456	19	0.238	Sheffield-run-1
CD010438	NF	22193152	20	0.236	Sheffield-run-1

- **Sheffield-baseline** In this run the list of PubMed abstracts are randomly ordered. This is intended to represent the scenario in which the results of the Boolean query are simply evaluated in the order in which they are retrieved without any attempt to identify those most likely to be relevant. This situation simulates common practise within many systematic review projects in which reviewers examine each of the retrieved abstracts in turn. The score of each abstract is calculated using the following equation:

  $$score = \frac{n - r + 1}{n} \tag{1}$$

  where n is the total number of abstracts returned by the Boolean query and r the abstract's rank in the random ordering.
- **Sheffield-run-1** Abstracts returned by the Boolean query are ranked by comparing them against only the topic title.
- **Sheffield-run-2** Abstracts are compared with the topic title and terms extracted from the Boolean query.
- **Sheffield-run-3** Abstracts are compared with the topic title and both terms and MeSH terms extracted from the Boolean query.
- **Sheffield-run-4** This run is the same as Sheffield-run-2 except that the PubMed stop-words list [12] is used rather than the one from sklearn.
- **Sheffield-run-5** Abstracts are compared against the topic title and MeSH terms extracted from the Boolean query. (This run is the same as Sheffield-run-3 except that terms extracted from the Boolean query are not included when computing the similarity.)

4 Results and Discussion

Task 2 consists of two formal evaluations: simple evaluation and cost-effective evaluation. The University of Sheffield participated only in the simple evaluation setup and did not attempt to optimise the approach for the cost-effective evaluation. Evaluation was carried out using the script provided by the task organisers[6].

4.1 Development Dataset

The development dataset contains of 20 DTA topics (see Section 3.1). Tables 3 and 4 present the results for the approaches described in Section 3.3 applied to this dataset for the abstract and content qrels respectively.

[6] https://github.com/leifos/tar

As expected, all of the implemented methods outperform the simple baseline approach. This demonstrates that even straightforward ranking techniques provide potential benefit to systematic reviewers by ensuring that documents more likely to be relevant are placed higher in the rankings. We have previously demonstrated a similar results for a single systematic review [5] and that finding is supported by these results which represent a substantially larger dataset.

The best result of the submitted runs for the abstract qrels (Table 3) was achieved by Sheffield-run-4 which achieved the average precision (ap) score of 0.223, an improvement of 0.173 against the baseline. It also achieved the best results for work saved over sampling (wss) and area under the cumulative recall curve normalized by the optimal area (norm_area) metrics. It is also close to the best result for the average of the minimum number of abstracts returned to retrieve all relevant ones (last_rel) metric.

For the content qrels (Table 4), both Sheffield-run-4 and Sheffield-run-5 are strong. Sheffield-run-4 produced the best scores for last_rel and norm_area and close to the best result of wss. Sheffield-run-5 achieved the best score for ap and wss_95.

Results from the development dataset suggest that including terms extracted from the Boolean query is beneficial (e.g. compare Sheffield-run-1 and Sheffield-run-2). However, the usefulness of MeSH terms extracted is less clear. Performance decreases when these are added to the title and query terms (e.g. compare Sheffield-run-2 and Sheffield-run-3). Results are mixed when they are used instead of query terms (e.g. compare Sheffield-run-1 and Sheffield-run-5), there is no improvement for the abstract evaluation but some benefit for the content evaluation.

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

RUN-ID	ap	last_rel	wss_100	wss_95	norm_area
Sheffield-baseline	0.05	7121.65	0.036	0.033	0.495
Sheffield-run-1	0.188	5793.7	0.138	0.385	0.815
Sheffield-run-2	**0.223**	**4449.65**	0.184	0.434	0.836
Sheffield-run-3	0.217	4768.85	0.17	0.415	0.83
Sheffield-run-4	**0.223**	4496.85	**0.188**	**0.442**	**0.839**
Sheffield-run-5	0.182	5866.6	0.135	0.344	0.808

Table 4. Results of runs evaluated against development dataset using content qrels

RUN-ID	ap	last_rel	wss_100	wss_95	norm_area
Sheffield-baseline	0.01	6575.3	0.104	0.077	0.465
Sheffield-run-1	0.094	2204.95	**0.574**	0.61	0.855
Sheffield-run-2	0.104	2097.2	0.549	0.589	0.867
Sheffield-run-3	0.095	2141.35	0.533	0.593	0.859
Sheffield-run-4	0.107	**1999.35**	0.568	0.611	**0.875**
Sheffield-run-5	**0.108**	2701.7	0.545	**0.615**	0.855

4.2 Test Dataset

The development dataset contains of 30 DTA topics (see Section 3.1). Tables 5 and 6 show the results for the abstract and content qrels respectively.

The highest ap scores were achieved using Sheffield-run-2 and Sheffield-run-4 for both the abstract and content qrels (Tables 5 and 6). The overall pattern of results suggest that Sheffield-run-4 is the best performing run on the test data.

Results from the development and test datasets indicate the strong relative performance of Sheffield-run-4. This indicates that including terms extracted from Boolean query and using the PubMed stop-words list are benefical for this task.

Table 5. Results of runs evaluated against test dataset using abstract qrels

RUN-ID	ap	last_rel	wss_100	wss_95	norm_area
Sheffield-baseline	0.045	3727.433	0.039	0.031	0.483
Sheffield-run-1	0.17	2678.333	0.31	0.422	0.818
Sheffield-run-2	**0.218**	2441.7	0.385	**0.493**	0.845
Sheffield-run-3	0.199	2404.967	0.384	0.473	0.841
Sheffield-run-4	**0.218**	**2382.467**	**0.395**	0.488	**0.847**
Sheffield-run-5	0.158	2650.8	0.303	0.423	0.809

Table 6. Results of runs evaluated against test dataset using content qrels

RUN-ID	ap	last_rel	wss_100	wss_95	norm_area
Sheffield-baseline	0.023	3307.793	0.088	0.067	0.478
Sheffield-run-1	0.12	**1801.724**	0.517	0.544	0.844
Sheffield-run-2	0.176	1928.828	0.534	0.58	0.87
Sheffield-run-3	0.153	1902.586	0.524	**0.588**	0.866
Sheffield-run-4	**0.177**	1846.586	**0.543**	0.587	**0.874**
Sheffield-run-5	0.114	1922.103	0.487	0.541	0.836

There were some relevant documents in the test data set for which our approach assigned a score of 0 and this caused NCG@100 scores to be less than 1. This was observed at both the content and abstract level for the development and test datasets. The scoring script treats these documents as not being included in the ranking. The problem could be resolved by adding a small delta value to each score.

5 Conclusion and Future Work

This paper described the University of Sheffield's approach to CLEF 2017 Task 2. Information from the review title and Boolean query was used to rank the abstracts returned by the query using standard similarity measures. The title and terms extracted from the Boolean query were found to be the most useful information for this task. All of the submitted runs outperform a baseline approach based on random ordering.

In future we plan to refine the techniques for extracting terms and MeSH terms from the Boolean query (Section 3.2) by taking account of the query structure and MeSH hierarchy. We also plan to develop techniques to minimise the cost of identifying relevant evidence and make use of ActiveLearning to improve the ranking based on feedback from reviewers.

References

1. D. Gough, S. Oliver, and J. Thomas, *An Introduction to Systematic Reviews*. Sage, 2012.
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, 2010.
4. 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.
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. I. Shemilt, N. Khan, S. Park, and J. Thomas, "Use of Cost-effectiveness Analysis to Compare the Efficiency of Study Identification Methods in Systematic Reviews," *Systematic reviews*, 2016.
7. A. O'Mara-Eves, J. Thomas, J. McNaught, M. Miwa, and S. Ananiadou, "Using Text Mining for Study Identification in Systematic Reviews," *Systematic reviews*, 2015.
8. E. Kanoulas, D. Li, L. Azzopardi, and R. Spijker, "CLEF Technologically Assisted Reviews in Empirical Medicine Overview," 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.
9. 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.
10. V. Nisenblat, P. M. Bossuyt, C. Farquhar, N. Johnson, and M. L. Hull, "Imaging Modalities for the Non-invasive Diagnosis of Endometriosis," *Cochrane Database of Systematic Reviews 2016*, vol. 2, no. CD009591, 2016.
11. C. Williams, N. Henschke, C. Maher, M. van Tulder, M. Koes, P. Macaskill, and L. Irwig, "Red Flags to Screen for Vertebral Fracture in Patients Presenting with Low-back Pain," *Cochrane Database of Systematic Reviews 2013*, vol. 1, no. CD008643, 2013.
12. "[table, stopwords] - pubmed help - ncbi bookshelf." [online] Available at: https://www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T.stopwords/ [Accessed 7 May 2017].